Distinctive Characteristics of Cancer-associated Genes and Point Mutations Driving Carcinogenesis Through Computational Modelling

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine, and Health

2023

Amro Safadi

School of Biological Sciences

Table of Contents

Word Count: 36,521

List of Figures

List of Tables

List of Acronyms

Abstract

The distinctive nature of cancer as a disease prompts an exploration of the special characteristics the genes and mutations implicated in cancer exhibit. Currently, we have no clear explanation for why patterns of replacements of amino acids are frequent in cancer, and what their effects may be on the protein. Such patterns would be expected to provide an understanding of how these amino acid replacements drive cancer progression and reveal the properties that distinguish them from replacements that are non-cancer associated. Moreover, the identification of cancer-associated genes and their characteristics is crucial to further our understanding of this disease. These characteristics can be used to recognise and prioritise therapeutic drug targets with an enhanced likelihood of success. However, the rate at which cancer genes are being identified experimentally is slow. Applying predictive analysis techniques, through the building of accurate machine learning models, is potentially a useful approach in enhancing the identification rate of these genes and their characteristics. In this work, we identified certain amino acid residues and replacements to be highly enriched in cancer. In particular, we highlight 17 substitutions showing high enrichment rates also we find that very frequently in cancer a residue is replaced with either a Cys or an aromatic residue. We explained the role of Cys in forming disulphide bonds and the aromatic amino acids in forming stacking interactions; both are known to be vital in binding activities highly enriched in cancer-associated gene functions. We also identified properties, such as protein stability and hydrophobicity that have distinguished patterns in these cancer-associated replacements compared to other non-cancer-associated mutations. We used these properties to train a machine learning model predicting cancer-associated replacements related to specific protein using only the amino acids residue position and physico-chemical properties. In terms of cancer-associated genes, we investigated gene essentiality and found that essentiality scores tend to be higher for cancer-associated genes compared to other protein-coding human genes. We built a dataset of extended gene properties linked to essentiality and used it to train a machine-learning model; this model reached 89% accuracy and > 0.85 for the Area Under Curve (AUC). The model showed that essentiality, evolutionary-related properties, and properties arising from protein-protein interaction networks are particularly effective in predicting cancer-associated genes. We were able to use the model to identify potential candidate

genes that have not been previously linked to cancer. Prioritising genes that score highly by our methods could aid scientists in the identification of novel genes and targets for further research.

Declaration Statement

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and she has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses.

Acknowledgements

All praise and thanks are due to Allah SWT.

"Read in the name of your Lord, who created" Quran (96:1)

"And they encompass not a thing of his knowledge except what he wills" Quran (2:255)

I dedicate this work to my children. May they and their children one day contribute to science for the good of the world.

I also dedicate this work to every person suffered from cancer especially the innocent Syrian children and adults whom were exposed to chemical agents during the recent uprising and had to suffer these atrocities alone.

I would like to express my sincere thanks and gratitude to my supervisors Prof. Simon Lovell and Prof. Andrew Doig. Their support, constructive feedback and suggestions allowed this research to exist. I also would like to thank Prof. David Robertson - the University of Glasgow for his support when I started my work.

1. Chapter 1: Introduction

Cancer is the second leading cause of death in the UK with more than one in four of all deaths in 2019 reported here by Cancer Research (UK) attributed to cancer. Cancer comprises a collection of diseases with shared biological base that could be described as a breakdown of specific controls that are normally imposed on multicellular systems (*1*). In particular, if processes such as growth and apoptosis are directly affected (*1*).

Research into tumor initiation and progression has established a number of hallmarks that most cancers exhibit (*2*). These include (i) proliferative signalling: the most recognisable feature of cancer cells is their ability to proliferate with no restriction. The growth promoting or inhibiting factors that normal cells use as means of control are abolished in tumor cells. Tumor cells maintain proliferation by producing their own growth factors and they also stimulate normal tumor-associated cells (stroma), which in turn provide the growth factors to the cancer cells. Moreover, cancer cells can become extra responsive to any growth signalling. Eventually cancer cells would become self-regulating (*3*). (ii) Evading growth suppressors: if proliferation inhibitors are mutated, this can lead to the cell to proliferate without control; an example of this is the impact of mutated proliferation inhibitor TGF-β (*2*). (iii) Resisting cell death: tumor cells can become invulnerable to apoptosis by avoiding mechanisms that in normal cells initiate apoptosis. An example is mutated p53 causing the malfunction of the mechanism that normally detects irreparable DNA damage and initiate apoptosis (*2*) . They can also avoid apoptosis by the expression of anti-apoptotic proteins such as Bcl-2 or by the down-regulation or mutation of pro-apoptotic proteins such as Bax (*4*). (iv) Enabling replicative immortality: normal cells replicate for limited number of times due to a shortening of telomere length while tumor cells overcome this mechanism by overexpressing telomerase (an enzyme that maintains telomere length) and thus can potentially replicate indefinitely (*5*). (v) Inducing angiogenesis: angiogenesis is the formation of new blood vessels from pre-existing ones in response to chemical signals. Tumors need a dedicated blood supply to provide the oxygen and nutrients to grow. Tumor cells use abnormal secretion of various growth factors to induce blood vessel growth (*2*). (vi) Invasion and metastasis: these processes start when cancer cells acquire the ability to penetrate the neighbouring tissues leading into spreading

cancer into distant organs. Metastasis usually occurs at an advanced stage of tumor development and is considered responsible for an increased chance of mortality (*2*). Two further potential hallmarks were proposed in recent research: reprogramming of energy metabolism and evading immune destruction (*2*).

Most hallmarks of cancer are principally caused by the existence of genetic variations and genome aberrations. These changes are determined in the downstream proteins and altered pathways that were proven to be associated with cancer initiation and progression (*6*). This underlines the important of studying cancer not just by only looking at individual components (genes, proteins, etc.) but to link this with an understanding of their interaction networks and cellular pathways.

1.1 Cellular processes affected by cancer associated mutations:

Biomedical research that described the varied cellular processes in humans has provided cancer researchers with the crucial knowledge to recognise the core cellular processes that are affected by cancer (*7*). Furthermore, cancer researchers are now able to identify the individual signalling pathways that are altered because of cancer mutations (*8*). These mutations (somatic or germline) are mainly responsible for causing the start of the cascade of steps that disrupt and break specific controls within core cellular processes leading to the transformation of normal cells into cancer cells (*6*). There are primarily 12 signalling pathways that are particularly affected by cancer associated mutations. These regulate three core cellular processes: cell fate, cell survival, and genome maintenance (*6*). It is important to remember that these pathways are not entirely separate from each other; as a gene that is implicated in one of these pathways might have a protein product that interacts with a protein that is involved in another pathway.

1. **Cell fate:** In many eukaryotes, differentiated cells have their distinctive set of properties defined by gene expression (*9*). The mechanism that maintains the genes' expression levels must be maintained during cell division. Multiple signalling pathways were shown to be involved in determining cell fate (*9*). In cancer, epigenetic variations can alter the process controlling how cell differentiate and

divide and ultimately changing the tissue architecture (*10*). Two pathways that are particularly regulated by the expression of genes are chromatin modification, and transcriptional regulation. DNA-binding proteins can be responsible for the increase or decrease of level of transcription of specific genes. Other pathways include:

a) **The Wnt/APC signalling pathway:** The Wnt pathway has the largest number of mutations in human tumors involving several tissues and cancer types (*11*). This pathway has complex branches, intersections and connections with other pathways and is part of a variety of biological processes, such as cell cycle progression and proliferation, inhibition of apoptosis, cell growth and cell migration (*12*). There are many genes implicated in this pathway encompassing a large variety of functions, including cell kinase regulation, cell adhesion, hormone signalling and transcriptional regulation. At least 20 genes have been identified that encode proteins that activate the cell cycle and/or proliferation. In particularly, the genes *WNT* and *APC* which encode the protein APC (adenomatous polyposis coli) are found often mutated here due to its interaction with several other proteins (*11*).

b) **The NOTCH pathway:** The Notch protein and its ligands are transmembrane proteins that are important regulators involved in many developments, cell fate and survival processes (*13*). Genes that code for the Notch proteins, such as *NOTCH2* and *NOTCH4,* were often found mutated in leukaemia, breast cancers and in several common cancers (*11*).

c) **The Hedgehog pathway (Hh):** The hedgehog signalling pathway is closely associated with the primary cilium delivering it alongside Gli Zinc finger proteins into the nucleus to activate target genes (*14*). Two proteins, patched (Ptch) and smoothened (Smo), are essential in initiating Hh signalling (*14*). Mutations in genes encoding signalling proteins of the hedgehog pathway, such as *PTCH1* and *SMO,* are most frequently observed among basal cell and brain carcinomas (*10*).

2. **Cell survival:** When certain mutations acquired by cancer cells lead to uncontrolled proliferation, these cells will consume the resources in their environment and thrive in a way other cells cannot. Many genes can be involved in this process where for

example, growth factors or receptors (part of a signalling pathway) become constantly activated due to mutations in these genes, this in turn can lead to proliferation (*10*). Moreover, genes that directly regulate cell cycle and apoptosis could have mutations that enhance the survival of the cancer cell. Therefore, tumor suppressor genes such as *CDKN2A*, *MYC*, and *BCL2s*, are often mutated in cancers (*10, 11*). Another gene whose mutations enhance cell survival is *VHL*, the product of which stimulates angiogenesis through the secretion of vascular endothelial growth factor (*15*). The signalling pathways affected by cell survival processes include:

a) **The RAS/MAPK pathways:** *RAS* is the generic term used for three oncogenes found to be responsible for driving tumor initiation and progression. *RAS* was among the earliest oncogenes to be detected in the history of cancer research. Three genes (*HRAS, KRAS* and *NRAS*) code for small G-proteins (GTPRas switch) that are anchored in the plasma membrane (*16*). These protein product of *RAS* genes are regarded as a molecular switches of signal transmission where the signal is transferred from the membrane into the cells by series of interactions. Oncogenic mutations in *RAS* genes cause the Ras protein to be constitutively active in its function as a signal transmitter disabling activation and inactivation mechanisms in the unmutated form (*17*). The MAPK (mitogen-activated protein kinase) pathway feature two important proteins coded by the genes *KRAS* and *BRAF*. The pathway can be divided into two main sections: upstream and downstream, based on the interaction between Ras and BRaf proteins (*18*). The upstream starts with the activation of a receptor tyrosine kinase leading to an activation of membrane bound protein Ras. Activation of BRaf by Ras protein initiates the downstream section (kinases and transcription factors). BRaf then phosphorylates several cytosolic proteins and thus transmits the signal into the cell. When the *RAS* gene carries certain point mutations, the mutated Ras protein cannot be inactivated by the GTPase protein and is constitutively active, even in the absence of an upstream signal. In some cases, the *BRAF* gene is mutated, and the dysregulation of the pathway occurs in the downstream section. This pathway is very important in human tumors. Indeed, it is dysregulated in more than 50 % of all human tumors (*11*). The MAPK pathway mediates

cellular processes, such as activation of the cell cycle and proliferation, and has been a major area of anticancer drug development research (*18*).

b) The PI3K/ACT pathway: This plays an important role in regulating growth, survival and division of cells and is activated by extracellular signals primarily through receptor tyrosine kinases (*19*). This pathway has multiple circuits and not all of the circuits functional involvement is yet fully explored. Like in the MAPK pathway, the Ras protein can play a role in PI3K/ACT pathway where phoshatidylinositol-3 kinase (PI3K) can be activated by binding to the GTP-bound form of the membrane protein Ras (*10*). Alternatively, PI3K can be activated by binding directly or via the protein IRS-1 to the activated receptor tyrosine kinase. Active PI3K leads to activation of the serine-threonine kinase AKT. Active AKT phosphorylates many proteins, which leads to the inhibition of apoptosis and to the activation of translation and proliferation (*19*). The genes *PIK3CA* and *AKT1/AKT2* are frequently found to be mutated in human cancer (*19*).

c) **The TGF-β pathway:** TGF (Transforming Growth Factor) beta signalling regulates diverse cellular processes, including cell proliferation, differentiation and apoptosis. Its dysfunction can result in various kinds of diseases, such as cancer and tissue fibrosis (*20*). TGF-β signalling is tightly regulated at different levels along the pathway. TGF-β signalling is initiated by the binding of TGF-β to its serine and threonine kinase receptors on the cell membrane. Modulation of receptor activity is a critical step for TGF-β signalling regulation (*20*). Although much effort has been made to understand the regulatory mechanisms of TGF-β receptor activity and stability, many questions still await to be addressed. Several genes such as *TGFBR2, ACVR2* and *SMAD4* that codes for receptors or other components of this pathway are frequently observed mutated in tumors (*20*).

d) The JAK-STAT pathway: The Janus kinase/signal transducer and activator of transcription (JAK/STAT) signalling pathway constitutes a membrane-to-nucleus signalling module and is regarded as one of the central communication hubs in the cell (*21*). It is involved in in cellular proliferation and differentiation, organ development, and immune homeostasis. More

than 50 cytokines and growth factors have been recognised to play a role in the JAK/STAT signalling pathway where JAKs mediate tyrosine phosphorylation of receptors, and recruit one or more STAT proteins (*22*). The dysregulation of the JAK/STAT pathway is associated with various cancers and autoimmune diseases. For example, *JAK2* V617F mutation frequently occurs in myeloproliferative neoplasms (MPN) (*22*).

e) **The NF-kB pathway:** Nuclear Factor kappa B (NF -kB) is a transcription factor in a pathway that affects cellular responses, such as proliferation, differentiation, and survival (*23*). Mutations affecting this pathway occur mainly in the B-cell cancerous lineage, but also in other subsets of cancers. For example, mutations in *NFKBIA* and *NFKBIE* genes are usually implicated in causing the pathway to malfunction leading to B-cell lymphomas (*24*).

3. **Genome maintenance:** Genome maintenance is realised through cell cycle apoptosis pathway where DNA damage control checkpoints within the cell are applied. For example, the mitotic checkpoint safeguards against a mistake during cell division and check a full complement of chromosomes are received by daughter cells (*25*). These control checkpoints make sure that cells with mutations resulting from mistakes during duplication or from exposure to external factors are terminated and mitigate the risk of genome instability (*10*). Bypassing these checkpoints can lead to diseases like cancer. In a cancer cell, mutations found in specific genes (e.g., *TP53*) allow cancer cells to avoid these measures applied in normal circumstances where a damaged cell would be killed (apoptosis) (*10*).

Further understating of the interrelation between these pathways and the functional role the downstream protein products play could be vital for successful development of targeted therapies. This knowledge could provide an explanation into why some drugs have limited success despite initial encouraging trials (*8*). For example, drugs that inhibit mutant BRAF kinase activity show less degree of success in colorectal cancers. This was found to be because expression of EGFR in this type of cancer counteracts the growth inhibitory effects of the BRAF inhibitors (*26*).

1.2 The role of genetic variations in tumourogenesis

Tumor analysis points to multi sequential phases involving several oncogenes and tumor-suppressor genes in cancer cells (*27*). Multiple variants often in multiple affected genes are found necessary to accumulate overtime in order to allow for cancer initiation (*27*).

1.2.1 The different categories of cancer-associated mutations

Alterations to the DNA that encodes a gene can occur in several ways and these changes will lead to a change of the mRNA produced. In turn, the altered mRNA may lead to the production of a protein that no longer functions properly affecting cell processes and contributing to the ultimate transformation of the normal cell into cancerous one. These changes can be either loss or gain of a function. The most common type of mutations found in cancer cells is point mutation. Point mutation may rise during DNA replication where a change of one nucleotide may lead to change of the amino acid codon. The location where the alteration occurs on the gene is important in determining the effects on the protein. If the mutation occurs in the region of the gene that is responsible for coding for the protein, a change in the sequence of amino acids may occur and can cause a change in the function of the protein. The vast majority of the alterations in noncoding regions are presumed to have a passive effect, although changes in regulatory sequences have the potential to alter function. Point mutations are classified based on their impact into two categories; non-synonymous or missense alterations are variants that alter the protein sequence. In some cases, these alterations affect the entire downstream protein product structure or function. The other type are silent or synonymous changes; here the change does not change the amino acid and so is unlikely to affect the function of the protein.

There are other types of mutations found in tumors and shown to contribute to the progression of the disease; frameshift mutations are where an insertion or a deletion of a single base pair occurs. As the gene is translated using triplet-based codons, an insertion or deletion can change the reading frame, resulting in a completely different translation from the original. This effect makes this type of mutations among the most deleterious alterations that can change the protein.

Most solid tumors also display widespread changes in chromosomes (aneuploidy), as well as deletions, inversions, translocations, and other genetic abnormalities (*28*). In cancer, translocations may occur as a result of the fusion of two genes to create an oncogene (such as *BCR-ABL* in chronic myelogenous leukemia) (*6*). In a small number of cases, it can also inactivate a tumor suppressor gene by truncating it or separating it from its promoter. As with point mutations, the majority of translocations appear to have no phenotypic effect (*6*). Homozygous deletions often involve just one or a few genes, and the target is mainly a tumor suppressor gene. Amplifications contain an oncogene whose protein product is abnormally active simply because the tumor cell contains 10 to 100 copies of the gene per cell, compared with the two copies present in normal cells. Studies to date indicate that there are roughly 10 times fewer genes affected by chromosomal changes than by point mutations (*6*).

1.2.2 Cancer driver mutations identification

Identifying cancer related genes and the consequences of genetic variations detected in tumors induce at the protein level could be vital to successfully allow for tailored and effective therapies (*29*). An example of such importance is the case of Vemurafenib, an inhibitor of V600E-mutant BRAF protein. The identification of the effect of V600E mutation on the BRAF protein and the role mutant BRAF protein plays in the majority of melanomas allowed the successful targeting of this protein (*29*). In the trial, the majority of patients with melanoma (49 patients) showed complete or partial tumor regression emphasising the potential of oncogene-targeted therapy for this disease (*29*).

Also, detecting cancer-associated genes and their mutations is key to pinpointing genetic aberrations implicated directly in causing cancer (drivers). Although a sole genetic change is almost never observed in malignant tumors, not all genetic changes found in cancer cells are directly responsible for tumor development (*30*). Many mutations are found to not be implicated directly in the initiation or the progression of the tumor and these mutations are often labelled as ‘passenger’ mutations (*30*). Therefore, continuous efforts in the last 3 decades of cancer research was dedicated to identifying the genes and mutations that drive carcinogenesis and distinguish them from the remaining mutations seem to confer no selective growth advantage.

The individuality nature of each tumor and the vast number of different mutations implicated within same tumor type makes any attempt to pinpoint these aberrations very difficult (*31*). A simple approach using the mutation rate of occurrence (the number of times the mutation was observed in tumor samples) was tested (*30*). However, this approach proved to be ineffective due to the high number of mutations shown to have a deleterious impact driving the tumor progression despite their very low frequency (i.e. false negatives) while others with high occurrence rate were shown to be playing a passive role (false positives) in the tumor initiation and progression (*30*).

Due to the heterogeneous nature of cancer, the mutation profile detected can vary from tumor to tumor even if they belong to the same cancer type. In some instances, the mutation profile of a tumor more closely matches to other tumors found in different organs compared to tumors from the same tissue based type (*32*). Mutation profiles can be clustered based on the locations and relevance of the mutation within a constructed gene network instead of the organ the tumor has originated from. These networks are based on coupling genes if they are reported to participate in the same biological process. This approach was shown to enhance the ability to predict driver genes and mutations implicated in cancer formation and progression (*30*). Also, combining cross-tumor data increased the size of the data sample studied and revealed shared oncogenic pathways across cancer sub-types in different organs and in some cases different roles the same family of genes plays depending on the organ (*33*).

The shared genomic aspects of cancer-associated mutations and the patterns of cross-cancer type data support the initiative of exploring shared characteristics that cancer-associated mutations exhibit. This, coupled with an understanding of their impact on the protein, could further our understanding of cancer and lead to better prediction of potential cancer-associated mutations yet to be identified.

In this thesis, I aim to highlight some of these shared characteristics across point mutations associated with cancer. I hypothesise that any distinctive patterns, if found, could highlight the specific biological processes crucial in the initiation and progression of many cancer

types. I aim to show that some of these characteristics could prove useful to train computational models to recognise cancer-related point mutations and therefore be used to identify novel cancer-associated point mutations.

1.2.2.1 Cancer genetic variations impact at the protein level

Proteomics has the potential to directly inform and affect cancer treatment. An example is the case of the inhibitors used in some new drugs and therapies of cancer (*34*). Inhibitors are molecules that stop or slow the proliferation of the cell by blocking specific proteins or disrupting its function. For instance, tyrosine kinase inhibitors (TKIs) block tyrosine kinase enzymes. These enzymes play a role in sending growth signals and obstructing them stops or slow cell growth (*34*). The MEROPS database (http://merops.sanger.ac.uk/) contains categorisation of inhibitors. Furthermore, mapping of protein-protein interactions can identify the relationship between proteins and should help us understand the full effects any new drug or intervention might have on linked processes in the cells (*35*).

Understanding the functional effect of a mutation on the gene product (i.e., the protein) is vital to the understanding of the pathological nature of a disease such as cancer. This is because variations in protein isoforms, protein quantity and structure are directly associated with disease phenotype and cannot often be predicted from genomics alone (*36*). Figure 1.1 demonstrates the effect mutation location has on the function of the protein. In this case, one amino acid change caused the loss of channel selectivity of the G protein-activated inward rectifier potassium channel. A somatic mutation in the oncogene *KCNJ5* results in the replacement of the amino acid Gly with amino acid Arg at position 151 of the protein sequence. This change happens to be at the mouth of the channel altering the channel selectivity causing the protein to be dysfunctional. This variant is associated with most adrenal gland cancers cases

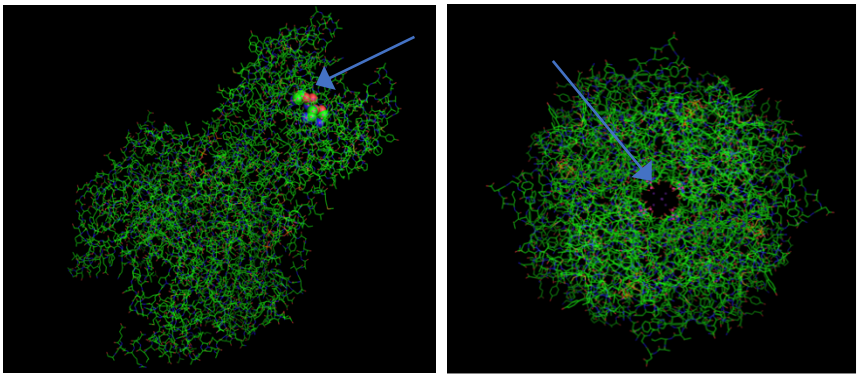

Figure 1.1. A 3D structure of G protein showing the location of mutation G151R

Another example is the constitutive activation of the PI3K pathway leading to increased cell proliferation and survival (one of main cancer hallmarks). The catalytic subunit p110α of PI3K protein coded by *PIK3CA* gene loses the ability to interact with its controlling substrate. This is caused directly by the E545K mutation occurring in the helical domain of the protein. This mutation causes the p110α to interact with a substrate IRS1, stabilizing it, resulting in constitutive activation of the PI3K pathway (*37*). Clearly both the location of the mutation and also its enablement to stabilise the protein were vital in this case.

Here, I expand the study of cancer point mutations impact on oncoprotein stability to other genes especially in cancer-associated genes that have been found frequently mutated in tumor samples. We also determine patterns where specific amino acid residues and replacements are enriched within cancer-associated replacements.

Changes caused by variants alerting specific biological functions seem to occur cross cancer types. For example, the *IDH1* gene with a single substitution, R132H, causes the protein encoded by this gene to alter its activity (*38*). The affinity of the mutated enzyme to bind is decreased substantially by this substitution. This alteration has been shown to promote gliomas and acute myeloid leukaemia (AML) cancers (*38*).

We link some of the enriched amino acid residue found in cancer-associated replacements to specific functions and describe the effects that could result from replacing certain amino acid with another.

The prediction of these replacements could potentially be significantly improved by studying the effect on protein function and properties. Each amino acid has distinct physico-chemical properties such that a replacement of an amino acid with another may lead to change in protein conformation and its function to different degrees. This would reveal the link between the genetic variations detected and the change in proteins functions leading to tumor development. A change of protein function can be manifested in multiple ways, for example, a change in how it binds with a given enzyme or in the amount of free energy of folding. The challenge here is to establish a way of scoring the functional change impact caused by each mutation.

For example, phosphorylation-related single nucleotide variants (pSNVs) that appear in most tumors show a greater functional effect than other protein-coding mutations (*39*). The authors developed a computational method (named ActiveDrive) and used it to identify functional sites in proteins (signalling sites, protein domains, regulatory motifs) that are specifically and significantly mutated in cancer genomes. It showed the ability to predict a subset of mutations (29%) from about 150 gene candidates with high pSNV recurrence due to their involvement in removing phosphorylation or changing kinase target sites to rewire signalling pathways. Thus, these mutations are predicted to be driver mutations with high confidence (*39*). The association with certain protein domains was also investigated, highlighting location persistent mutations. The findings showed that those replacements found to be enriched in specific regions or protein domains**.** 140 driver mutations have been identified and validated so far by studying the sustained locations that these replacements occur at (*6*).

Although several methods were developed to predict cancer-related replacements, we still do not have a model that can evaluate any amino acid replacement. We hypothesise that this could be reached if shared characteristics linked to the impact highly enriched replacements have on oncoprotein are used. Prime candidates are the physico-chemical properties of amino acids.
These properties characterise all amino acid residues and the change can be easily measured in related replacements.

In this thesis, I focus on understanding the change the cancer enriched single amino acid replacements introduce at the protein level. I investigated whether the measured change in physico-chemical properties alongside their position in the protein sequence could be used by machine learning technique to predict cancer-associated amino acid replacements.

1.2.2.2 Notable protein features

There are protein functions where a particular amino acid or amino acid group plays a vital role. These are of particular interest to study when identifying certain amino acids or amino acid groups to be highly enriched in cancer associated replacements. Below we list some of these features:

Stacking Functions: The biological function of proteins is often linked to interactions with their ligands and substrates. Knowledge of the molecular mechanisms of protein-ligand interactions, particularly in the spatial structure of the protein-ligand complex, can help understand the functional properties of proteins and their role in biochemical pathways in the living cell (*40*). Among all the various types of interactions in biomolecular complexes (such as hydrogen bonds, salt bridges, etc.), the stacking of aromatic substances plays a notable role in achieving successful binding to protein targets and many nucleoproteins use aromatic stacking to recognize binding site on DNA or RNA (*40*). Aromatic stacking is involved in the process of mismatch repair, strand separation, degradation and RNA cap binding (*41*). Aromatic stacking can be defined as interactions rising from the attractive force between the π-electron clouds in the aromatic groups where the stacking is attained between aromatic residues and the bases in the nucleotides (*41*) .

Disulphide Bonds: Disulphide bonds (also known as disulphide bridges) are type of covalent bonds (often between two cysteine amino acids) linking different components of a protein. They are known to stabilize the protein structure helping proteins fold and remain in their tertiary and quaternary shape (*42*). These bonds also play a key role in proteins and enzymes activities, in particular, stimulating cell proliferation through receptors regulating cellular growth (*43*).

Oxidative stress: refers to elevated intracellular levels of reactive oxygen generation rate. Oxidative stress is known to cause damage to lipids, proteins, and DNA (*44*). Oxidative protein modifications can cause partial unfolding of the protein thus, oxidation is shown to be a cause of altered protein functions (*44*). Both the unfolding and the direct oxidation of functional amino-acid side chains may lead to an impaired protein function. Among all the amino acids, Cysteine (Cys) is more prone to oxidation because of its high nucleophilic property (*45*) .Oxidative stress has been implicated in a number of human diseases (*46-49*) and cancer cells are known to produce elevated reactive oxygen generation rate (*50*).

1.3 Cancer-associated genes

According to Catalogue of Somatic Mutations database (COSMIC), approximately 3.5% of the 20 thousands human protein coding genes are identified to have mutations that drive the onset of the cancer, i.e. they are cancer driver genes (*51*). The COSMIC Cancer Gene Census (CGC) is an expert-curated list of the cancer-associated genes used by cancer researchers worldwide. The CGC (version 86, August 2018) listed 719 cancer-driving genes. More genes are being added to the CGC in every new release by verified contributions from researchers in this field. A gene is classified as an oncogene in CGC based on evidence that the activity of the gene product is related to cancer hallmarks and that the variants resulting in gain of function (loss of function for tumor suppressor genes) are observed in tumor samples. In the CGC (version 86, August 2018) the list of genes was divided into tiers where a second tier of genes (145 genes) is extracted from studies that show supportive but less detailed indications of a role in cancer. The extensive review and level of evidence required for adding a gene to CGC makes the CGC database one of the most trusted and accurate lists depicting genes related to cancer and is used as the prime source for both the genes and mutations in our study.

Cancer driver genes are often implicated in multiple cancer-related functions contributing to multiple key cancer hallmarks, making the functional annotation of each of these genes varied and complex (*51*). In turn, different mutations reported for each gene could affect different cellular processes related to cancer hallmarks. Moreover, not all cancer genes are

simply either oncogenes or tumor suppressor genes (TSGs). The genes *ATR* and *RB1* were shown to be involved in certain circumstances in tumor development in addition to their tumor suppressing function (*51*). This means that a simple identification of gene based on a single function or type of mutation found would not yield always an accurate result.

If we are able to identify genes related to cancer through a combination of features and computational methods, we could potentially provide an opportunity to enrich CGC with more candidates of genes that cancer researchers would consider in their work (possibly in a separate tier). Identifying cancer-associated genes using a machine learning based framework would expedite the rate at which cancer-related genes are being identified. Such a framework could also help prioritise targets for therapies and drug development without being limited by the lengthy and complex process of tumor samples acquiring and sequencing. Several studies have attempted to build models to identify human disease-related genes. Computational models built using sets of evolutionary and protein network based properties showed great potential and success in predicting disease genes in general (*52, 53*). We apply a similar approach that combine evolutionary and protein network-based properties to predict cancer related genes.

1.3.1 Important Oncogenes and Tumor Suppressor genes

There are hundreds of genes that are implicated in tumor initiation and progression. Here we provide a brief description for some key cancer associated genes.

PTEN: This gene codes for the phosphatase tumor suppressor protein PTEN (phosphatase and tensin homologue) that plays an important role in the PI3K and MAPK pathways (see section 1.1) and frequently altered in tumors (*54*). PTEN can bind to and activate or inactivate several nuclear proteins. Through these effects PTEN activates DNA repair and inhibits proliferation and thus acts as an important tumor suppressor protein (*54*).

BRAF: This gene encodes a protein belonging to the RAF family of serine/threonine protein kinases. This protein enables several functions, including protein binding activity; protein kinase activity; and scaffold protein binding activity. This protein plays a role in regulating

the MAP kinase/ERK signalling pathway, which affects cell division, differentiation, and secretion (see section 1.1). Mutations in this gene, most commonly the V600E mutation, are the most frequent cancer-causing mutations in melanoma, and have been identified in various other cancers as well (*29, 55*).

P53: This gene encodes a tumor suppressor protein comprising transcriptional activation and DNA binding domains. The encoded protein is involved in several processes, including intracellular signal transduction and regulation of apoptosis process, thus playing a vital role in inducing cell cycle arrest, apoptosis and DNA repair. Mutations in this gene are associated with a variety of human cancers (*11*).

KRAS: a Kirsten ras oncogene homolog from the mammalian ras gene family, encodes a protein that is a member of the small GTPase superfamily (*56*). The protein product of this gene is found in cytoplasm and plasma membrane of the cell and enables protein binding activity (see section 1.1). This gene is involved in regulation of cell population proliferation and regulation of metabolic process (*56*). A single amino acid substitution is responsible for an activating mutation. The mutated protein is considered a biomarker for carcinoma and is implicated in several diseases, including gastrointestinal cancer, glioma, and lung cancer (*57*).

JAK2: This gene encodes a non-receptor tyrosine kinase that plays a central role in growth factor signalling and enables several functions, including SH2 domain binding activity (*58*). Mutations in this gene are associated with numerous inflammatory diseases and malignancies. Disregulation of the JAK2/STAT3 signalling pathways produces increased cellular proliferation (see section 1.1). A nonsynonymous mutation in the pseudo kinase domain of this gene disrupts the domains inhibitory effect and results in constitutive tyrosine phosphorylation activity (*58*). This gene is implicated in several diseases, including lung non-small cell carcinoma and gastrointestinal cancer.

EGFR: The protein encoded by this gene is a transmembrane glycoprotein that is a member of the protein kinase superfamily (*59*). This protein is a receptor for members of the epidermal growth factor family. EGFR is a cell surface protein that binds to epidermal

growth factor, thus inducing receptor dimerization and tyrosine autophosphorylation leading to cell proliferation (*59*). Mutations in this gene are associated with several diseases, including colorectal cancer, pancreatic cancer, and prostate cancer (*26, 34, 59*).

1.3.2 Prediction of cancer driver genes

Properties used in computational models to predict cancer driver genes can be grouped into three main categories: mutational frequency, network-based (specially protein–protein interaction) and function-based groups (*60*). The frequency-based approach identifies candidate driver genes based on the assumption that their mutation rates are higher than the background mutation rate (BMR) found across tumor samples. This assumption is not accurate for all cancer driver genes; some of the verified genes implicated in cancer formation have a low mutation rate (*61*). Additionally, precise background mutation frequency evaluation is not always possible. As for the function-based approach, not all functions of all genes are known to same level of detail. Also, the lack of a suitable numeric basis, which is often necessary for accurate computational models, hinders this approach further. This problem was evident when reviewing methods developed to predict cancer driver genes to date. Network-based algorithms preform the best overall on average (*60*) and the HotNet2 (*62*) method using protein–protein interaction network properties (PPI) exhibited the best overall performance in this category. Using the *ROC curve*, the accuracy of prediction can be evaluated through AUC (area under the curve) with the larger the area under the curve, the more accurate the model is. An AUC of 0.5 suggests that predictions based on this model are no better than a random guess and the closer AUC to 1.0 the stronger the model is. HotNet2 was reported to achieve AUC = 0.81 and the average AUC achieved by all other Network-based algorithms were 0.77. A notable exception to the low performance of frequency-based methods was the driverMaps algorithm (*63*). Despite being a frequency based algorithm, driverMaps AUC was reported to reach 0.94 (still *sensitivity* was much higher in HotNet2 compared to driverMaps) while the average AUC for the rest of frequency based methods was lower than 0.55.

Method	**Approach**	**AUC**
driverMaps	Frequency based	0.94
HotNet2	Network based	0.81
MutPanning	Function based	0.62

Table 1.1 The top cancer driver genes prediction methods in each approach

Other factors could be influencing the model's performance in addition to the data used to train the models. Sample size and machine learning techniques would have an impact and should be evaluated. However, reviewing the change in sample size on the performance for available methods showed little impact (*60*).

Despite numerous attempts, we still do not have one central model that can be reliably used to predict cancer-related genes (*60*). More than 12 different models were developed using different approaches with on average only low to medium level of accuracy attained (*60*). The poor performance could be attributed to the smaller data sample available in the past. However, for more recent attempts this cannot be the case and it is likely due to the actual data used to train the models. Two notable exceptions where the prediction models showed excellent performance (Table 1.1) were driverMaps and HotNet2. Both model reported achieving AUC > 0.8 (*60*). Part of the challenge is then to identify the distinctive features of cancer-related genes required to achieve this goal. Properties associated with PPI networks are the most successful to date (*60*).

In this thesis, I contribute to these on-going efforts by illustrating the influence of some genes properties in predicting cancer association. I show that identifying properties that have a distinctive characteristic in cancer-associated genes could result in superior model performance.

1.4 Cancer as an evolutionary process

As Theodosius Dobzhansky put it: "Nothing in biology makes sense except in the light of evolution" and evolution can indeed answer many questions about cancer (*64*) such as: why do humans get cancer? why not at higher or lower rate? Somatic cell selection, the evolution of tumor suppressing genes and the limitations imposed on these genes are all evolutionary concepts that can shed more light and help answering the above questions (*64*). The interest in the evolutionary mechanism of cancer has increased since it became increasingly clearer that the approach of trying to find a global genetic variation pattern for

each cancer type is not the answer as it was thought it would be when only few common mutations were known to cause cancer (*65*).

Cancer has the characteristics of typical evolutionary processes (*65*). Most cancers are thought to arise from a single mutant precursor cell. As that cell divides, the resulting 'daughter' cells may acquire different mutations over time and exhibit different behaviours from the originals cell and from their sister cells. Those cells that gain an advantage in division or resistance to cell death will tend to take over the population and become eventually cancerous cells. For instance, genetic variation can influence differential fitness (survival and growth) of cancer cells which are known in turn for their clonal expansion and their competitiveness for space and nutrition resources. Thus, understanding cell evolution and how it underlines cancer's initiation and progression provides great prospect for studying cancer (*65*). Concepts that are currently the main focus of studying cancer in the context of evolution include the evolving of malignant cells via 'somatic cell selection', the influence different ecological factors within the cell microenvironment have on the growth and apoptosis of these cells and the tumor supressing mechanism, which is found to be the result of early evolution dating back to the emergence of the multicellular organisms. Tumor suppression in human has limits and differences compared to other species indicating the existence of different trade off model across species (*64*).

Another link to evolution is the association of some cancer genes and evolutionary events. For example, in humans the whole genome is thought to have been duplicated in two events some 500 MY ago (*66*). The association with genes retained from Whole Genome Duplication (WGD) events suggest that evolution was strong factor in contributing towards the emergence of specific class of cancers (*66*).

1.4.1 Somatic Cell Selection

The heterogeneity of tumors forms the base on which somatic selection works. Despite not all the mutations found in tumors being critical for the progression of the tumor, the number of mutations is important and is linked to the progression of cancer and may even influence the consequences of treatments (*64*). The faster progression of cancer and higher

aggressiveness have been linked to heterogeneity (*67*). Also, a high mutation rate is a strong indication of possible early relapses following treatments. Chemotherapy not only allows resistant cells to be prevalent but also provides space for them to form colonies (*68*). This understanding might allow for customization of treatment based on the rate of heterogeneity observed and possibly lead to higher survival rate.

1.4.2 Ecology theory

The microenvironment of a cell may influence tumor suppression or progression (*69*). An association was reported between cancer risk and the behavioural changes observed in neighbouring cells. These cells were showed to play a role in providing cancer cells with growth signals and fitness enhancing factors (*69*). Understanding how to control the ecological factors surrounding tumor could help greatly in treating and preventing cancer. The influence of the microenvironment could be controlled to also slow or stop malignant cells progression by inhibiting the growth (*70*).

1.4.3 Tumor Suppression

Evolution not only helps explain how cancer forms but provides the means to understand how the organism evolved to protect itself. Tumor suppression could have evolved as a co-opting a mechanism where termination of a replicated cell is necessary due to inefficient in environment resources (*64*).

There are several processes identified by researchers that play a role in preventing cancer formation. Processes such as DNA repair, cell cycle checkpoint, apoptosis in addition to particular tissues architectures that can control proliferation (*64*). Also, there are further mechanisms known to halt cancer progression to the metastatic phase such as cell cycle arrest, cell adhesion, asymmetric cell division and apoptosis. All these mechanisms are the products of evolutionary processes and the selective advantage delivered to the organism (*64*).

Cancer is not always supressed and there are several theories that explain this from the evolutionary perspective. One theory argues that the constraints on selection such as path-

dependence in conserved cell cycles might contribute to vulnerability to cancer (*64*). Another theory highlights that cancer cells are derived from normal cells, which limit the immune system effectiveness and increase cancer susceptibility. For example, maintaining cell division capabilities such as tissue repair, which involves the abilities to proliferate and rapid generation of blood vessels is crucial for most cells (*71*). Other examples such as the necessity for cells to be able to invade tissues during gastrulation, and the fast growth rate at the age of sexual maturation could be linked to elevated risk of early mutation linked to cancer formation later in life (*72*).

1.4.4 The impact of cancer evolutionary characteristics on research and therapies

The study of evolutionary aspects of cancer biology is influencing research into drug discovery and in particular, the acquired therapeutic resistance observed in many cancer treatments (*67*). Understanding cancer formation and progression in the context of evolution could impact treatment, management, and prevention of cancer and further our understanding of this disease (*65*). Moreover, these evolutionary aspects may help in the identification of genes implicated in cancer. However, using evolutionary related properties to train machine learning model predicting cancer genes is not yet fully explored (*60*).

In this thesis, I combine various properties of cancer genes including evolutionary related measures to train a machine-learning model to identify cancer related genes. We aim to use measures indicating selection pressure in relation to genetic variants. Also, I highlight any correlation with evolutionary events such as whole genome duplication. We hypothesise that these measures would show important impacts on predicting genes involved in cancer initiation and progression.

1.5 Computational tools and Cancer research

Bioinformatics in cancer research is playing a significant role in advancing our understanding. This role is expected to grow as a multitude of bioinformatics tools are developed and used for this purpose by researchers (*73*). The databases and computational tools available to support the efforts in analysing the constantly increasing large genomic

data sets are providing advantageous contributions. For instance, it is possible to perform predictive analysis and build machine-learning models attaining some key discoveries such as pattern finding, clustering and interpretation of research results. The recent acceleration in acquiring genome data and the high-throughput technologies available to store and share this data will drive in turn an acceleration of collaboration between researchers worldwide and that in turn is expected to have a great impact on advancing cancer research (*74*). An example of this is the aforementioned COSMIC database that provides researchers access to manually curated and verified results implicating genes and mutations in cancer initiation and progression (*51*). Furthermore, bioinformatics tools could facilitate and contribute effectively to challenges faced by researchers allowing precise evaluation of the margin of errors in the results and better analysis of experiments outcomes (*74*).

Machine learning techniques are showing great potential in supporting cancer research enabling scientists to not only predict the most likely targets to investigate but also to understand the influence of each feature that is used to train the model on that prediction. Various open-source algorithms can be used to build classification and regression type models, most of which are accessible and easy to implement. Several bioinformatic tools allow the building of several models using different techniques and configurations that can then be compared. Performance metrics and validation methods are available to ensure that the models are reliable, and their performance would be sustained for new data as it is acquired (*60*). Moreover, machine-learning methods show more flexibility than traditional statistical methods because they rely on fewer statistical assumptions. For instance, ordinary least squares regression requires that the Gauss Markov assumptions (list of conditions the modelled data should adhere to) be supported, to ensure that the model is unbiased and efficient. Traditional statistical regression techniques rely on formal hypothesis testing for variable significance and feature selection (e.g., t-test, p-value, standard error). These statistical tests tend to have assumptions about distribution shape and independence that may not be supported by the data. Machine learning methods, on the other hand, are more flexible in defining the model structure, which typically results in better model performance (*75*). Machine learning includes methods that do not need to have formal hypothesis testing nor distributional assumptions to demonstrate model validity.

In this thesis, I perform a parallel heuristic search for the best model or ensemble of models from a repository of open-source models of different types (e.g., Random Forests, Neural Networks, etc) ensuring that the selected model is not arbitrary.

Machine-learning methods are not free of challenges. Several pitfalls should be avoided when using these techniques in particular within genomic-based research due to the complexity and interconnectivity of biological data (*76*). Several guards should be implemented to lessen the chance of making one of these mistakes. One of the main important issues is the possibility of 'leakage' where a feature used in the training data would not be fully formed until the outcome has occurred. For instance, predicting whether the gene is cancer-associated using the number of tumor samples as one of the training data inputs may cause an information leak if -by definition- the gene found mutated in tumor sample is a cancer-associated gene. This could create a false level of accuracy and such correlation should be detected and eliminated before the model is built. Another pitfall is the unaccounted dependency within the examples, creating false enhanced performance. For instance, predicting protein functions from a protein interaction database that links two proteins if they share a functional annotation. Another common problem is the existence of confounder variables used to train the model, creating a false link between some of the features and the outcome. Understanding the data used to build the model is vital to avoid such problems. Using the model prediction in different conditions to the ones the model was trained or tested under is also a common mistake (*76*). Here, we ensure that several guardrails are implemented when we create our models. I implement a leakage detection method and use cross fold validation minimising the chance of model over fitting. Also, separate testing dataset extracted under the same conditions with similar distribution to the training and validation sets is used ensuring appropriate evaluation of the model performance. Only models that have the same level of performance across training, validation and testing were selected.

1.6 Thesis Outline

Many mutations have been collated and described in COSMIC to be cancer related. The data showed that different variants in different genes and in some occasions even within the

same gene (*51*) may have distinct effects on oncogenesis. Observations of this nature raise several questions: are there any shared characteristics between the cancer-associated mutations? Can certain amino acid residues show higher than expected enrichment? Can the affinity of cancer-associated genes to certain biological processes such as 'binding' explain these enrichments? Are there certain properties that can be used to predict cancer-associated genes accurately? Could a machine-learning model be implemented to discover not just for genes but also in mutations the degree certain characteristics play a role in carcinogenesis. Despite the importance of manual curation to describe how these alterations affect the physiological processes that drive cancer, bioinformatic approaches are crucial to support the analysis of vast amount of information available to date. Utilising the CGC, COSMIC data *in silico* methods may prove to be the best way to discover the genetic fingerprints across cancer, discover new targets as well as to using these patterns in guiding future therapeutic research.

The success of prior studies in predicting human genetic disease-associated genes using genes' properties prompted us to replicate the approach when identifying cancer-associated genes candidates. We investigate the distinct characteristics that cancer-associated genes might exhibit. We then determine whether if these characteristics could be used to predict further potential gene candidates to be directly implicated in cancer. We also investigate whether certain amino acid residues and replacements are enriched in cancer and whether a change in the physico-chemical properties caused by the replacements are indicative of their cancer-associated involvement.

As understanding the biological foundations of cancer is crucial for development of new diagnostic and therapeutic measures, knowledge of why certain mutations cause dysfunction of the proteins and how this can drive the functional hallmarks of cancer is essential. Identifying the distinctive characteristics of cancer-associated mutations and their genes using the vast amount of data already collated would allow us to further understand the diseases and use machine learning techniques to identify cancer-associated genes and mutations implicated in cancer accurately and in practical-timed manner. Cancer is a complex disease, and it cannot be fully described by the mutations and genes driving the tumor formation as other factors such as the microenvironment play a role in the

development of tumors. However, the cancer-associated genes and their related mutations remain the primary causative factors in cancer initiation and progression and their identification remains a key and not yet fully resolved challenge.

The thesis is organised into five chapters. Outlines of these chapters are given below:

Chapter 1: introduces the background of this research, discuss different prior studies and provide an overview of this research.

Chapter 2: Here I study the impact variants in most mutated cancer associated genes have on the stability of the protein. This is done using the free folding energy and the comparison between the half-life of the protein in wild type and the constraint on number of variants in the gene. It shows the distinct impact (mainly neutral or stabilizing) these variants have on stability compared to non-cancer deleterious mutations.

Chapter 3: I compare the enrichment of amino acids in cancer-associated replacements to the expected frequency based on the genetic code, to Blosum62 (*77*) and to other missense non-cancer mutations. It also links certain biological processes known to be cancer-related to these found enriched replacement residues (Cys and the aromatic amino acid group). Finally, it analyses the impact of these replacement on physic-chemical properties of the residues highlighting that hydrophobicity related properties and polarity are most affected. These changes were used to train a machine-learning model that can identify cancer-associated replacements.

Chapter 4: I identify the association between cancer genes and essentiality of the gene. It then shows how using essentiality related properties combined with protein-protein interaction network and evolutionary properties can be used to train a machine-learning model with high accuracy that is able to predict cancer-associated genes.

Chapter 5: This chapter aims to bring together all the results from previous chapters and discuss their implications and how they can be utilised in the cancer research field, highlighting any limitation.

1.7 References

1. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature* **458**, 719-724 (2009).
2. D. Hanahan, Robert A. Weinberg, Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674 (2011).
3. T. Gutschner, S. Diederichs, The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biology* **9**, 703-719 (2012).
4. S. Elmore, Apoptosis: A Review of Programmed Cell Death. *Toxicologic pathology* **35**, 495-516 (2007).
5. S. E. Artandi, R. A. DePinho, Telomeres and telomerase in cancer. *Carcinogenesis* **31**, 9-18 (2010).
6. B. Vogelstein *et al.*, Cancer Genome Landscapes. *Science* **339**, 1546-1558 (2013).
7. L. W. Elmore *et al.*, Blueprint for cancer research: Critical gaps and opportunities. *CA: A Cancer Journal for Clinicians* **71**, 107-139 (2021).
8. H. Y. K. Yip, A. Papa, Signaling Pathways in Cancer: Therapeutic Targets, Combinatorial Treatments, and New Developments. *Cells* **10**, (2021).
9. N. Perrimon, C. Pitsouli, B. Z. Shilo, Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb Perspect Biol* **4**, a005975 (2012).
10. R. Sever, J. S. Brugge, Signal transduction in cancer. *Cold Spring Harb Perspect Med* **5**, (2015).
11. F. Sanchez-Vega *et al.*, Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e310 (2018).
12. R. Nusse, Wnt signaling. *Cold Spring Harb Perspect Biol* **4**, (2012).
13. R. Kopan, Notch signaling. *Cold Spring Harb Perspect Biol* **4**, (2012).
14. P. W. Ingham, Hedgehog signaling. *Cold Spring Harb Perspect Biol* **4**, (2012).
15. Z. Lei *et al.*, Control of Angiogenesis via a VHL/miR-212/132 Axis. *Cells* **9**, (2020).
16. K. Zenonos, K. Kyprianou, RAS signaling pathways, mutations and their role in colorectal cancer. *World J Gastrointest Oncol* **5**, 97-101 (2013).
17. L. Li *et al.*, The Ras/Raf/MEK/ERK signaling pathway and its role in the occurrence and development of HCC. *Oncol Lett* **12**, 3045-3050 (2016).
18. D. K. Morrison, MAP kinase pathways. *Cold Spring Harb Perspect Biol* **4**, (2012).
19. B. A. Hemmings, D. F. Restuccia, PI3K-PKB/Akt pathway. *Cold Spring Harb Perspect Biol* **4**, a011189 (2012).
20. A. Korkut *et al.*, A Pan-Cancer Analysis Reveals High-Frequency Genetic Alterations in Mediators of Signaling by the TGF-β Superfamily. *Cell Systems* **7**, 422-437.e427 (2018).
21. D. A. Harrison, The Jak/STAT pathway. *Cold Spring Harb Perspect Biol* **4**, (2012).
22. X. Hu, J. li, M. Fu, X. Zhao, W. Wang, The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduction and Targeted Therapy* **6**, 402 (2021).
23. H. M. Shen, V. Tergaonkar, NFkappaB signaling in carcinogenesis and as a potential molecular target for cancer therapy. *Apoptosis* **14**, 348-363 (2009).
24. S. M. Wuerzberger-Davis *et al.*, Nuclear export of the NF-κB inhibitor IκBα is required for proper B cell and secondary lymphoid tissue formation. *Immunity* **34**, 188-200 (2011).

25. N. Rhind, P. Russell, Signaling pathways that regulate cell division. *Cold Spring Harb Perspect Biol* **4**, (2012).
26. B. Zhao *et al.*, Mechanisms of resistance to anti-EGFR therapy in colorectal cancer. *Oncotarget; Vol 8, No 3*, (2016).
27. C. M. Croce, Oncogenes and cancer. *N Engl J Med* **358**, 502-511 (2008).
28. M. Grade, M. J. Difilippantonio, J. Camps, Patterns of Chromosomal Aberrations in Solid Tumors. *Recent Results Cancer Res* **200**, 115-142 (2015).
29. K. T. Flaherty *et al.*, Inhibition of Mutated, Activated BRAF in Metastatic Melanoma. *New England Journal of Medicine* **363**, 809-819 (2010).
30. D. Tamborero *et al.*, Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Nature* **3**, 2650 (2013).
31. I. Martincorena, P. J. Campbell, Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489 (2015).
32. J. N. Weinstein *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nature* **45**, 1113-1120 (2013).
33. M. Hofree, J. P. Shen, H. Carter, A. Gross, T. Ideker, Network-based stratification of tumor mutations. *Nature* **10**, 1108-1115 (2013).
34. M. E. M. Noble, J. A. Endicott, L. N. Johnson, Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **303**, 1800-1805 (2004).
35. J. A. Alfaro, A. Sinha, T. Kislinger, P. C. Boutros, Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature* **11**, 1107-1113 (2014).
36. Q. Wang *et al.*, Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences* **108**, 2444-2449 (2011).
37. Y. Hao *et al.*, Gain of Interaction with IRS1 by p110α-Helical Domain Mutants Is Crucial for Their Oncogenic Functions. *Cancer Cell* **23**, 583-593 (2013).
38. L. M. Gagné, K. Boulay, I. Topisirovic, M.-É. Huot, F. A. Mallette, Oncogenic Activities of IDH1/2 Mutations: From Epigenetics to Cellular Signaling. *Trends in Cell Biology* **27**, 738-752 (2017).
39. J. Reimand, O. Wagih, G. D. Bader, The mutational landscape of phosphorylation signaling in cancer. *Nature* **3**, 2651 (2013).
40. T. V. Pyrkov, D. V. Pyrkova, E. D. Balitskaya, R. G. Efremov, The role of stacking interactions in complexes of proteins with adenine and Guanine fragments of ligands. *Acta Naturae* **1**, 124-127 (2009).
41. M. M. Rahman, Z. T. Muhseen, M. Junaid, H. Zhang, The aromatic stacking interactions between proteins and their macromolecular ligands. *Curr Protein Pept Sci* **16**, 502-512 (2015).
42. M. E. Ortiz-Soto, S. Reising, A. Schlosser, J. Seibel, Structural and functional role of disulphide bonds and substrate binding residues of the human beta-galactoside alpha-2,3-sialyltransferase 1 (hST3Gal1). *Scientific Reports* **9**, 17993 (2019).
43. R. Mor-Cohen *et al.*, Unique Disulfide Bonds in Epidermal Growth Factor (EGF) Domains of β3 Affect Structure and Function of αIIbβ3 and αvβ3 Integrins in Different Manner. *Journal of Biological Chemistry* **287**, 8879-8891 (2012).
44. V. Cecarini *et al.*, Protein oxidation and cellular homeostasis: Emphasis on metabolism. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1773**, 93-104 (2007).
45. S. Ahmad *et al.*, Protein oxidation: an overview of metabolism of sulphur containing amino acid, cysteine. *Front Biosci (Schol Ed)* **9**, 71-87 (2017).
46. C. E. Cross *et al.*, Oxygen radicals and human disease. *Ann Intern Med* **107**, 526-545 (1987).

47. V. J. Thannickal, B. L. Fanburg, Reactive oxygen species in cell signaling. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **279**, L1005-L1028 (2000).
48. G.-Y. Liou, P. Storz, Reactive oxygen species in cancer. *Free radical research* **44**, 479-496 (2010).
49. B. Halliwell, J. M. Gutteridge, C. E. Cross, Free radicals, antioxidants, and human disease: where are we now? *J Lab Clin Med* **119**, 598-620 (1992).
50. M. Schieber, N. S. Chandel, ROS function in redox signaling and oxidative stress. *Curr Biol* **24**, R453-462 (2014).
51. Z. Sondka *et al.*, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
52. N. López-Bigas, C. A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108-3114 (2004).
53. N. Spataro, J. A. Rodríguez, A. Navarro, E. Bosch, Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum Mol Genet* **26**, 489-500 (2017).
54. M. S. Song, L. Salmena, P. P. Pandolfi, The functions and regulation of the PTEN tumour suppressor. *Nature Reviews Molecular Cell Biology* **13**, 283-296 (2012).
55. P. A. Ascierto *et al.*, in *J Transl Med*. (England, 2012), vol. 10, pp. 85.
56. A. Ferreira *et al.*, Crucial Role of Oncogenic KRAS Mutations in Apoptosis and Autophagy Regulation: Therapeutic Implications. *Cells* **11**, (2022).
57. L. Huang, Z. Guo, F. Wang, L. Fu, KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy* **6**, 386 (2021).
58. K. Gnanasambandan, P. P. Sayeski, A structure-function perspective of Jak2 mutations and implications for alternate drug design strategies: the road not taken. *Curr Med Chem* **18**, 4659-4673 (2011).
59. P. Wee, Z. Wang, Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers (Basel)* **9**, (2017).
60. X. Shi *et al.*, Comprehensive evaluation of computational methods for predicting cancer driver genes. *Briefings in Bioinformatics* **23**, bbab548 (2022).
61. M. S. Lawrence *et al.*, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
62. M. D. M. Leiserson *et al.*, Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**, 106-114 (2015).
63. S. Zhao *et al.*, Detailed modeling of positive selection improves detection of cancer driver genes. *Nature Communications* **10**, 3399 (2019).
64. C. A. Aktipis, R. M. Nesse, Evolutionary foundations for cancer biology. *Evolutionary Applications* **6**, 144-159 (2013).
65. H. H. Q. Heng *et al.*, The evolutionary mechanism of cancer. *Journal of Cellular Biochemistry* **109**, 1072-1084 (2010).
66. T. I. Zack *et al.*, Pan-cancer patterns of somatic copy number alteration. *Nature* **45**, 1134-1140 (2013).
67. L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley, Cancer as an evolutionary and ecological process. *Nature* **6**, 924-935 (2006).
68. A. Lee, C. Swanton, Tumour heterogeneity and drug resistance: Personalising cancer medicine through functional genomics. *Biochemical pharmacology* **83**, 1013-1020 (2011).

69. K. O. Alfarouk, Tumor metabolism, cancer cell transporters, and microenvironmental resistance. *Journal of Enzyme Inhibition and Medicinal Chemistry* **31**, 859-866 (2016).
70. M. J. Bissell, W. C. Hines, Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nature* **17**, 320-329 (2011).
71. S. Guo, L. A. DiPietro, Factors Affecting Wound Healing. *Journal of Dental Research* **89**, 219-229 (2010).
72. S. D. Tyner *et al.*, p53 mutant mice that display early ageing-associated phenotypes. *Nature* **415**, 45-53 (2002).
73. D. Kihara, Y. D. Yang, T. Hawkins, Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. *Cancer Informatics* **2**, 25-35 (2006).
74. G. Gómez-López, A. Valencia, Bioinformatics and cancer research: building bridges for translational research. *Clinical and Translational Oncology* **10**, 85-95 (2008).
75. J. A. Cruz, D. S. Wishart, Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics* **2**, 117693510600200030 (2006).
76. S. Whalen, J. Schreiber, W. S. Noble, K. S. Pollard, Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics* **23**, 169-181 (2022).
77. S. R. Eddy, Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **22**, 1035-1036 (2004).

2. Chapter 2: Cancer-associated missense point mutations exhibit dissimilar impact on protein stability to other deleterious mutations

*** Statement of Authorship**

All the work presented in this chapter including analysis and method development were carried out by Amro Safadi with supervision of his supervisors prof. Simon C. Lovell. and prof. Andrew J. Doig. Both supervisors approved the figures and final wording of the manuscript providing corrections when needed.

2.1 Introduction

Nonsynonymous single nucleotide polymorphisms (SNPs) cause an alteration to amino acids within the protein sequence potentially leading to important changes in protein properties such as structure, solubility, and stability. These changes could result in functional changes (*1*). Stability is an important property for biological research. Understanding the stability of the protein and factors that might impact it is vital in evaluating protein functional changes. For instance, protein destabilization caused by deleterious mutations is known to be a primary factor in many Mendelian diseases (*1*).

Stability indicates the survival of protein over time and thus can be analysed by studying properties such as the change in the protein energy of folding and protein half-life. Any mutation that adds energy to the folded state is likely to destabilize the structure and make the protein more likely to be in its unfolded form, this was found to be a feature in some diseases (*2*). If no or very small change in stability is produced by the mutant, then the mutation is neutral. If the change produced by the mutation was negative, then the mutation is considered to have stabilising effect on the protein. The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis. It is used alongside the change in folding energy as an indication of the stability of the protein (*3*).

In some cases, if a destabilising mutation occurs, the protein may compensate and stay in its optimal stability zone by "recruiting" other mutations (*4*). Both types of mutations, functional and compensatory, are more likely to be fixed in the genome than neutral

mutations. This is known as cryptic epistasis in molecular evolution (*5*). It means that several mutations may need to occur to achieve a particular effect (e.g., positive effect in the case of functional adaptation or negative in the case of disease).

A study of 20 proteins affected by deleterious mutations in different diseases found that 18 were shown to be destabilized by diseased associated amino acid replacements (*6*). These findings pointed to destabilization of protein structure to be key pathogenicity factor caused by missense mutations, even in cases where reduced protein stability was not a trait associated with the nature of the disease itself. However, the impact caused by cancer mutations on the corresponding proteins and in particular their stability is still not yet fully understood. Understanding the impact of cancer nonsynonymous SNPs on protein stability could further inform us about the mechanisms involved in tumor initiation and progression.

Here we analysed the impact missense point mutations found in oncogenes have on protein stability and highlighted how the effects differ from other non-cancer missense mutations. The sample of oncogenes analysed were selected due to their relative importance as reported in literature and the number of mutations reported per oncogene across all cancer types. We also investigated whether epistasis in the same gene in relation to stability by missense point mutations is likely in oncogenes.

2.2 Materials and Methods

We obtained the curated list of mutations reported to be involved in cancer formation and progression. These were downloaded from COSMIC (*7*) release v77 (http://cancer.sanger.ac.uk/census). We used the UNIPROT (https://www.uniprot.org/) database as the source for all protein sequences FASTA files analysed and corresponding proteins 'solved' 3D structures. Files used for stability calculations and the 3D structures of studied proteins were extracted from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb).

The software Tableau (http://www.tableau.com) was used for the purpose of the integration and the analysis of the data. The Tableau software provides several features that allow data to be imported and integrated from multiple data sources and provides

advantages in terms of automatic data updates. Also, it facilitates the visualisation and configuration of data in a simple way.

Large-scale reference data sets of human genetic variation are required to act as the control group and background for comparison against the cancer mutations in this project. For that we utilised the Exome Aggregation Consortium (ExAC), which is a collation of exome sequencing data from a wide variety of large-scale sequencing projects, and population genetic studies that spans 60,706 unrelated individuals (*8*).

Proteins are generally found in two main states: folded (functional) and unfolded. Protein folding is a process by which proteins are folded into their biochemically functional three-dimensional structure. Unfolded and folded state, both can be depicted by their Gibbs free energy (G). To transfer from one state to the other, there needs to be a transfer of energy. The free energy of folding (ΔG) is defined as difference of the Gibbs free energies between the denatured and the native state with the unfolded being of higher energy than the folded state. Comparing the folding rate of wild type protein and the mutant form gives an indication of the stability. The difference between ΔG for the wild type and ΔG for the mutated protein is ΔΔG (kcal mol¯1).

$$\Delta\Delta G = \Delta G\,(wt) - \Delta G\,(mutant)$$

To calculate the stability of the protein we used FoldX software (*9*). FoldX calculates both ΔG (Gibbs free energy) of the wild type structure and the ΔΔG for the mutated variant of the protein. However, this is done one mutation at a time. We recognised the need to speed and partially automate the process of calculating the stability (folding energy) for the proteins with multiple variants using the FoldX software. We coded a program (calling it SpeedUp) that automatically runs FoldX on a provided list of mutations (rather than manually run it on each individual mutation at a time). This minimized the manual effort needed. The program first converts the missense mutations file (exported from ExAC) for each gene into format recognised by FoldX and then extract the chain id and first and last position on the amino acid sequence in the selected PDB file. Finally, the FoldX stability calculation is performed for each mutation on the list. The program does not calculate the stability for any mutation at positions outside the range covered by the PDB file and/or

when the ‘mutated to’ amino acid does not match the amino acid at that specific position on the file.

In general, stability effect can be interpreted using ΔΔG as:
If ΔΔG >1, then the mutation in question is of a destabilising effect.
If ΔΔG <-1, then the mutation in question is of a stabilising effect.
If 1> ΔΔG >-1, then the mutation in question is neutral.
The threshold for the stabilising effect might differ from protein to protein but a mutation is considered to have a significant destabilising effect if ΔΔG is >1 kcal/mol (*9*).

Unlike our ability to calculate the difference in the energy of folding between the wild and mutated type of the protein, half-life of the mutated proteins cannot be always calculated using computational methods. Instead, we carried out the comparison between the half-life of the wild-type protein (*in vivo*) and the ExAC Z constraint score. The ExAC Z constraint is a score that shows the deviation of observed number of variants from the expected number of variants. High Z scores indicate increased constraint (intolerance to variation) that is when the gene had fewer variants than expected (*8*). This comparison would indicate whether intolerance to variations were associated with shorter half-life (destabilisation effect).

To calculate the half-life of the protein wild type we used the tool ProtParam (*10*) (http://web.expasy.org/protparam/protparam-doc.html).
ProtParam allows the estimation of half-life of protein and relies on the "N-end rule", which relates the half-life of a protein to the identity of its N-terminal residue (the first part of the protein that exits the ribosome during protein biosynthesis). Experiments showed that the identity of the N-terminal residue of a protein plays an important role in determining its stability *in vivo* (*3*).

We integrated the data sources (UNIPROT, PDB, COSMIC, etc.) used in our work to obtain cancer mutations, protein sequences and protein structures data in Tableau to allow the analysis to be carried out (figure 2.1). The data was extracted from UNIPROT and PDB Using

the URL query function. Only proteins with resolutions <=3.0 Å was obtained ensuring acceptable accuracy.

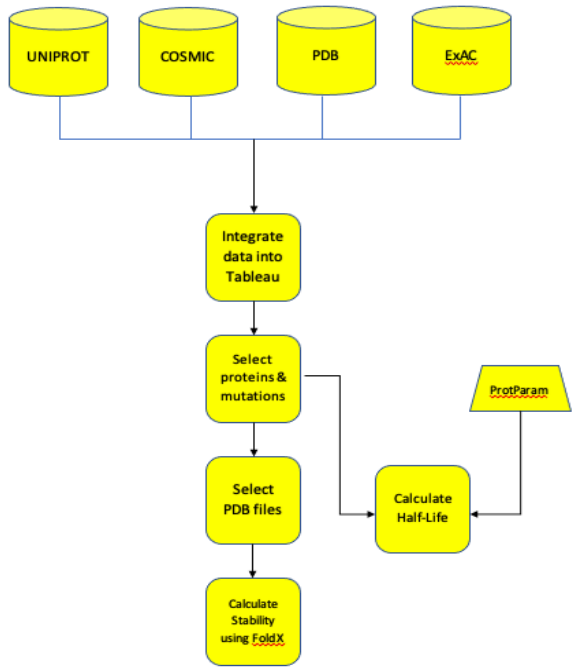

Figure 2.1. The outline of the data sources and methods used in this chapter

2.3 Results and Discussion

Due to the large number of missense SNPs and genes associated with cancer, we focused the analysis on the highly mutated genes. Tableau allowed us the creation of different views of the data based on number of samples, mutation types, etc. Therefore, we used these views to identify these frequently mutated genes. We first produced a view showing the most frequently mutated genes found in tumor samples as per COSMIC Sep-2016. In figure 2.2, mutations found in over 500 samples are coloured blue. Each mutation is labelled using the gene name, type of mutation and number of samples that this mutation was found in. The data showed specific cancer-associated genes to have significantly large number of mutations in various tumor samples. However, this could be driven by research focus on certain genes and certain cancer types. The results do not necessary reflect the unique number of cancer mutations found per gene.

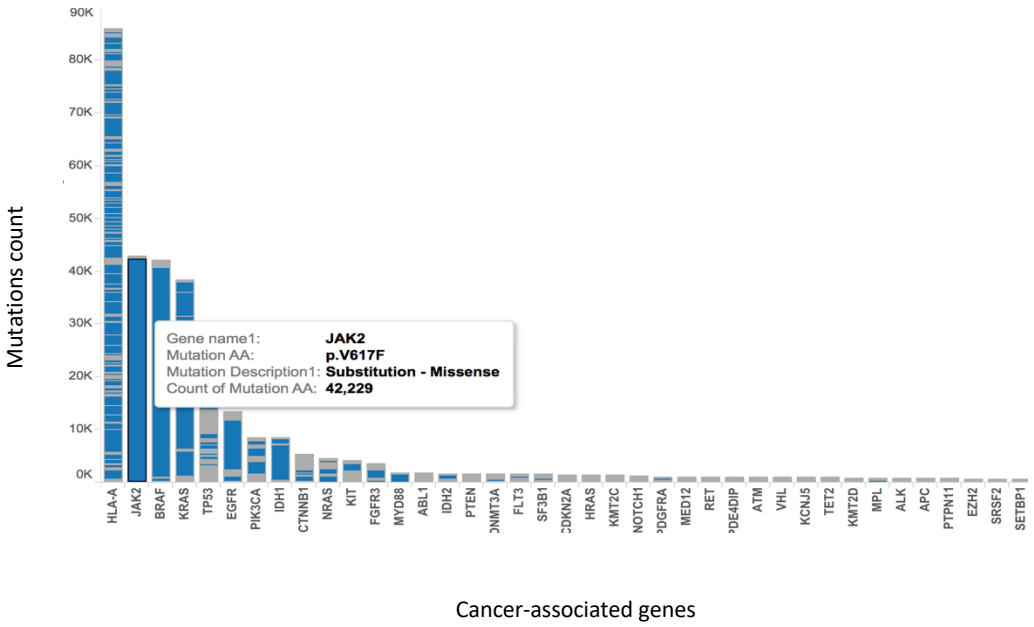

Figure 2.2 Most frequently mutated genes found in tumor samples as per COSMIC v77

Yet, this result identified the top frequently mutated genes found in tumors to target in our work. Using the oncogenes with available corresponding protein solved 3D structures (PDB – Sep 2016), we identified the following oncogenes to be suitable candidates for our study: *PIK3CA, IDH1, EGFR, KRAS, BRAF* and *JAK2*. We use this group of oncogenes and their reported missense mutations in ExAC/COSMIC to analyse the impact on protein stability.

2.3.1 The impact on proteins stability (energy of folding)

In the first phase, we established ΔG (the free energy of folding for each protein). For each entry on our frequently mutated genes list we look up the corresponding protein identification code in UNIPROT then we generated the FASTA sequence file while noting protein function and variants reported in UNIPROT. Using the FASTA file or UNIPROT id, we find the best PDB match (where organism is human sapiens, Resolution is equal or better than 3 Å and E value is less than 0. Finally, we downloaded the corresponding PDB file and ran the Foldx in repairPDB mode to find ΔG. The second stage was to calculate ΔΔG (the difference between the free energy of folding for each protein in wild and mutated state), we obtained all missense mutations reported in ExAC in addition to the cancer mutations reported in COSMIC for that specific gene and ran our program 'SpeedUP' calculating ΔΔG for all variants of a gene for that specific protein.

The distribution of ΔΔG found across all missense mutations for the oncogene *PIK3CA* is illustrated in figure 2.3. The x-axis represents the all missense mutations for this gene while the y-axis shows the ΔΔG for these mutations. Destabilising mutations (with positive ΔΔG values) are found to left end of the plot while the mutations with stabilising effect (negative ΔΔG values) are found to toward the right. The missense mutations implicated in cancer were highlighted and indicated by arrows. We noticed that the distribution is skewed towards the positive ΔΔG values indicating that majority of missense mutations had destabilising effect. The three cancer-related mutations reported in COSMISC for *PIK3CA* gene found to have a positive ΔΔG.

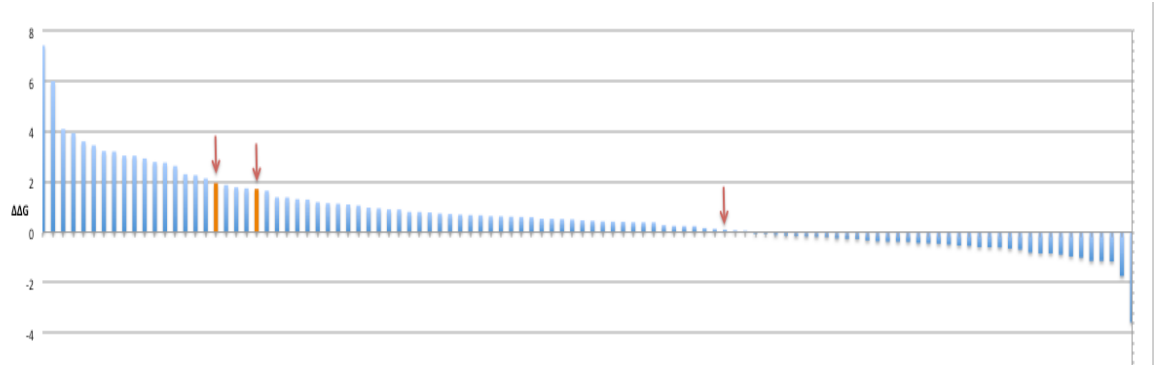

Figure 2.3 ΔΔG results for the PIK3CA gene' missense mutations indicating the positions of cancer variants.

This result concurs with expected destabilising effect of deleterious mutations on the protein (*4*). However, one of the cancer mutations (E542K) had ΔΔG = 0.1 value indicating neutral effect for that specific variant.

For the gene *IDH1*, again the values distribution was skewed towards positive ΔΔG indicating that most missense variants reported are of a destabilising effect. However, the cancer-associated mutation R132H showed neutral impact (ΔΔG = -0.18) on the stability of the protein (figure 2.4).

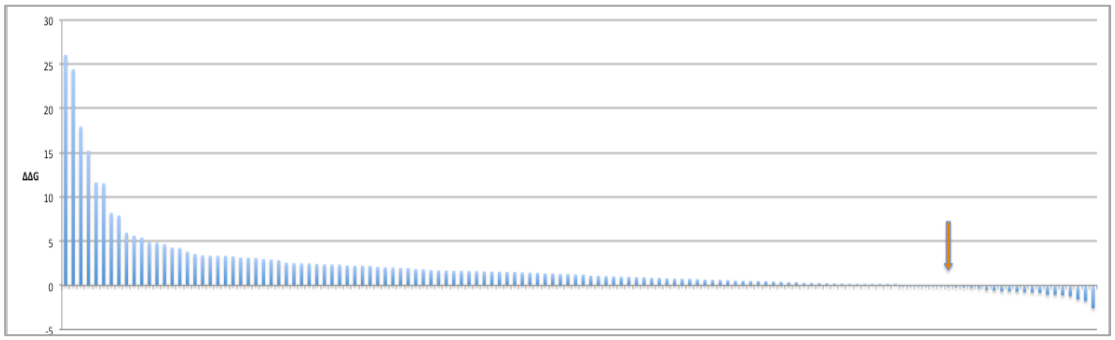

Figure 2.4 ΔΔG results for the IDH1 gene' missense mutations indicating the position of cancer variant.

KRAS is one of most studied cancer genes and has several mutations reported that play an important role in the initiation and progression of cancer (*11*). Six of *KRAS* known cancer mutations were analysed here. To calculate ΔΔG for all six cancer mutations in *KRAS* genes, two different PDB files had to be used (4LPK and 4QL3) encompassing the full protein sequence. Only one mutation (G13D) showed positive ΔΔG > 1 indicating destabilising effect on the protein and one mutation (G12R) had ΔΔG < -1 signifying stabilising effect while the other 4 mutations found to have neutral effect (figure 2.5, figure 2.6).

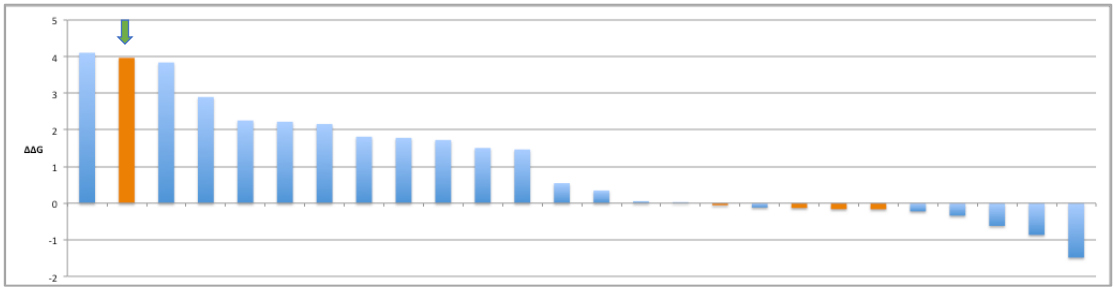

Figure 2.5 ΔΔG results (PDB: 4LPK) for the KRAS gene' missense mutations indicating the positions of cancer variants.

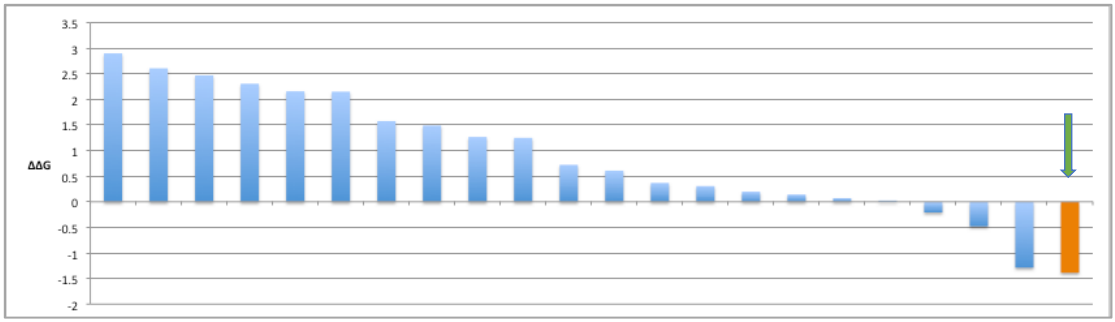

Figure 2.6 ΔΔG results (PDB: 4QL3) for the KRAS gene' missense mutations indicating the position of cancer variant.

In the case of the *BRAF* gene, there was one cancer mutation analysed (V600E) and found to have ΔΔG = 0.74 indicating neutral effect (figure 2.7).

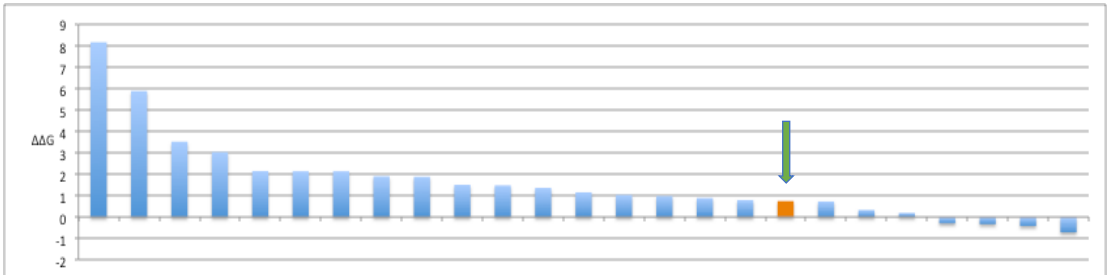

Figure 2.7 ΔΔG results for the BRAF gene' missense mutations indicating the position of cancer variant.

JAK2 gene had one cancer associated variant V617F that showed a stabilising effect (ΔΔG = -2.12). JAK2 V617F mutation was the most stabilising cancer mutation found in our sample (figure 2.8).

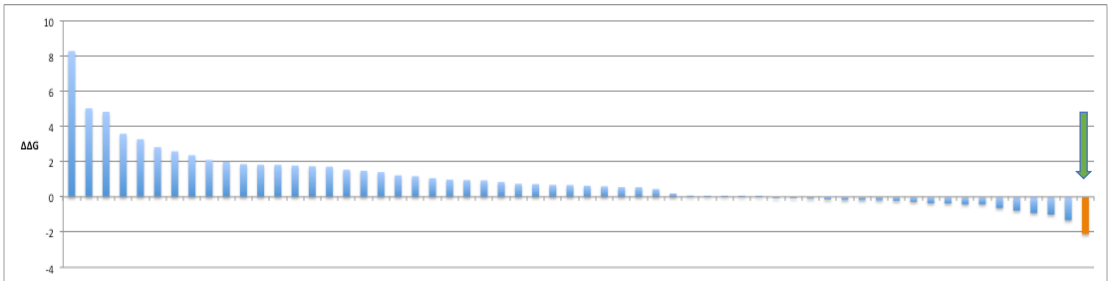

Figure 2.8 ΔΔG results for the JAK2 gene' missense mutations indicating the position of cancer variant.

Despite us not being able to analyse the stability effect of all missense mutations on all cancer genes due to the lack of availability of the corresponding PDB files (solved 3D structure), the results from the sample studied are pointing to different overall impact on stability between the cancer associated mutations and other missense mutations. The missense mutations are expected to have mainly a destabilising effect on the protein as found for other diseases. However, as our results showed most cancer mutations in our sample have a close to neutral effect. Although in every oncogene analysed most of the missense mutations had destabilising effect, only 3 of 13 cancer-associated mutations analysed had destabilising effect (figure 2.9).

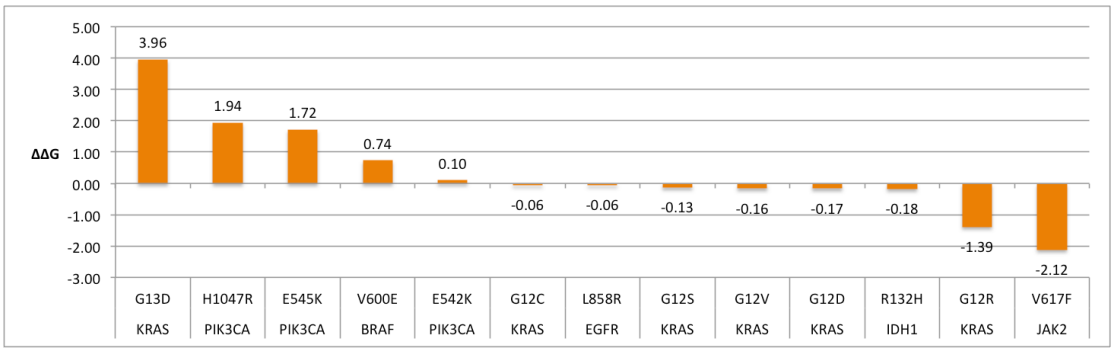

Figure 2.9 ΔΔG results indicating the impact on proteins' stability by all cancer variants studied.

2.3.2 The impact on proteins stability (half-life vs. Z constraint score)

We compared half-life of wild type proteins in our data sample as estimated by ProtParam and the ExAC Z constraint score reported for the related oncogene. The comparison between these two measures should provide an indication to the association between intolerance to variants and stability of the protein. On average, a gene is expected to have a shorter half-life the more it is intolerant to variants. In general, the half-life of the majority of proteins is less than 8 hrs (the median is estimated to be 7.1hrs) while the Z score ranges from 0 to 5 for the majority of the genes.

We found that only 2 out of 6 oncogene proteins studied show the expected effect on the half-Life in relation to the measure of Z score (figure 2.10, figure 2.11). The results indicate that genetic variants reported in the other oncogenes are not necessarily associated with destabilisation of their proteins. For instance, in the case of *JAK2*, the Z constraint is low (0.19) while the half-life is relatively long (30hrs) and for the *BRAF* gene the Z constraint is high (3.09) while half-life is relatively short (4.4hrs).

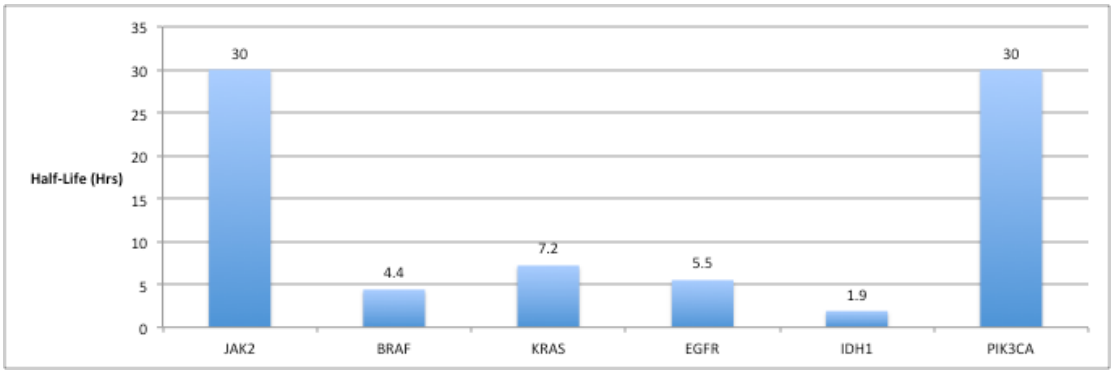

Figure 2.10 Half life in hours calculated for the protein coded by each of the selected cancer genes

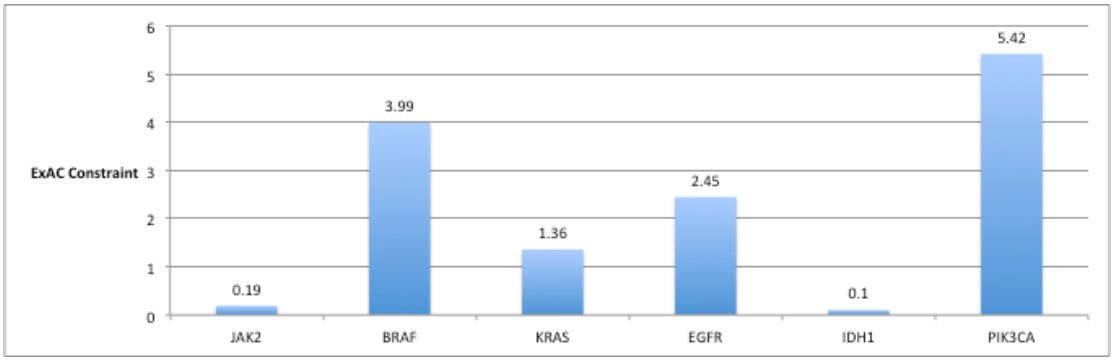

Figure 2.11 ExAC Z constraints reported for each of the selected cancer genes

That some oncogenes do not show a correlation between intolerance to the number of missense mutations and half-life (as a measure of stability for the protein), further indicates the distinct relationship cancer related oncogenes exhibit with stability.

If the neutral or stabilising effect of cancer point mutations is common, an overall destabilising effect might still occur through epistasis at the same gene level. The oncoprotein may have two or more mutations working altogether to alter the overall stabilising effect. We therefore investigated the number of point mutations from the same protein coding oncogene to occur together in a tumor sample as per COSMIC v77. We analysed point mutations found in sequenced tumor samples and found that less than 1% of the cancer-associated genes were reported to have more than one point mutation in one tumor sample (e.g., *CASCS*, *NOTCH1* and *USP6* genes all had more than one point mutation each in one tumor sample). Moreover, it is possible that an additive stability effect to occur in these rare cases. This finding shows that epistasis in one protein coding gene rarely exist

in cancer and that a cancer-associated point mutation with a neutral or stabilising effect is rarely countered by another point mutation.

Our result is pointing to a dissimilar effect of cancer-associated replacements on protein stability compared to other non-cancer deleterious replacements. This result could be explained in light of the altered function of the oncoproteins and its role in tumor initiation and progression. Factors such as the location of the mutation on the 3D structure of the protein and the type of amino acid being substituted play an important role to what constitute the effect of cancer mutations. For example, deleterious mutations with a destabilising effect are found mainly on the surface of the protein (*4, 12*). This could be different in the case of cancer mutations in general where the mutations that drive the tumor progressions may occur in varied locations on the protein 3D structure (*13*). Also, the distinct patters of amino acid replacements in cancer might explain this dissimilar effect. For example, the cancer associated E545K mutation occurs in the helical domain of the protein PI3K coded by *PIK3CA* gene. This mutation occurs in the p110α subunit stabilizing it, resulting in constitutive activation of the PI3K pathway leading to increased cell proliferation (*14*). Clearly both the location of the mutation and also the specific amino acid replacement (the amino acid E was replaced with the amino acid K) were crucial to allow the function alteration to take place. This could explain why the change in stability (particularly the destabilising effect) is not always observed on oncoproteins.

2.4 Novelty of results

Here I found that some of the most frequently mutated oncoproteins exhibit distinctive stability characteristics. Stability measured using the free energy of folding and the comparison between the half-life of the wild type protein and the constraint on number of variants shows that majority of cancer associated mutations have a neutral or stabilising effect on oncoprotein stability. This contrasts to the pattern found in other genetic diseases where deleterious variants have mainly destabilizing effects. This supports the main thesis hypothesis that cancer associated genes and proteins have distinctive characteristics.

2.5 References

1. S. Teng, A. K. Srivastava, C. E. Schwartz, E. Alexov, L. Wang, Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein protein interaction. *International Journal of Computational Biology and Drug Design* **3**, 334-349 (2010).
2. P. Yue, Z. Li, J. Moult, Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. *Journal of Molecular Biology* **353**, 459-473 (2005).
3. A. Bachmair, D. Finley, A. Varshavsky, In vivo half-life of a protein is a function of its amino-terminal residue. *Science* **234**, 179-186 (1986).
4. N. Tokuriki, D. S. Tawfik, Stability effects of mutations and protein evolvability. *Carbohydradtes and glycoconjugates / Biophysical methods* **19**, 596-604 (2009).
5. B. Lehner, Molecular mechanisms of epistasis within and between genes. *Trends in Genetics* **27**, 323-331 (2011).
6. R. L. Redler, J. Das, J. R. Diaz, N. V. Dokholyan, Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *Journal of Molecular Evolution* **82**, 11-16 (2016).
7. S. A. Forbes *et al.*, The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10.11 (2008).
8. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
9. J. Schymkowitz *et al.*, The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382-W388 (2005).
10. E. Gasteiger *et al.*, ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784-3788 (2003).
11. L. Huang, Z. Guo, F. Wang, L. Fu, KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy* **6**, 386 (2021).
12. N. O. Stitziel *et al.*, Structural Location of Disease-associated Single-nucleotide Polymorphisms. *Journal of Molecular Biology* **327**, 1021-1030 (2003).
13. J.-J. Liu *et al.*, The structure-based cancer-related single amino acid variation prediction. *Scientific Reports* **11**, 13599 (2021).
14. Y. Hao *et al.*, Gain of Interaction with IRS1 by p110α-Helical Domain Mutants Is Crucial for Their Oncogenic Functions. *Cancer Cell* **23**, 583-593 (2013).

3. Chapter 3: Characteristics of favoured amino acids found in point missense cancer-associated mutations

*** Statement of Authorship**

All the work presented in this chapter including analysis and method development were carried out by Amro Safadi with supervision of his supervisors prof. Simon C. Lovell. and prof. Andrew J. Doig. Both supervisors approved the figures and final wording of the manuscript providing corrections when needed.

3.1 Introduction

Analysing the overall spectrum of amino acid changes that impact tumorogenesis is of great importance. Our understanding of cancer would be enhanced if specific mutational patterns and certain characteristics could be identified in cancer-associated amino acid replacements. A single amino acid change could affect the protein structure and function and identifying amino acid replacements enriched in cancer could point towards certain properties that when altered could lead to tumor formations and progression (*1*). Expanding datasets of reported and verified cancer-associated mutations give an opportunity to discover new insights. A prime source is the constantly updated COSMIC database (https://cancer.sanger.ac.uk) (*2*) where these amino acid replacements can be accessed and used to potentially discover patterns in the mutational landscape of cancer.

At the nucleotide level, C:G>G:C changes in lung, ovarian and other cancers are strongly enriched at TpC/GpA dinucleotides (*3*). However, the biological basis of this mutational signature remains unknown (*3*). That some amino acid replacements are favoured in disease-associated variants was demonstrated in (*4*) where they looked at the 1000 Genomes Project for humans and found that the amino acid exchange matrix generated from the observed nucleotide variants is asymmetric and that disease-associated variants differ from other non-disease-associated variants.

Across human genetic diseases, amino acid replacements were shown to have the prevalence of Arg and Gly at the original residue (*5*). This study used the Online Mendelian Inheritance in Man (OMIM) database representing Mendelian diseases

(https://www.omim.org) (*6*) and did not confirm if the results pertain to cancer. The study demonstrated that the spectrum of replacement in disease correlates well with the amino-acid replacement frequencies based on the genetic code (normalized by the mutation frequencies) (*5*). The probability of a mutation to cause a genetic disease goes up with an increase in the degree of evolutionary conservation at the mutation site and a decrease in the solvent-accessibility of the site (*5*).

In a study that did look specifically at cancer amino acid changes (*1*), replacements that are most frequent are identified as those likely to lead to the cause or progression of cancer (i.e., drivers) while the least frequent are identified as passengers (playing passive a role in tumor progression). The R → H substitution was shown to be favoured in drivers followed by R → Q and R → C, whereas E → K has the highest frequency in passenger mutations. It also demonstrated that substitution of Arg is frequently found in many cancer types. Although this study analysed the mutational landscape at the amino acid level, the frequency of replacements could be affected by several factors not related to cancer intrinsic processes. These replacements could be found frequent due to selection pressure or in genetic diseases in general. It is yet unclear how these favoured replacements impact cancer onset and progression (*1*).

There is a need to utilise the expanded dataset of cancer-associated mutations collated in COSMIC and use computational methods to confirm the enrichment of certain amino acid residues and replacements free from the influence of the number of samples analysed. This should be achieved via comparison against control groups that can reveal the cancer specific enrichments of these residues and replacements. Identifying these highly enriched replacements specific to cancer would allow us to link their distinct activities to physiological processes driving cancer highlighting the importance of these activities. It would also allow us to use amino acid properties in an informed way to train machine-learning model enabling the prediction of other replacements yet to be identified as cancer-associated.

In this study, we analysed missense cancer point mutations reported in the COSMIC cancer database. Unlike previous technique explained in (*1*), we avoided relying on the occurrence frequency based on samples analysed and focused on mutations reported and verified in the literature. This approach reduces the likelihood of bias that might rise if the frequency of occurrence is used. The occurrence frequency of a mutation reported could lead to miscalculation in the results due to several unrelated factors, such as number of tumors sampled for one cancer patient or the tendencies to focus on certain types of tumor samples that are easier to obtain. We used the probability of each expected amino acid replacement based on the genetic code, and transition and tranversion rates to demonstrate that certain amino acid replacements are occurring at a higher frequency than expected in cancer. This contrasts with Mendelian diseases where they were shown to occur at the expected frequencies confirming that cancer mutations are under selection. We analysed both the amino acids that were replaced and the amino acids that they were replaced with (we term these 'original' and 'replacement' residues respectively). We confirm that Arg is the amino acid with the highest likelihood of mutating in cancer. We also compared to mutation frequencies in ExAC containing variants from other genetic diseases (*7*) and Blosum62 containing expected amino acid enrichment rates based on protein sequence alignment (*8*). We found that Cys and Trp are highly enriched in cancer mutations as the replacement residues and that the amino acid replacement patterns in cancer are more diverse than previously thought (*1*). We highlighted that there are 17 particularly favoured amino acid replacements; these replacements have a strong presence of the aromatic amino acid group (about 30%) in the replacement residues and a similar presence of the 'Stop' codon in the same position. The strong enrichment of certain amino acids in cancer replacements could be explained in the light of the widespread involvement of cancer genes in the 'binding' biological process as found in Gene Ontology Consortium (http://www.geneontology.org/) (*9*). Aromatic amino acids group are specifically responsible for forming 'stacking interactions' (see section 1.2.2.2) critical in recognizing binding sites while potential disulphide bonds formed by Cys residue could lead to constant activation of certain pathways necessary to cell proliferation to take place. 'Stop' codon on the other hand would likely terminate or shorten the produced protein.

We also found highly enriched cancer-associated single amino acid variants preferentially occur early in the protein sequence indicating the potential importance of these sites in influencing functions and binding affinity related to cancer. We also found that these variants exhibit on average increase in hydrophobicity and decrease in polarity in comparison to other non-cancer missense mutations. The different characteristics of cancer-associated single amino acid replacements prompted us to build a machine learning model trained using the amino acid physico-chemical properties to predict if a replacement is cancer-associated or not. This could be applied for each oncoprotein assuming there are enough number of replacements reported to use to train the machine learning model. An example used here was the protein PTEN, the model showed good performance (F1 score of 0.76) in distinguishing cancer associated replacements using amino acids properties and can be used without limitation compared to a model reliant on protein structure data. However, the model predictability could significantly increase by combining the two approaches. These findings may assist further in discovering novel cancer-associated mutations and further understand the protein functional changes caused by these mutations in the initiation and progression of cancer.

3.2 Materials and Methods

We used the Chi-square test to estimate the significance of linear correlations between the expected and observed probabilities and t-test for the rest of the dataset's comparisons with p-value < 0.0008 throughout. Calculated ratios were rounded up to three decimal points.

3.2.1 Cancer-associated amino acids enrichment ratios list

We have compiled a list in excess of 60K distinct cancer-associated missense mutations. These mutations belong to 590 cancer genes as listed in the COSMIC Census dataset of October 2017 (*2, 10*). We then counted the number of times each amino acid was found in a mutation differentiating between the mutated amino acid in the 'original residue' position and the amino acid resulting from the mutation (the 'replacement residue' position).

The following conditions were observed when including the mutations: (i) silent mutations are excluded. (ii) A cancer-associated mutation is only counted once no matter how many samples it was found in. This will ensure equal weight given to each mutation regardless of number of samples being analysed. (iii) A cancer mutation is only counted if it occurs as a single codon change as this study focuses on missense point mutations only (point mutations are the most common type of genetic variation).

3.2.2 Expected amino acids enrichment ratios in mutations

Genetic code frequencies were used to determine the initial amino acid expected occurrence rates. We then used transition or transversion rates (*11*) to calculate the expected frequency for each amino acid to be replaced by another. We started with all 61 non-stop codons and applied every possible one-step nucleotide base change.

We then listed all possible cases, where a case is a pair consisting of an initial amino acid and all the possible amino acids that it could mutate to (e.g., F → Y), we multiply the frequency of each starting codon in the human genetic code by the probability of each transition or transversion to find the probability of that specific pair. The results are a list of all possible codon changes from point mutations with a probability attached to each of these pairs.

The list contains duplicated amino acid pairs (this because the same amino acid might rise from more than one codon. For example, ACC→CCC and ACG →CCG both result in the replacement T → P); we therefore summed the probabilities of matching pairs to find the final list of amino acid pairs and their probabilities.

3.2.3 Amino acids properties dataset

We started with a list of general physical properties of amino acids such as polarity, charge, and volume in addition to using two widely known measures of hydrophobicity; the Kyte-Doolittle scale and the Janin scale (*12, 13*).

We then extended the properties to use physico-chemical, energetic, and conformational properties of the 20 amino acids to quantify the mutation impact on protein properties obtained (*14*). These properties have previously been shown to be important in further understanding the folding and stability of proteins (*15*). The amino acid properties were normalized between 0 and 1 using the expression: Pnorm(i)=[P(i)−Pmin]/[Pmax−Pmin]

where P(i), Pnorm(i) are, respectively, the original and normalized values of amino acid i for a particular property, and Pmin and Pmax are, respectively, the minimum and maximum values.

For each oncoprotein and for each replacement found there we subtract the property value of the amino acid at the original residue position from the amino acid at the replacement residue position recording the difference for all properties. This allows for the resulting dataset to be used to train a machine-learning model to predict the probability of any replacement to be a cancer associated for that specific oncogene. The full list of amino acids properties and their definitions are available on the web at http://www.iitm.ac.in/bioinfo/fold_rate/.

3.2.4 ExAC mutations dataset

The Exome Aggregation Consortium (ExAC) database – Oct 2017 spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. It recently became part of the gemoAD database (*7, 16*). We extracted all mutations annotated as 'Missense' recorded for all the 590 genes that are identified as cancer genes in COSMIC of Oct 2017 from the ExAC database. We then excluded all the mutations that were already reported in COMIC as cancer-associated mutations. This left us with a set of missense variants (~80K) that were not yet identified as cancer associated mutations but were found in the ExAC population.

The process of collecting the missense mutations from ExAC started with collecting all missense mutations from all identified cancer genes, isolating the amino acids that appear as the original and replacement residues, recording the frequency rate and then normalising

the results to produce two ordered lists of enriched amino acids (the original and replacement residues).

3.2.5 Machine learning method

To create a machine learning model that predict the likelihood of a certain mutation to be cancer associated we used the DataRobot platform (https://www.datarobot.com) where the dataset described in 3.2.3 was used to train the model. The dataset was submitted to the software following a short preparation to ensure that the dataset file had one of the following accepted delimiters: comma (,), tab (\t), semicolon (;), or pipe (|). All dataset feature names must be unique, and dataset must be a single file.

Following the login to the software web application, a new project is created once our full dataset is uploaded using the button provided on the first page. The data fields are automatically assigned an appropriate data type (e.g., numeric or categorical, etc) by the software. The dataset fields are then listed with basic data analysis automatically performed. This analysis includes for numeric fields values such as mean, standard deviation, median, min and max. The software excludes any duplicate or empty fields and performs automatic data quality checks that detect any outliers.

The software performs automatic data preparation including imputation of missing values and ordinal encoding of categorical variables. All transformations are listed in the model log provided by the software. In our tree-based model, we kept the default setting of the software where numeric missing values are imputed with an arbitrary value (-9999) and for categorical variables, missing values are treated as an additional level in the categories.

The next step would be selecting the field within the dataset that contain the outcome of each row. To build a supervised machine-learning model, it is necessary to identify the predicted classes and individual outcome to every row in our dataset. The outcome in our dataset is binary (true or false), indicating whether a single amino acid replacement for the protein in question is cancer associated. For our model, this binary field represents a mutation being either cancer associated (value = 1) or not (value =0). The training dataset

included all instances of missense replacements that were reported for the PTEN protein and whether each replacement was deemed cancer associated (as reported on COSMIC database). The software detects the type of classification required (binary classification in this case) once the predicted class is selected. DataRobot displays a histogram providing information about the target feature's distribution and list available performance metrics that would be calculated to select the best performing model. The recommended performance metric for our project was logistic loss.

Once the model building phase is completed, the software indicates the best performing models based on the recommended optimization metric. DataRobot searches through a repository of possible combinations of algorithms based mainly on 'open source' libraries for supervised learning algorithms such as Python-sklearn and tests several parameters values before producing final results. Also, DataRobot uses heuristic logic to recommend the best performing model. The top performing model (recommended by the software) in our project was Gradient Boosting. This model achieved the best logistic loss across both training and validation datasets. Only models that have the same level of performance across training, validation and testing are displayed on the leaderboard page.

To avoid over-fitting, the best practice is to evaluate model performance on out-of-sample data. If the model performs very well on in-sample data, (the training data), but poorly on out-of-sample data, that is an indication that the model is over-fit. The k-fold cross-validation is a standard technique used to validate model performance and ensure that over-fitting does not occur. DataRobot uses a 5-fold cross-validation framework as the default option to test the out-of-sample stability of a model's performance. DataRobot automatically divides the original dataset into the respective training and validation sets. We kept the default 5 folds rather than choosing a smaller number of partitions as the size of the dataset allow for that and thus, we ensure more through testing. In addition to the cross-validation partitioning, a holdout sample (test sample) is used to further test out-of-sample model performance ensuring appropriate evaluation of the model performance and reducing likelihood of over-fit. 20% of the training data is set aside as a holdout dataset. This dataset is used to verify that the final model performs well on data that has not been touched throughout the training process, while the remainder of the data is divided into 5

cross validation partitions. Because the distribution of the target's values in a binary classification project may be imbalanced, the validations' partitions were randomly selected using a stratified sample approach (this is the default option in the software) where sub-populations within the data are always represented in each partition to preserve the distribution of the target's values for each partition.

Our model algorithm is Gradient Boosting Machines (or Generalized Boosted Models, 'GBM'). GBM is a cutting-edge algorithm for fitting extremely accurate predictive models (*17*). GBMs are a generalisation of Freund and Schapire's adaboost algorithm (1995) modified to handle arbitrary loss functions. They are very similar in concept to random forests, in that they fit individual decision trees to random re-samples of the input data, where each tree sees a bootstrap sample of the rows of the dataset and N arbitrarily chosen columns, where N is a configurable parameter of the model. GBMs differ from random forests in a single major aspect: rather than fitting the trees in parallel, the GBM fits each successive tree to the residual errors from all the previous trees combined. This is advantageous, as the model focuses each iteration on the examples that are most difficult to predict (and therefore most useful to get correct). Due to their iterative nature, GBMs are almost guaranteed to over-fit the training data, given enough iterations. The two critical parameters of the algorithm, therefore, are the learning rate (or how fast the model fits the data) and the number of trees the model is allowed to fit. It is critical to cross-validate these two parameters. When done correctly, GBMs are capable of finding the exact point in the training data where over-fitting begins, and halts one iteration prior to that. In this manner, GBMs are usually capable of producing the model with the highest possible accuracy without over-fitting (*17*).

Our model uses logistic loss and early stopping to determine the best number of trees. Early stopping is a method for determining the number of trees to use for a boosted trees model. The training data is split into a training set and a validation set; in each iteration the model is scored using the validation set. If validation set performance decreases for 200 iterations, the training procedure stops, and the model returns the fit at the best tree seen so far. Note that the early stopping validation set will be a 90/10-train/validation split within the training data for a given model. The model will therefore internally use 90% of the available training

dataset and 10% of the data for early stopping. Since the early stopping test set was used to find the optimal termination point, it cannot be used for training.

To set the hyperparameters used by the model, the DataRobot platform performs an internal "grid search" with pre-set hyperparameters values ensuring optimum accuracy. The default setting of the platform avoids a 'brute force' strategy where every possible value of a parameter is tested. The platform strategy relies on setting these parameters for the model to be built in reasonable timeframe. We have used the hyperparameters values selected by the platform without alternation.

In our model, several guardrails were implemented to mitigate possibilities of data labelling bias. We ensured the dataset used for training does not carry any overrepresentation for any feature or group and used several performance metrics to evaluate the model performance such as Logistic Loss to eliminate any chance of overfitting or underfitting. Bias is usually detected when the difference between a model's predictions for different populations (or groups) is evident. DataRobot implement bias mitigation techniques for reducing model bias for a predicted class. The model can be tested using *Proportional/Equal Parity (also known as Demographics Parity)* where the platform shows the probability of receiving favourable predictions for one of the predicted classes from the model or what is the total number of records with favourable predictions from the model for each class. Other bias mitigation metrics can be selected on the platform that evaluate the bias based on equal error of the predicted classes such as *Favourable Predictive Value Parity*.

DataRobot automatically mitigates any bias found using the metrics above by applying *Pre-processing Reweighing* where row-level weights are used as a special model input during training to attempt to make the predictions fairer. This was not needed in our model as no bias was detected by the platform. Finally, to eliminate a chance of implicit bias in our model, the model is built for one oncoprotein at a time and results should not be generalised to all other proteins ensuring no overgeneralisation bias can occur.

Unaccounted dependency within the examples and confounder variables used to train the model may create a false enhanced performance by creating a false link between some of

the features and the outcome. The performance metric GiniNorm was used to discover any circularity that may exist in the data. A matrix was calculated showing the GiniNorm values between all feature pairs and all features and predicted outcome class. Any GiniNorm value of 0.85 or more indicated abnormal correlation and that feature was automatically eliminated. This threshold is preselected by the DataRobot platform but can be changed. We have kept the default GiniNorm threshold used by the platform to detect and eliminate any circularity in the data. We also paid special attention to understanding the data itself used to build the model and understand what the features represent to avoid problems such as circularity.

Another common machine learning pitfall that we tested our model for is the possibility of 'leakage' where a feature used in the training data would not be fully formed until the outcome has occurred. For instance, predicting whether the replacement is cancer-associated using data that is only known following the replacement detection. This could create a false level of accuracy and such correlation should be detected and eliminated before the model is built. DataRobot implements a leakage detection method based also on GiniNorm metric and if such case is at hand, the software alerts the user before starting the modelling process, giving the user the ability to remove the field(s) that are causing the leakage.

Once the models' building process is completed, the software automatically makes available evaluation results for each competing models. Insights available include the model workflow, the features selected and their relative impact on prediction. Also produced is the accuracy related information such as the *ROC curve*, *sensitivity*, and *specificity* (our model accuracy results are discussed below in the results section 3.3). The software also provides the ability to upload new dataset (with no known outcome) to calculate the predictions for.

3.3 Results

For any specific patterns in cancer-associated mutations to be revealed, it is necessary to compare against a control group. We chose three different control groups reflecting

different selection scenarios. The first serves as our 'Null' model where the expected probability of mutations is calculated assuming they are under no selection. The second control group was using Blosum62 reflecting the effect of natural selection and the third is the ExAC dataset containing missense mutations for other genetic diseases. These calculations are explained in the next sections.

3.3.1 Amino acid residues probabilities of occurrence under no selection

The results here reflect the probability of an amino acid to feature in a mutation if assumed to be free from evolutionary selection pressure. This enables us to identify any elevated replacement rates arising by selection for cancer.

Each mutation can be represented by two amino acids (e.g., A→C); we call the initial amino acid on the left side of the pair 'original residue' and the one on the right 'replacement residue'.

By summing up all the probabilities of pairs (calculated using codon frequencies and transition or transversion rates), then normalising the numbers (considering the total sum to equal 1 and assigning each amino acid a frequency accordingly), we obtain the table 3.1:

Original Residue	**Normalised Frequency**	**Replacement Residue**	**Normalised Frequency**
L	0.084	R	0.085
S	0.082	S	0.083
E	0.072	L	0.073
A	0.066	V	0.065
P	0.065	P	0.064
K	0.058	A	0.063
G	0.056	T	0.062
V	0.054	G	0.057
Q	0.054	Stop	0.049
R	0.051	I	0.047
T	0.050	D	0.042
D	0.049	N	0.040
I	0.044	E	0.040
F	0.043	H	0.039
N	0.037	K	0.037
H	0.030	F	0.033
Y	0.030	Q	0.032

M	0.028	Y	0.027
C	0.026	C	0.027
W	0.017	M	0.023
Stop	0.003	W	0.011

Table 3.1: The normalised frequency of expected amino acid residues using codon frequencies, and transition or transversion rates.

3.3.2 Amino acids frequencies based on the genetic code vs. their frequencies in cancer mutations

Figure 3.1 shows the ratio of the frequency of the amino acids in cancer missense point mutations at the replacement residue position to the expected frequency; similarly, figure 3.2 shows the ratio of the frequency of the amino acids in cancer missense point mutations in the original residue to the expected frequency. We found that in addition to the stop codon, Trp and Cys and Lys are often enriched at the replacement residue position in cancer mutations with the following enrichments ratios respectively (1.98, 1.77 and 1.54). Pro, Ala and Gly are disfavoured at the replacement residue position in cancer mutations. In the original residue, we found that Arg is highest, followed by Gly and Glu), while Phe, Ile and Leu are disfavoured.

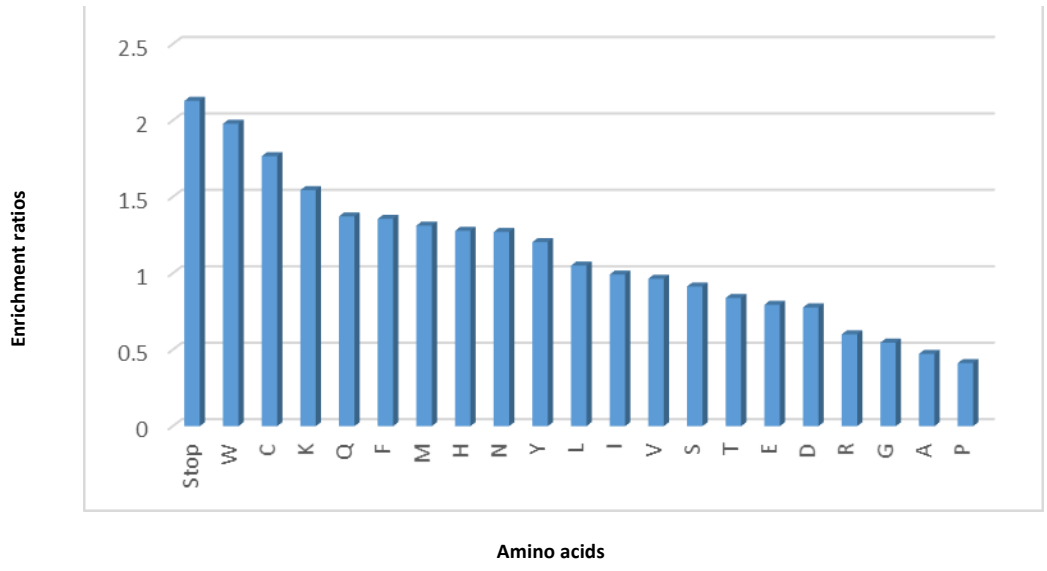

Figure 3.1 Enrichment ratios of 'replacement residues' in cancer-associated mutations when compared to frequencies based on the genetic code

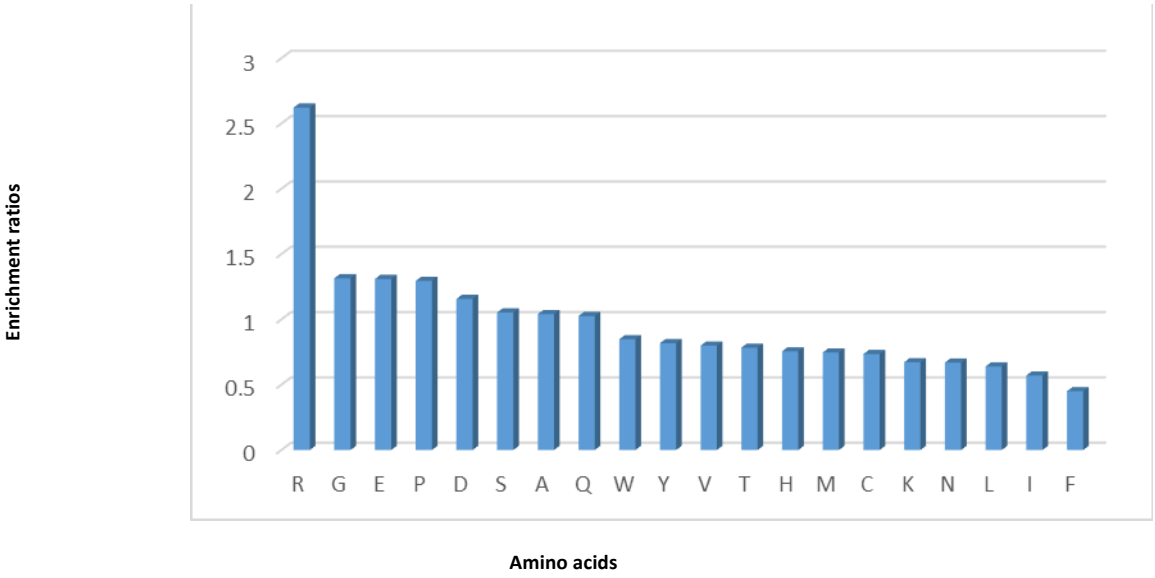

Figure 3.2 Enrichment ratios of original residues in cancer-associated mutations when compared to frequencies based on the genetic code

We also calculated the coefficient of determination (figure 3.3) for the normalised frequencies of amino acids at the 'replacement residue' position and their expected frequencies based on the genetic code. This showed an acceptable degree of association between the two sets at for this type of experiments.

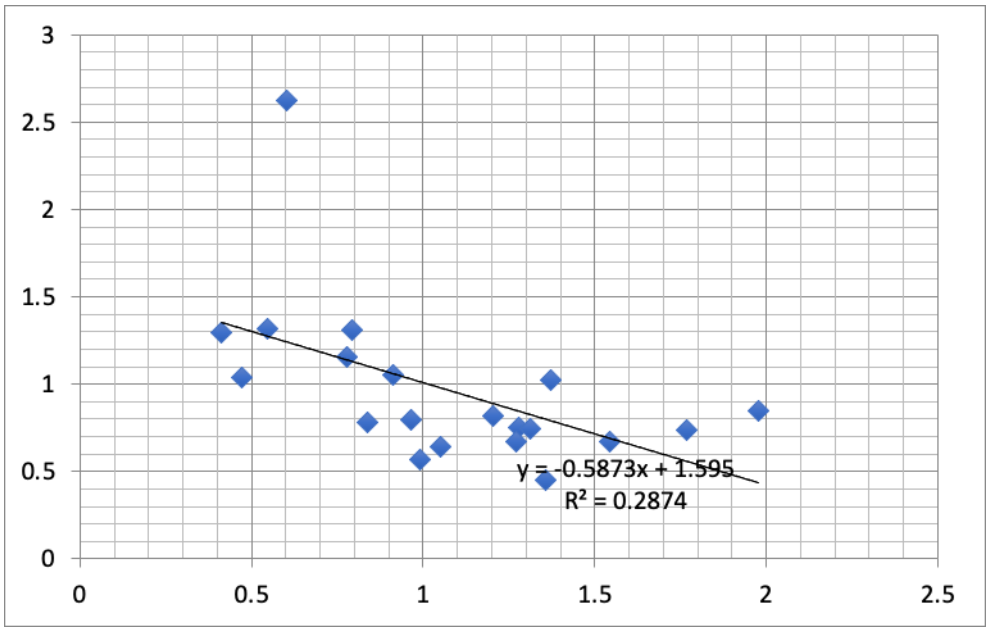

Figure 3.3 Coefficient of determination calculated for the replacement residues

When we repeated the calculation counting all the instances a mutation was repeated in tumor samples, the results were different. Val replaced Arg as the amino acid that is most

mutated in the original residue position while F and E became favoured at the 'replacement residue' position. This indicates the importance of excluding these repetitions due to the bias in results that they would produce.

Figure 3.2 confirms that Arg is the amino acid with highest probability to mutate in cancer-associated mutations. This result concurs with previous study that reported Arg to be highly mutable in human disease-associated amino acid variants (*4*). They suggested that because four out of the six codons for Arg include CpG sequences and CpG dinucleotide in DNA are known to mutate at high rates then this might explain the higher probability of Arg to mutate (other amino acids like Pro, Thr, Ser and Ala do have CpG sequences but only in one codon compared to four codons in Arg).

In an attempt to provide additional possible explanation linked to cancer, we mapped all cancer genes in our study to the molecular functions available using the Gene Ontology Consortium (http://www.geneontology.org/). We found that a higher than average involvement in the 'Bindings' functions at fold enrichment (for all binding types) >1.35 and in some binding functions > 9 such as damaged DNA binding (*18*). This could be explained by a high likelihood of mutations in DNA binding proteins to be drivers of disease onset and progression (*19*). This in turn may explain why Arg is the most likely amino acid to mutate in cancer missense mutations, as Arg is found to be quite frequent in binding sites and plays a key role in the stability of the protein; replacement of Arg is likely to have a detrimental effect on the function and structure of the DNA binding complex (*20, 21*).

Furthermore, we have listed all the oncogenes In COSMIC database that reported to have a cancer point mutation that feature the amino acid Cys and Trp at the 'replacement residue' position and used UNIPROT to find the identified function of their proteins product. For example, the oncogene *LRP1B* is the oncogene with the highest number of mutations with the amino acid Cys as the replacement residue and the protein product is a cell surface protein that bind and internalize ligands in the process of receptor-mediated endocytosis. The second highly enriched oncogene with mutations that feature the amino acid Cys at the replacement residue position is *FAT4* and the protein is calcium-dependent cell adhesion protein. The same affinity to binding activities pattern was emerging for oncogenes enriched

with mutations where Trp feature as the replacement residue. The special characteristics of these two amino acids might be directly linked and effect the biological job the oncogene protein is implicated in.

Whole replacement analysis

We found that the E→K and R →H substitutions have the largest enrichment ratio among all cancer genes agreeing with the results from (*1*) where machine learning approach was used to determine the enriched replacements. E→K replaces a negatively charged amino acid for a positively charged one; this is likely to substantially alter the electrostatics, conformation, stability and interactions of the protein (*22-24*). An R →H substitution replaces an always protonated amino acid for one that can be positive or neutral thus potentially affecting the pH sensitivity (*25, 26*) and function (*27*) of the protein. Why such changes are particularly frequent in cancer is unclear.

However, when we compared the cancer dataset substitution frequencies to those based on codon frequencies and re-ranked the substitutions accordingly (Table 3.2), we found that R→H is most enriched in cancer mutations followed by R→Q, R→ Stop and R → C. E→K while is still enriched, its enrichment ratio was less than half of R→H and R→Q ratios.

Amino Acid Replacement	**Enrichment ratio**
R→H	6.75
R→Q	6.62
R→Stop	5.76
R→C	5.28
T→M	4.70
R→W	3.80
E→Stop	3.60
S→L	2.86
R→I	2.82
E→K	2.72
G→Stop	2.54
S→Stop	2.52
D→N	2.36
S→F	2.33
D→Y	2.28

R→L	2.19
Q→Stop	2.08

Table 3.2 Cancer- associated replacements with ratios > 2 in when compared to expected frequencies based on codon frequencies

In total, there were 17 replacements with an enrichment ratio > 2. Over 40% of these had Arg at the original residue position. We also found a strong presence of the aromatic amino acid group (about 30%) in the replacement residue position and a similar presence of the 'stop' codon in the 'replacement residue' position. Our results here agree with a previous study that highlighted R→H, R→Q and R → C as driver mutations in cancer (*28*). However, our results differ in showing that there are many more replacements that should be considered as cancer drivers including E→K that belongs to enriched replacements in cancer-associated mutations and we recommend reviewing its label stated in (*28*) as a passenger mutation.

3.3.3 Amino acids frequencies in Blosum62 vs. their frequencies in cancer mutations

Blosum62 provides a substitution matrix with scores for all possible exchanges of one amino acid with another based on aligned protein sequence segments and it is shown to provide an improvement in sequences alignment compared to other methods (*29*). We correlated the frequencies of amino acids appearing in our cancer mutations dataset with the normalised (each row in the matrix made to equal 1) values of Blosum62 matrix. The comparison reveals different substitution frequencies in cancer mutations compared to the Blosum62 matrix, confirming that cancer mutations are under selection. The results here also show a similar pattern of replacement residues to the findings when comparing with the expected frequency based on the genetic code (Table 3.2 and Figure 3.1). We find that also Arg is favoured here as the original residue and Cys and Trp as the replacement residues (Table 3.3).

Top Replacements	**Enrichment ratio**
R→C	55.44
R→W	54.38
Y→C	25.70
W→C	24.04
P→L	22.60

Table 3.3 The most enriched cancer-associated replacements when compared to frequency found in Blosum62

3.3.4 Amino acids frequencies in ExAC (non-cancer missense mutations) vs. their frequencies in cancer mutations

As our ExAC dataset contains human variants (missense mutations) that are associated with other diseases (cancer-associated mutations found in COSMIC are excluded), our comparison of the frequencies between the two datasets aimed to reveal the certain amino acids patterns in missense mutations that pertain specifically to the cancer mutations.

Replacement Residue Amino Acid	Cancer to non-cancer disease ratio
F	1.52
Y	1.50
K	1.36
N	1.17
D	1.04
C	1.02
W	1.01
H	1.00
L	0.97
E	0.9
Q	0.89
I	0.89
S	0.84
P	0.84
M	0.8
V	0.72
T	0.72
G	0.67
R	0.63
A	0.59

Table 3.4 Ratios of amino acid frequencies as replacement residue in Cancer mutations compared to non-cancer missense mutations

The results (table 3.4) show that aromatic amino acids (Phe and Tyr) have a higher enrichment in cancer mutations at the replacement residue position when compared to

other diseases. This further emphasises the role of aromatic groups in cancer onset and progression. We hypothesis that the aromatic group involvement in forming stacking interactions (see section 1.2.2.2) which is essential for achieving successful binding in both DNA and RNA sites (a biological process shown to be affected by most common cancer mutations(*30*)) can explain their high enrichment in cancer missense mutation.

We have also compared the amino acids frequencies in the ExAC dataset of missense mutations to the expected amino acids frequencies in mutations calculated using codon frequencies (table 3.5). This would reveal the amino acids that are prominent in genetic diseases in general. We wanted to see if these differ from the results when comparing to cancer. We found that similar to the comparison with cancer mutations dataset, amino acids Trp and Cys feature again (figure 3.1) as likely amino acids at the replacement residue position when compared to the expected spectra calculated based on codon frequencies. These findings indicate the important role the specific properties of Trp or Cys amino acids in missense mutations driving the onset of genetic diseases.

Replacement residue Amino Acid	**Enrichment ratio**
W	1.97
C	1.73
M	1.65
Q	1.54
V	1.34
H	1.28
T	1.17
K	1.13
I	1.11
S	1.09
L	1.08
N	1.08
R	0.95
F	0.89
E	0.89
G	0.89
Y	0.8
A	0.8
D	0.75

P	0.49

Table 3.5 Ratios of amino acid frequencies as replacement residues in non- cancer missense mutations compared to those expected from codon frequencies

In addition, the aromatic amino acids, we see Cys is highly enriched in cancer mutations. The amino acid Cys is known to be involved in forming disulphide bonds (see section 1.2.2.2) crucial of correct folding of the protein (*31*). If a replacement occurs with Cys at the 'replacement residue' position, then there is chance that a disulphide bond will form and, in some cases, altering (generally increasing) the stability of the protein. These disulphide bonds were shown to have an impact on protein structure and function, a study of hST3Gal1 showed that removing a Cys residue abolished the enzyme activity (*32*). Another example is Epidermal Growth Factor (EGF), EGF is a small protein that stimulates cell proliferation and is shown to have disulphide bonds formed by the Cys residues (*33*). Integrins which are transmembrane receptors with EGF – like domains play a key role in regulating cellular growth, proliferation and signalling and introducing Cys mutations were also shown to cause αIIbβ3 (a subfamily of Integrins) to be constitutively activated (*33, 34*). The constant 'turned on' state observed in some signalling pathways leading to proliferation of the cell (one of cancer main hallmarks) could be induced by the impact of these unintended bonds.

3.3.5 The analysis of the physical properties of amino acid replacements highly enriched in cancer

We analysed the hydrophobicity, polarity, charge, and volume properties of the amino acids in enriched cancer-associated replacements by calculating the difference in value of the amino acid at the replacement residue position from the value of the amino acid at the original residue position for each of these properties. Hydrophobicity measure was obtained based on two of the most used methods (Kyte-Doolittle and Janin) to measure hydrophobicity (each method is known to excel at measuring hydrophobicity for certain protein types). Table 3.6 shows these changes for replacements deemed enriched in cancer when compared to the expected frequencies based on genetic code. Table 3.7 shows these changes for replacements deemed enriched in cancer when compared to Blosum62.

Amino Acid	**Hydrophobicity (Kyte-**	**Hydrophobicity (Janin scale)**	**Polarity**	**Charge**	**Volume**

Replacement	Doolittle scale)				
R→H	1.3 ↑	1.3↑	0	0	-32.8↓
R→Q	1 ↑	0.7↑	-1↓	-1↓	-43.4↓
R→Stop	2.9 ↑	1.1↑	-2↓	-1↓	-69 ↓
R→C	7 ↑	2.3↑	-2↓	-1↓	-87.8↓
T→M	2.6 ↑	0.6↑	-1↓	0	46.2 ↑
R→W	3.6 ↑	1.7↑	-2↓	-1↓	36.1 ↑
E→Stop	1.9 ↑	0.4↑	-2↓	1↑	-17.5↓
S→L	4.6 ↑	0.6↑	-1↓	0	69.6 ↑
R→I	9 ↑	2.1↑	-2↓	-1↓	-27.3↓
E→K	-0.4↓	-1.1↓	0	2↑	26.3 ↑
G→Stop	-1.2↓	-0.6↓	0	0	57.5 ↑
S→Stop	-0.8↓	-0.2↓	-1↓	0	27.8 ↑
D→N	0	0.1↑	-1↓	1↑	8 ↑
S→F	3.6 ↑	0.6↑	-1↓	0	97.3 ↑
D→Y	2.2 ↑	0.2↑	-1↓	1↑	80.2 ↑
R→L	8.3 ↑	1.9↑	-2↓	-1↓	-27.2↓
Q→Stop	1.9 ↑	0.4↑	-1↓	0	-25.6↓

Table 3.6 The changes recorded for hydrophobicity, polarity, charge and volume properties for each of the enriched cancer-associated replacement (from comparison with frequencies in the genetic code).

Amino Acid Replacement	Hydrophobicity (Kyte-Doolittle scale)	Hydrophobicity (Janin scale)	Polarity	Charge	Volume
R→C	7 ↑	2.3↑	-2↓	-1↓	-87.8↓
R→W	3.6↑	1.7↑	-2↓	-1↓	36.1↑
Y→C	3.8↑	1.3↑	-1↓	0	-92.1↓
W→C	3.4↑	0.6↑	0	0	-123.9↓
P→L	5.4↑	0.8↑	0	0	41.8↑

Table 3.7 The changes recorded for hydrophobicity, polarity, charge and volume properties for each of the enriched cancer-associated replacement (from comparison with Blosum62).

Both set of results in table 3.6 and table 3.7 showed highly enriched cancer-associated replacements to exhibit increase in hydrophobicity and decrease in polarity and charge. There are few exceptions to this pattern, but the majority were shown to conform. No specific pattern emerged when analysing the volume differences.

The hydrophobicity change rate between the most and least favoured cancer-associated replacements (based on the comparison with expected replacements frequencies in the

genetic code) is shown in (Figure 3.4). The change is on average much higher in favoured replacements (with replacement residues having bigger hydrophobicity value) when compared to the least favoured replacements. On the other hand, polarity is found to be higher for the least favoured the amino acid replacements (Figure 3.5).

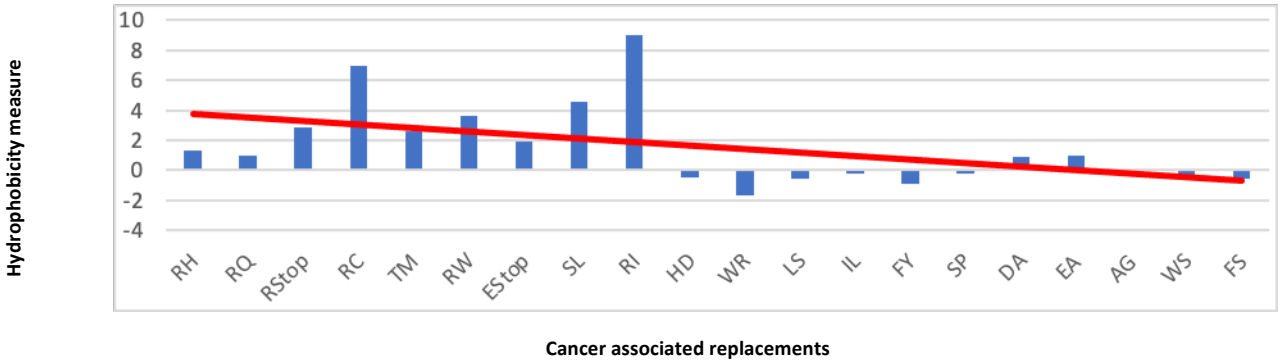

Figure 3.4 The Hydrophobicity (Kyte-Doolittle scale) change rate of cancer associated replacements

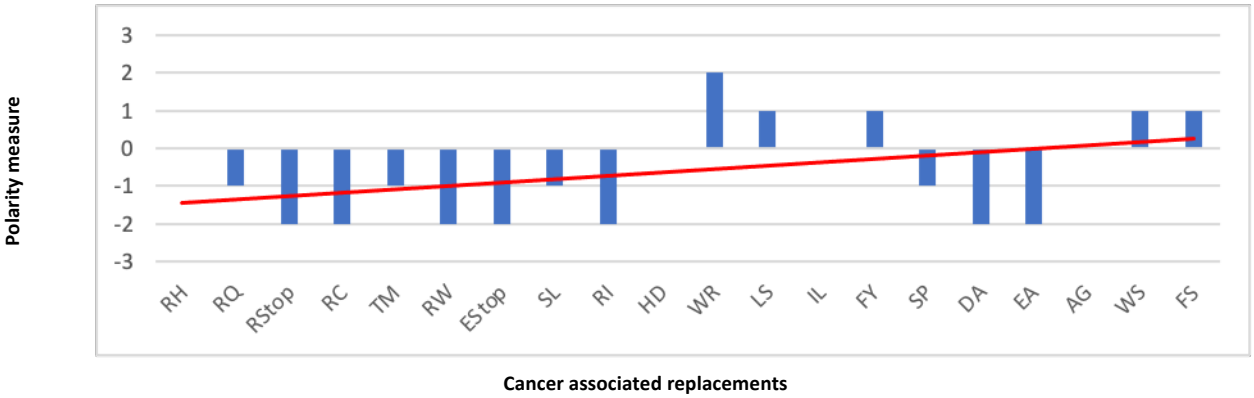

Figure 3.5 The Polarity change rate of cancer associated replacements

3.3.6 Predicting cancer-associated amino acid replacements using physico-chemical and conformational properties

Despite our findings when analysing the physical properties of amino acid replacements, it was not yet clear to what extent these properties can be utilised to discover novel cancer mutations. To answer this, we implemented a machine-learning approach to predict cancer mutations simply using amino acid physico –chemical properties, without using any additional genomic information. The model should only be trained per oncogene/oncoprotein as impact of the replacement location and the effect the change of

the physico-chemical properties per replacement may be significantly different from protein to protein. Here, we chose PTEN protein as an example to apply this technique to. However, the model can be applied to any oncoprotein/oncogene assuming that there are a sufficient number of replacements to produce a reliable model. In our example, the PTEN protein dataset included over 3000 replacements, a third of these were cancer associated. The model's evaluation metrics show a significant difference when comparing the training and the validation datasets if the dataset (number of replacements) was not sufficient.

Several different modelling configurations and algorithms were tested on the data to ensure the selection of the best performing approach. These algorithms include tree-based classifiers such as Gradient boost trees and Random-forests, Neural Networks classifiers such as Keras Slim Residual NN Classifier and Generalized Additive2 Model (the list of these can be found in Appendix A - Table A.1 along with their performance metrics). The performance metric used to rank the models was Area Under Curve (AUC). AUC is an appropriate performance measure when the model is of a binary-classification type. The larger the area under the curve, the more accurate the model is. An AUC of 0.5 suggests that predictions by that model are no better than a random guess. An AUC of 1.0 suggests that the model predictions are perfect. Of course, a model with AUC of 1.0 is an indicator of a flawed set up where some of the data used to train the model are only known after the outcome event and reveal the actual outcome (usually referred to as target leakage). Other performance metrics for our models were also calculated and can be found in the Appendix A - Table A.1. The training dataset was divided to 5 folds where in each iteration the model is trained using 4 folds and validated on the 5^{th} fold. This was repeated so the model was validated on all the datasets. The average AUC of all validated segments is called 'cross validation'. In addition to the validation datasets, 20% of the original dataset was left out of the model training to be used later as an external test (holdout) where the outcome was removed for scoring. The AUC is calculated for all validation and test datasets to ensure that the model is not over-fitting (Table 3.8). This calculation showed close performance across validation and test (holdout) sets where AUC value ranged from 0.67 to 0.72 ensuring no over-fitting.

Scoring Type	Score (AUC)
One Validation	0.72
Cross validation (average of all sets)	0.68
Holdout	0.71

Table 3.8 The AUC calculated for the model validation and holdout segments.

The confusion matrix (Table 3.9) shows the actual versus predicted values for both true/false categories for our training dataset (80% of the total dataset). The model statistics show the model reached just over 76% specificity and 50% sensitivity in predicting cancer mutations. This means that we are able to detect over half of cancer mutations successfully while misclassifying around 24% of non-cancer mutations within the training/validation datasets. The positive predictive value (Precision) was 0.76.

		Predicted		
		-	+	
Actual	-	806 (TN)	252 (FP)	**1058**
	+	784 (FN)	808 (TP)	**1592**
		1590	**1060**	

Table 3.9 The model's Confusion Matrix (where TP is true positives, TN is true negatives, FP is false positives and FN is false negatives)

Both the confusion matrix and the model's AUC (0.72) indicate a moderate performance by our model. Our work here signifies that it is possible to predict cancer mutations from only the location and the amino acid properties. However, a significant number of misclassifications are present and combining this data with other genomic and protein structure related data could yield a stronger more practical model.

The False Positives

Of particular interest are those replacements that were predicted to be cancer-associated but were not yet classed as such in the original training dataset. The dataset used to train the model was extracted from COSMIC Oct -2017. In order to confirm the model's ability to predict cancer-associated mutations, we extracted the somatic missense replacements confirmed implicated in cancer from COSMIC 2022 and compared the list of our false positives to the updated list of cancer-associated mutations. We found that 47% of the replacements in our false positives list are now included in COSMIC. If we only check against replacements that scored > 0.8 then the percentage increases to over 60%. This further confirms the model's ability to predict novel candidate cancer mutations. We recommend considering the false positive replacements provided in the Appendix A - Table A.3 when researching novel mutations as these could be experimentally confirmed later.

3.3.7 Amino acid physico-chemical properties ranked by their impact

The properties used to train the machine-learning model emerged from the original list of features as a result of applying a feature selection procedure. Only features (properties) that are useful for the prediction are selected. These selected features contribute to the calculation of the likelihood scores to different extents. Each property is ranked by its importance in relation to predicting whether the mutation is a cancer associated or not. This importance can be measured by the impact on how much worse a model's error score would be if the model made predictions after randomly shuffling the values of each property (while leaving other values unchanged). Each impact is then normalised, showing the features ranked by their usefulness for the prediction. The impact is normalised so that the value of the most important feature is 100% and the other subsequent features are normalised to it. This process identifies those properties that are particularly important in relation to predicting cancer mutations in our model and would aid in further our understanding of the biological aspects that underline the propensity of a mutation to be a cancer associated mutation.

The location of the substitution is showing to have the highest impact followed by hydrophobicity and solvent accessibility (Figure 3.6). Although other properties are also

important, location of the replacement has several times the impact on the prediction compared to any other property. This could be explained if the early segments of the oncoprotein primary structures play a particular role in functions (e.g., binding) important in carcinogenesis or if the oncoproteins are in general smaller in size compared to other proteins coded by non-associated cancer genes. The results confirm the importance of the hydrophobicity when predicting cancer mutations. Also, solvent accessibility has been shown to have a correlation with hydrophobicity (e.g., small hydrophobic residues will on average have a small accessible surface).

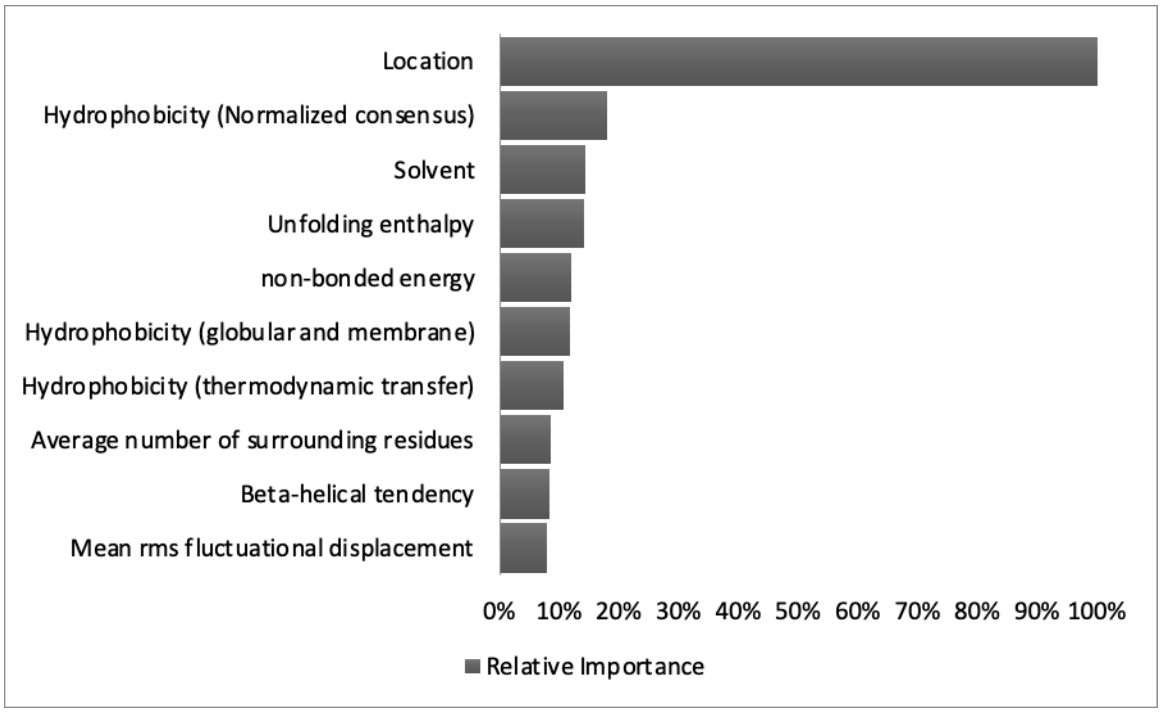

Figure 3.6 The top 10 properties ranked by their relative importance when predicting cancer-associated mutation in our model.

We have also retrained our model using the same dataset while excluding 'location' of the amino acid replacement. We found that a reduction in model's performance was evident. The model trained on this dataset achieved an AUC of 0.62 for the cross validation while the AUC for the holdout was 0.65. The model reached just 76% specificity and 40% sensitivity in predicting cancer mutations (i.e., sensitivity was reduced by 10%). This showed that the model predictability using the physico-chemical properties of the amino acids is still important indicating that information implied by these properties play an important role in tumourogenesis in relation to this specific gene.

Figure 3.7 shows that that the predicted likelihood of a mutation to be cancer-associated on average here is higher by 15% when comparing the replacements with lowest (negative) hydrophobicity change and the replacements with highest increase in hydrophobicity change. For this figure and figure 3.8, the yellow line depicts the marginal effect of this data feature on the target variable after accounting for the average effects of all other predictive features. The orange line with circles depicts how a change in this feature's value, while keeping all other features as they were, impacts a model's predictions. The model result here agrees with our previous findings (figure 3.4 and table 3.6). The model shows a small increase in the likelihood of a mutation to be cancer-associated the higher the increase in hydrophobicity value between the replacement residue and the original residue.

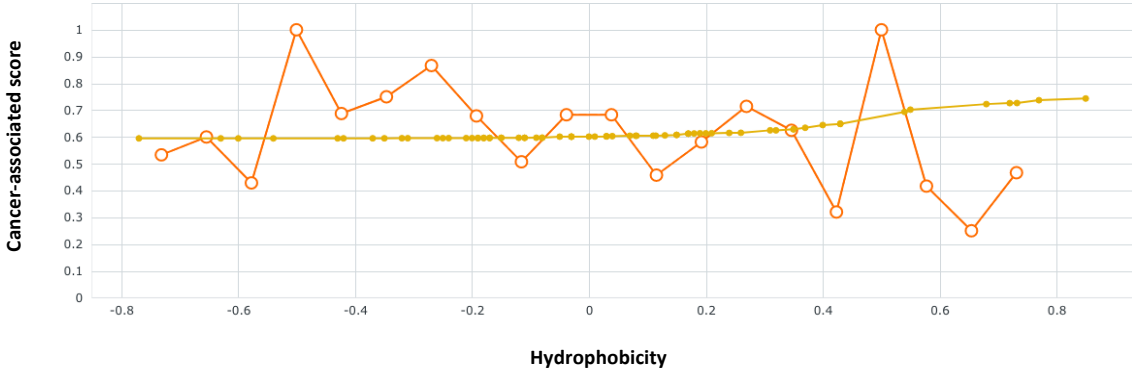

Figure 3.7 Distribution of Hydrophobicity change values in relation to the likelihood of mutation to be cancer-associated

Figure 3.8 illustrates how the position of the replacement in the protein sequence influences the likelihood of a mutation to be cancer associated. The result shows that the likelihood of replacement to be cancer-associated is average decreased by more than 40% when comparing this likelihood for mutations with sequence position < 200 and mutations with sequence position > 2000.

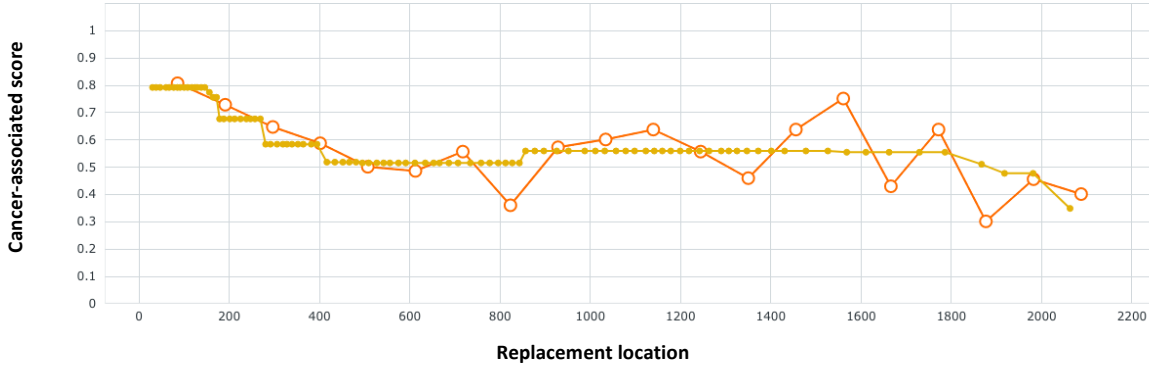

Figure 3.8 Distribution of replacements locations on the protein sequence in relation to their likelihood of being cancer-associated

3.3.8 Comparison with other prediction method

CanSavPre is a machine learning model developed recently (June 2021) and relies on protein structure to predict cancer-related single amino acid variations (*35*). The main limitation of this method comes from the limited availability of protein structures for all isoforms. The study relied on several examples of oncoproteins where the resolved protein structures are available (only certain isoforms were available). However, the model showed great potential in predicting amino acid variations where one of its configurations reached over 89% for accuracy (compared to 64% reached by our model) and 0.81 for F1 score (compared to 0.76 by our model). Unfortunately, and unlike our model, CanSavPre showed significant difference in accuracy between the validation dataset and the independent testing dataset where the high accuracy of 89% was only achieved in the training/validation dataset. The limited number of cases studied and the small number of protein isoforms with a structure available currently prevents CanSavPre from being tested on a wider sample set. Our model does not rely on any data that might not be available for all proteins and so can be used in all instances. However, our model needs to be applied to each oncoprotein individually and the performance may differ from protein to another. Despite the apparent close accuracy scores reported by our model and CanSavPre, this should be treated with some caution. As we only trained our model for one protein (PTEN). Applying our method to a wider range of oncoproteins and including the proteins studied by CanSavPre is required for meaningful comparison. Additionally, a statistical measure needs to be implemented once the accuracy scores are known for the same protein by CanSavPre and our method to determine whether the difference between the scores is significant.

3.4 Novelty of results

As discussed in the literature review (chapter 1 – section 1.2), determining specific amino acid residues and replacements that are enriched in cancer associated mutations could

highlight the specific biological processes crucial in the initiation and progression of many cancer types. We were able to conclude here that that aromatic amino acid group plus Cys are the most enriched amino acids as 'replacement residue' linked to cancer-associated mutations in comparison to all other control groups. This expands the previous view reported in the literature (*1, 28*) to show that all aromatic amino acids should be considered as a highly enriched category in cancer.

These patterns are likely pertained to oncoprotein functions. As aromatic amino acids are often critical for forming protein-nucleotide complexes realised through interactions between aromatic residues and the bases in the nucleotides(*36, 37*), we provided a possible explanation of the highly enriched aromatic amino acid group that highlight the importance of aromatic stacking (necessary to recognize binding sites on DNA or RNA). Furthermore, we linked the prevalence of Cys amino acid in cancer associated mutation to its role in forming disulphide bonds and to Cys oxidation as likely explanation for the enrichment of Cys in cancer associated replacement residues. Our findings narrow down the focus to several protein features such as the aromatic stacking, disulphide bonds and Cys oxidation when investigating cancer associated mutations and their impact on protein functions. Our proposed explanations for the high enrichment of certain amino acid residues provide a missing link in numerous studies that reported on these findings (*1, 3, 5, 28*), connecting these patterns to biological processes related to carcinogenesis.

When investigating the whole replacement (e.g., R → W) rather than the individual amino acids that make the replacement, we showed that there are 17 amino acid replacements highly enriched in cancer-associated mutations. This extended list of highly enriched replacements found in cancer changes the perceived view that only a handful of cancer driver mutations are frequently found in tumors (*1, 28*), emphasising the complex nature of the disease. Our analysis showed that these enriched amino acid replacements in cancer exhibit on average an increase in hydrophobicity and decrease in polarity in comparison to less enriched replacements. This result underlines the different characteristics of these replacements compared to other non-cancer-associated reported replacements.

We demonstrated that physico-chemical properties of amino acids can be used to train a machine-learning model predicting if a replacement is cancer-associated. We showed a

relatively acceptable accuracy can be achieved when building a model predicting these replacements for the PTEN protein. In particular, our model highlights the impact of the position of the replaced amino acid where a significant increase in likelihood of an amino acid replacement to be cancer-associated if it is positioned earlier in the sequence of the protein. It could also be that the early segment of the oncoprotein primarily influences key functions (e.g., binding) that are vital in tumor initiation and progression.

3.5 References

1. Z. A. Szpiech *et al.*, Prominent features of the amino acid mutation landscape in cancer. *PLOS ONE* **12**, e0183273 (2017).
2. S. A. Forbes *et al.*, The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10.11 (2008).
3. C. Greenman *et al.*, Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
4. T. A. P. de Beer *et al.*, Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLOS Computational Biology* **9**, e1003382 (2013).
5. D. Vitkup, C. Sander, G. M. Church, The amino-acid mutational spectrum of human genetic disease. *Genome Biology* **4**, R72 (2003).
6. J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, A. Hamosh, OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789-798 (2015).
7. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
8. S. R. Eddy, Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **22**, 1035-1036 (2004).
9. M. Ashburner *et al.*, Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).
10. Z. Sondka *et al.*, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
11. D. W. Collins, T. H. Jukes, Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics* **20**, 386-396 (1994).
12. J. Janin, Surface and inside volumes in globular proteins. *Nature* **277**, 491-492 (1979).
13. J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105-132 (1982).
14. K. Tomii, M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Peds* **9**, 27-36 (1996).
15. M. M. Gromiha, A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class Information. *Journal of Chemical Information and Modeling* **45**, 494-501 (2005).
16. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
17. J. H. Friedman, Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232 (2001).
18. H. Mi *et al.*, PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* **49**, D394-D403 (2021).
19. Y. Shiroma, R.-U. Takahashi, Y. Yamamoto, H. Tahara, Targeting DNA binding proteins for cancer therapy. *Cancer science* **111**, 1058-1064 (2020).
20. R. Fan *et al.*, The effects of L-arginine on protein stability and DNA binding ability of SaeR, a transcription factor in Staphylococcus aureus. *ScienceDirect* **177**, 105765 (2021).

21. K. Yao *et al.*, The Arginine/Lysine-Rich Element within the DNA-Binding Domain Is Essential for Nuclear Localization and Function of the Intracellular Pathogen Resistance 1. *PloS one* **11**, e0162832-e0162832 (2016).
22. Z. Zhang, S. Witham, E. Alexov, On the role of electrostatics in protein–protein interactions. *Physical Biology* **8**, 035001 (2011).
23. Y. Zheng, Q. Cui, Microscopic mechanisms that govern the titration response and pKa values of buried residues in staphylococcal nuclease mutants. *Proteins: Structure, Function, and Bioinformatics* **85**, 268-281 (2017).
24. J. A. Ubersax, J. E. Ferrell Jr, Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology* **8**, 530-541 (2007).
25. C.-H. Choi, B. A. Webb, M. S. Chimenti, M. P. Jacobson, D. L. Barber, pH sensing by FAK-His58 regulates focal adhesion remodeling. *Journal of Cell Biology* **202**, 849-859 (2013).
26. B. A. Webb *et al.*, A histidine cluster in the cytoplasmic domain of the Na-H exchanger NHE1 confers pH-sensitive phospholipid binding and regulates transporter activity. *Journal of Biological Chemistry* **291**, 24096-24104 (2016).
27. E. L. DiGiammarino *et al.*, A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant p53 tetramer. *Nature Structural Biology* **9**, 12-16 (2002).
28. P. Anoosha, R. Sakthivel, M. Michael Gromiha, Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *ScienceDirect* **1862**, 155-165 (2016).
29. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915 (1992).
30. H. Nishi *et al.*, Cancer Missense Mutations Alter Binding Properties of Proteins and Their Interaction Networks. *PLOS ONE* **8**, e66273 (2013).
31. E. Lee, D. H. Lee, Emerging roles of protein disulfide isomerase in cancer. *BMB reports* **50**, 401-410 (2017).
32. M. E. Ortiz-Soto, S. Reising, A. Schlosser, J. Seibel, Structural and functional role of disulphide bonds and substrate binding residues of the human beta-galactoside alpha-2,3-sialyltransferase 1 (hST3Gal1). *Scientific Reports* **9**, 17993 (2019).
33. R. Mor-Cohen *et al.*, Unique Disulfide Bonds in Epidermal Growth Factor (EGF) Domains of β3 Affect Structure and Function of αIIbβ3 and αvβ3 Integrins in Different Manner. *Journal of Biological Chemistry* **287**, 8879-8891 (2012).
34. M. Popielarski, H. Ponamarczuk, M. Stasiak, C. Watała, M. Świątkowska, Modifications of disulfide bonds in breast cancer cell migration and invasiveness. *American journal of cancer research* **9**, 1554-1582 (2019).
35. J.-J. Liu *et al.*, The structure-based cancer-related single amino acid variation prediction. *Scientific Reports* **11**, 13599 (2021).
36. M. M. Rahman, Z. T. Muhseen, M. Junaid, H. Zhang, The aromatic stacking interactions between proteins and their macromolecular ligands. *Curr Protein Pept Sci* **16**, 502-512 (2015).
37. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Research* **29**, 2860-2874 (2001).

4. Chapter 4: Essentiality, Protein-Protein Interactions and Evolutionary Properties are Key Predictors For Identifying Cancer-associated Genes Using Machine Learning

*** Statement of Authorship**

All the work presented in this chapter including analysis and method development were carried out by Amro Safadi with supervision of his supervisors prof. Simon C. Lovell. and prof. Andrew J. Doig. Both supervisors approved the figures and final wording of the manuscript providing corrections when needed.

4.1 Introduction

The identification of cancer-related genes (referring to both oncogenes and tumor suppressor genes) remains a key challenge. Among all human genes, approximately 3.5% have been directly implicated in cancer initiation and progression (*1*), although it is likely that many remain to be found. Accurate identification of genes potentially related to cancer would provide an opportunity to advance both personalised treatment of cancer and aid drug discovery by providing new targets. The Cancer Gene Census of COSMIC (*1*) provides an expert-curated dataset of cancer-associated genes, relying on tumor sample analysis to identify cancer genes. This provides a high standard in accurately identifying these genes. However, the expert-curation is a lengthy and complex process due to several factors including the availability of tumor samples and the difficulty in sequencing them. Several studies have attempted to build models to identify human disease-related genes. Computational models built using sets of evolutionary and protein network-based properties showed great potential and success in predicting disease genes (*2, 3*). Using protein-protein interactions properties also showed great potential in cancer gene prediction when compared to the frequency of mutations based approach (*4*). However, the goal of accurately predicting cancer genes still eludes us, despite multiple approaches that have been attempted to date.

One viable approach may be to define and enrich the set of properties that characterise these genes and combine these properties to reach a more reliable prediction method. Several characteristics may be correlated with the likelihood of a gene being associated with cancer. A prime candidate is essentiality. A gene is considered essential when loss of its function compromises the viability of an individual (*5*). Essentiality is a quantitative measure and not a simple divide between essential versus non-essential, as defining it as such would

be impossible due to the changeable nature of essentiality based on the genetic and microenvironment context. The identification of essential genes in multiple organisms has provided researchers with vital insights into the mechanisms of biological processes (*6*). For example, essential genes are likely to encode hub proteins in protein–protein interaction networks, signifying more interacting partners than non-essential genes. Furthermore, essential genes are more likely to be abundantly and ubiquitously expressed in cells and tissues and have smaller-sized introns (*7*). Also, several studies determined the relationship between evolutionary conservation and the degree of essentiality in genes with variations in findings across species (*7*). The general findings in human genes point to a relationship whereby the more essential the gene is, the less likely it is to show enrichment of missense mutations. In contrast, the number of synonymous mutations is not dependent on essentiality. This indicates that purifying selection acts more stringently on essential genes (*5, 8*).

One could argue that genes implicated in driving and initiating tumors, which generally do not compromise viability in a direct manner, are thus unlikely to score high on the essentiality spectrum. However, there are indications that human genes associated with genetic disease are likely to be essential (*6*). Cassa et al (*8*) investigated heterozygous protein-truncating variants in over 60,000 individuals from the Exome Aggregation Consortium (ExAC) dataset (*9*) using the 'shet' essentiality score (a metric that provides Bayesian estimates of the selection coefficient against heterozygous loss-of-function variation) and were able to predict phenotypic severity, age of onset and penetrance for Mendelian disease-associated genes. In addition, genes involved in neurological phenotypes, including autism, congenital heart disease and inherited cancer risk, seem to be under more intense purifying selection, which may indicate essentiality. Overall, quantitative estimates of essentiality appear to be particularly useful in Mendelian disease gene discovery efforts.

Here, we identify combinations of gene properties that have not been previously used to assess the likelihood of a gene to be cancer-associated. We study whether cancer-associated genes are more likely to be essential than non-cancer genes and check whether an uplift in predicting a gene to be a cancerous can be achieved by using essentiality-related

properties. These findings might also indicate if these genes are more likely to be under stronger selection than other non-cancer related genes. We were able to build a relatively accurate machine-learning model predicting cancer genes using essentiality-related properties. Using this machine learning approach, we were able to identify further candidate genes for cancer, in addition to those currently reported in COSMIC census (October 2018).

4.2 Materials and Methods

4.2.1 Datasets

A total list of 18,000 human protein-coding genes and various properties (focusing on essentiality) were obtained by combining data from different data sources and data obtained from previous studies (*3, 5*). Below are the different sources of data used:

4.2.1.1 Essentiality scores

We obtained several different essentiality scores calculated for human genes from (*5*) to use in our dataset. Petrovski's 'residual variation intolerance score' (RVIS) (*10*) and Rackham's EvoTol (*11*) relate the amount of common loss-of-function variation to that of the total gene variation. Other scores are based on the work of Samocha et al. (Missense Z-score)(*12*), which sets up a baseline expectation of mutation count per gene based on the sequence context, local mutation rate, sequencing depth and, most importantly, sample size. Fadista's LoFtool (*13*) combines the neutral mutation rate of Samocha et al. and the evolutionary information in EvoTol. The baseline neutral expectation is compared with the observed counts of loss-of-function variants in the Missense Z-score, in Bartha's probability of haploinsufficiency (Phi) (*14*) and in Lek's probability of loss-of-function intolerance (pLI) (*15*). Finally, recent work by Cassa et al. (*8*) describes a metric (shet) that provides Bayesian estimates of the selection coefficient against heterozygous loss-of-function variation. The various scores were developed or updated using the Exome Aggregation Consortium (ExAC) sample of 60,706 human exomes described in (*15*). These scores show high correlations with one another (*5*).

4.2.1.2 *Evolutionary profile and genomic related properties*

We used gene properties provided and constructed in (*3*) including genomic location, protein network parameters and summary statistics of neutrality for human genes.

The genomic location properties we used in our work were: Chr, Start, End and Strand and additionally dN/dS values that indicate neutrality and selection pressure (multiple species). All were extracted from Ensembl Biomart Genes (*16*).

Group property divides genes into three different mutually excluding groups: (i) Complex-Mendelian (CM) genes, (ii) Mendelian Non-Complex (MNC) genes, and (iii) Complex Non-Mendelian (CNM) genes.

Data also include measures of genetic variation at intra-species level and measures for proportion of rare variants, such as Tajima's D exons, Tajima's D regulatory, Fay and Wu's H exons and Fay and Wu's H regulatory (*3*).

4.2.1.3 *Protein network properties*

The human protein–protein interaction network (PIN) was reconstructed from the interactions available in the BioGRID database version 3.1.81 (*17*). Properties such as degree were computed as the total number of interactions in which a protein is involved, while betweenness and closeness centralities were computed using the NetworkX Python library (*18*).

4.2.1.4 *General gene properties*

We enriched the dataset with general gene properties in addition to the properties compiled from the sources above. These properties were directly extracted from Ensembl Biomart Genes (*16*) such as Gene % GC content, Transcript count, Gene Length, while some were calculated, such as StdDev Transcript length, Average Transcript length, Min Transcript length, Max Transcript length and Exon Count. Also, a list of all human Ohnolog genes with strict and intermediate score was downloaded from this database: (http://ohnologs.curie.fr/).

4.2.1.5 Outcome

To build a supervised machine-learning model, we need to identify what the model is trying to predict and add that outcome to every row in our dataset. The outcome in our dataset is binary (true or false), indicating if this gene has been identified as a cancer gene. We did this by identifying if this gene has been added to the COSMIC 's Cancer Gene Census.

The properties we have used to construct our dataset are not inclusive of all possible features that can relate to genes essentiality. Other studies carried out on mice investigated an extended list of essentiality properties, where the subset of features we selected here was shown to be of particular interest (*19*). Expanding the number of properties used would be an option to explore in the future.

4.2.2 Machine learning method

This is identical to the method discussed in chapter 3 (section 3.2.5). The only exception is that the binary prediction classes here are reflecting whether the individual record (containing data related to a single gene) is cancer associated (value =1) or non-cancer associated (value = 0) and these were discussed in section 4.2.1.5.

4.3 Results

4.3.1 Cancer-associated genes and essentiality scores

We first determined whether cancer-related genes are likely to have high essentiality scores. We aggregated several essentiality scores calculated by multiple metrics (*5*) for the list of genes identified in the COSMIC Census database (Oct 2018) and for all other human protein coding genes. Two different approaches to scoring genes' essentiality are available. The first group of methods calculates the essentiality scores by measuring the degree of loss of function caused by a change (represented by variation detection) in the gene. It uses the following methods: residual variation intolerance score (RVIS), LoFtool, Missense-Z, the probability of loss-of-function intolerance (pLI) and the probability of haplo-insufficiency (Phi). The second group (Wang, Blomen and Hart- EvoTol) studies the impact of variation on cell viability. For all methods above measuring essentiality, a higher score indicates a higher degree of essentiality and each method is described in detail in (*5*).

We find that on average the cancer genes exhibit a higher degree of essentiality compared to the average scores calculated for all protein coding human genes and all metrics. We find that genes associated with cancer have higher essentiality scores on average in both categories (intolerance to variants and cell line viability) compared to the average scores across all human genes. P-values consistently < 0.00001 (Table 4.1).

We also investigated whether Tumor Suppressor Genes (TSGs) as a distinct group of genes would show different degrees of essentiality. We found that no significant difference in the degree of essentiality on average for that group compared to the set of all cancer genes (Table 4.1).

Method	Mean Essentiality Score for all genes	Mean Essentiality Score for cancer genes	Ratio (cancer genes/all genes)	P-Values	Mean Essentiality Score for TS genes	Ratio (TS genes/all genes)	P-Values
Phi	0.27	0.63	2.34	< .00001	0.58	2.17	< .00001
Wang	0.42	0.62	1.48	< .00001	0.65	1.55	< .00001
S_het	0.06	0.12	2.07	< .00001	0.12	2.07	< .00001
LofTool	0.50	0.70	1.40	< .00001	0.71	1.42	< .00001
Missense-Z	0.69	1.86	2.70	< .00001	1.85	2.69	< .00001
RVSI	50.0	68.3	1.37	< .00001	68.9	1.38	< .00001

Table 4.1 The comparison between the mean essentiality scores of cancer genes and all other human genes.

The results are particularly of interest in the context of cancer, as essential genes have been shown to evolve more slowly than nonessential genes (*20-22*) although some conflicts have been reported (*22*). A slower evolutionary rate indicates less probability to evolve resistance to a cancer drug. This is particularly important in the case of anticancer drugs as it was reported that these drugs cause a change in the selection pressure when administrated, leading to increased drug resistance (*23*).

4.3.2 Cancer-associated genes prediction analysis results

This association between cancer-related genes and essentiality scores prompted us to develop methods to identify cancer-related genes using this information. We used a machine-learning approach, a range of open-source algorithms were applied and tested to produce the most accurate classifier. We focus on properties related to protein-protein interaction networks, as essential genes are likely to encode hub proteins, i.e., those with highest degree in the network (*21, 24*).

A total of 9 different modelling approaches (or configurations) were run on the data to ensure the selection of the best performing approach (the list of these can be found in Appendix A - Table A.2 along with their performance metrics). The performance metric used to rank the models was Logarithmic Loss (LogLoss), LogLoss is an appropriate and known performance measure when the model is of a binary-classification type. The LogLoss measures confidence of the prediction and estimate how to penalize incorrect classification. The selection mechanism for the performance metric takes the type of model (binary classification in this case) and distribution of values into consideration when recommending the performance metric. However, other performance metrics were also calculated and can be found in the Appendix A - Table A.2. The performance metrics are calculated for all validation and test (holdout) sets to ensure that the model is not over-fitting (4.2). The particular model with best performance result (LogLoss) in this case was: eXtreme Gradient Boosted Trees Classifier with Early Stopping. The model shows very close LogLoss values for training/validation and holdout (20% of the data was left out of the model training and validation datasets to be used as a blind test) data sets ensuring no over-fitting.

Scoring Type	Score (LogLoss)
One Validation	0.097
Cross validation (average of all sets)	0.098
Holdout	0.099

Table 4.2 The LogLoss scores for our model validations and holdout segments.

The model development workflow (i.e., the model blueprint) is shown in figure 4.1:

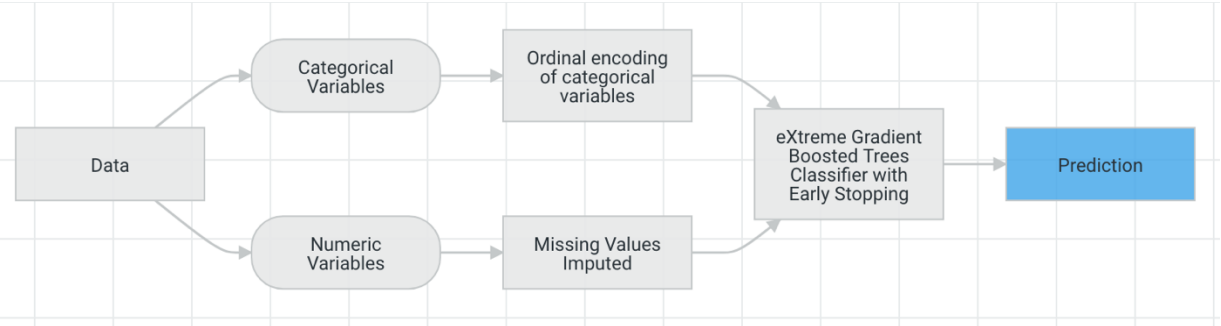

Figure 4.1 Model development stages.

The model blueprint (figure 4.1) shows the pre-processing steps and the algorithm used in our final model and illustrates the steps involved in transforming input into a model. In this diagram, 'Ordinal encoding of categorical variables' converts categorical variables to an ordinal scale while the 'Missing Values Imputed' node imputes missing values. Numeric variables with missed values were imputed with an arbitrary value (default -9999). This is effective for tree-based models, as they can learn a split between the arbitrary value (-9999) and the rest of the data (which is far away from this value).

To demonstrate the effectiveness of our model, a chart was constructed (figure 4.2) that shows across the entire validation dataset (divided into 10 segments or bins and ordered by the average outcome prediction value) the average actual outcome (whether gene has been identified as cancer gene or not) and the average predicted outcome for each segment of the data (order from lowest average to highest per segment). The left side of the curve indicates where the model predicted a low score on one section of the population while the right side of the curve indicates where the model predicted a high score. The "Predicted" blue line displays the average prediction score for the rows in that bin. The "Actual" red line displays the actual percentage for the rows in that bin. By showing the actual outcomes alongside the predictive values for the dataset, we can see how close these predictions are to the actual known outcome for each segment of the dataset. Also, we can determine if the accuracy diverges in cases where the outcome is confirmed cancer or when it is not, as the segments are ordered by their average of outcome scores.

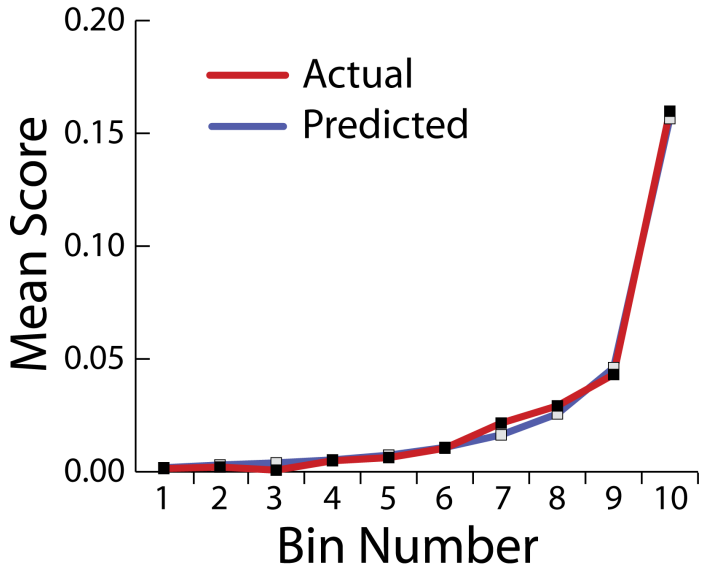

Figure 4.2 The Lift Chart illustrating model's accuracy.

In general, the steeper the "actual" line is, and the more closely the "predicted" line matches the actual line, the better the model. A close relationship between these two lines is indicative of the predictive accuracy of the model; a consistently increasing line is another good indicator of satisfactory model performance. The graph we have for our model indicates strong accuracy of our prediction model.

Moreover, the confusion matrix (Table 4.3) and the summary statistics (Table 4.4) show the actual versus predicted values for both true/false categories for our training dataset (80% of the total dataset). The model statistics show the model reached just over 89% specificity and 60% sensitivity in predicting cancer genes. This means that we are able to detect over half of cancer genes successfully while only misclassifying around 10% of non-cancer genes within the training/validation datasets. The summary statistics (Table 4.4) also shows the F1 score (harmonic mean of the precision and recall) and Matthews Correlation Coefficient (MCC is the geometric mean of the regression coefficient) for the model. The low F1 score reflects our choice to maximise the true negative rate (preventing significant misclassification of non-cancer genes).

		Predicted		
		-	+	
Actual	-	12493 (TN)	1490 (FP)	**13983**
	+	159 (FN)	243 (TP)	**402**
		12652	1733	

Table 4.3 The model's Confusion Matrix (where TP is true positives. TN is true negatives. FP is false positives. FN is false negatives)

F1 Score	True Positive Rate (Sensitivity)	False Positive Rate (Fallout)	True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Matthews Correlation Coefficient
0.23	**0.61**	**0.11**	**0.89**	**0.14**	**0.99**	**0.89**	**0.25**

Table 4.4 Summary of the model's performance statistics

The False Positives

To further confirm the model's ability to predict cancer genes, we used the model on 190 new cancer genes that had been added to the COSMIC' Cancer Census Genes between October 2018 and April 2020. Applying the model, we were able to predict 56 genes out of the newly added 190 genes as cancer genes, all of which were among the false positives detected by the model. This indicates that the model is indeed suitable to use to predict novel candidate cancer genes that could be experimentally confirmed later.

Another way to visualise the model performance and determine the optimal score to use as a threshold between cancer and non-cancer genes is the 'prediction distribution' graph (Figure 4.3) that illustrates the distribution of outcomes. The distribution (in purple) shows the outcome where gene is not classified as cancer gene while the second distribution (in green) shows the outcomes where gene is classified as cancer gene. The dividing line represents the selected threshold at which the binary decision is optimal (creating a desirable balance between true negatives and true positives). Figure 4.3 shows how well our model discriminates between prediction classes (cancer gene or non-cancer gene) and

shows the selected score (threshold) that could be used to make a binary (true/false) prediction for a gene to be classified as a candidate cancer gene. Every prediction to the left of the dividing line is classified as non-cancer associated and every prediction to the right of the dividing line is classified as cancer associated.

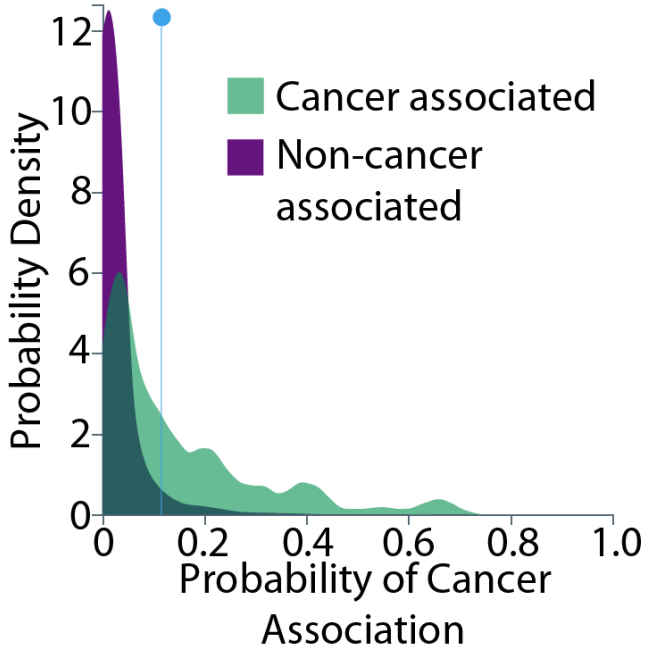

Figure 4.3 The prediction distribution graph showing how well the model discriminates between cancer and non-cancer genes.

The prediction distribution graph can be interpreted as follows: purple to the left of the threshold line, is for instances where genes were correctly classified as non-cancer (true negatives). Green to the left of the threshold line is for instances were incorrectly classified as non-cancer (false negatives). Purple to the right of the threshold line, is for instances were incorrectly (according to the current training/validation dataset) classified as cancer gene (false positives). Green to the right of the threshold line, is for instances were correctly classified as cancer genes (true positives). The graph again confirms that the model was able to accurately between cancer and non-cancer genes.

Using the receiver operating characteristic curve (ROC) curve produced for our model (Figure 4.4), we were able to evaluate the accuracy of prediction. The AUC (area under the curve) is a metric for binary classification that considers all possible thresholds and

summarizes performance in a single value, with the larger the area under the curve, the more accurate the model. An AUC of 0.5 suggests that predictions based on this model are no better than a random guess. An AUC of 1.0 suggests that predictions based on this model are perfect, (this is highly uncommon and likely flawed indicating some features that should not be known in advance are being used in model training and thus revealing the outcome). As the area under the curve is of 0.86, we conclude that the model is accurate. The circle intersecting the ROC curve represents the threshold chosen for classification of genes. This is used to transform probabilities scores assigned to each gene into binary classification decision where each gene would be classified into potential cancer gene (true) or not cancer gene (false).

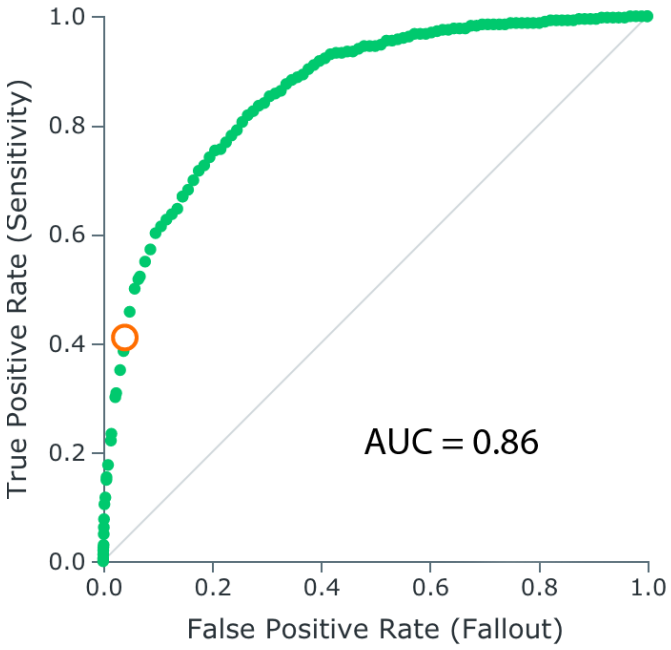

Figure 4.4 The receiver operator characteristic (ROC) curve indicating model performance

Feature Impact

Feature impact measures how much worse a model's error score would be if the model made predictions after randomly shuffling the values of one field input (while leaving other values unchanged) and thus shows how useful each feature is for the prediction. The scores were normalised so that the value of the most important feature column is 100% and the other subsequent features are normalised to it. This helps identify those properties that are

particularly important in relation to predicting cancer gene in our model and would aid in further our understanding of the biological aspects that might underline the propensity of a gene to be a cancer gene.

'Closeness' and 'degree' are ranked as the properties with the highest feature impact (4.5). Both are protein–protein interaction network properties, indicating a central role of the protein product within the network. We find that both correlate with likelihood of cancer association., Other important properties such the 'phi' essentiality score (probability of haploinsufficiency compared to baseline neutral expectation) and Tajima's D regulatory (measures for genetic variation at intra-species level and for proportion of rare variants) show that increased essentiality accompanied with occurrence of rare variants increase the likelihood of pathological impact and for the gene to be linked to cancer initiation or progression. We also note that greater length of a gene or transcript increases the likelihood of a somatic mutation, so increasing the chance of a mutation within that gene, thus increasing the likelihood of it being a cancer gene.

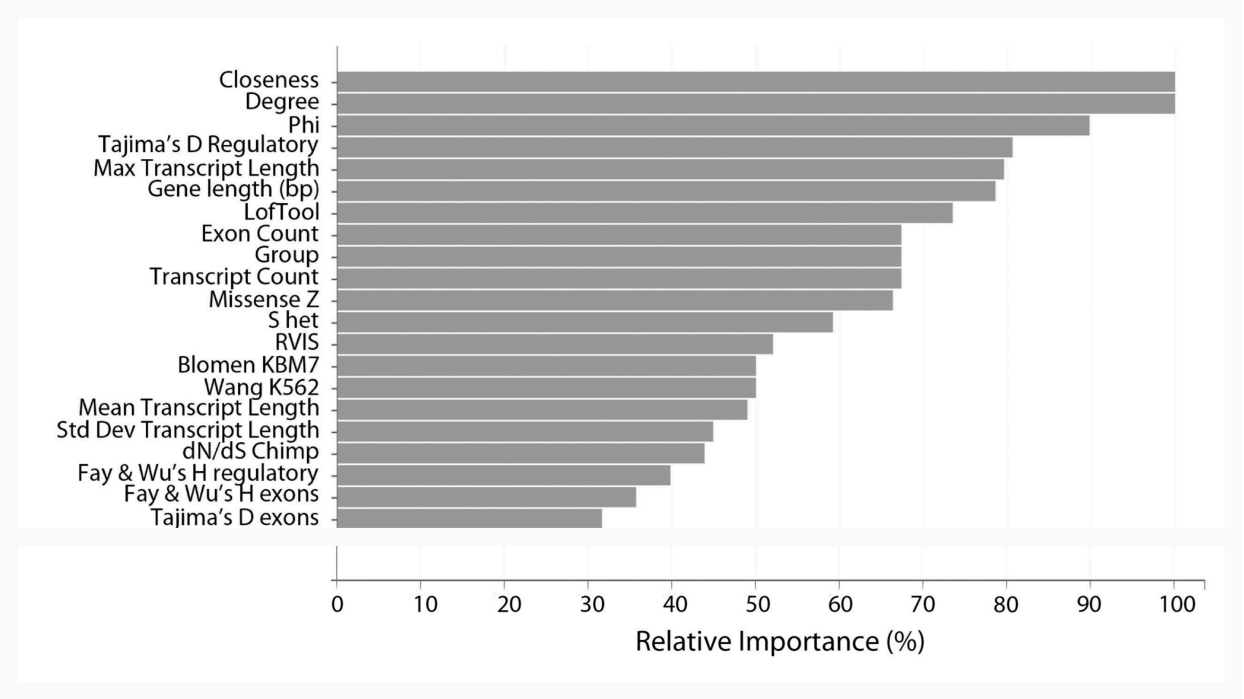

Figure 4.5 The top properties ranked by their relative importance used to make the predictions by the model

To confirm that the selected model performance is optimal based on the input data used, we created a new blended model combining the best 2nd and 3rd modelling approaches from all modelling approaches tested within our project and compared the performance metric (AUC) of our selected model with the new blended model. We found that improvement (despite the added complexity) is small (0.008) where the blended model achieved an AUC of 0.866 and our selected model achieved an AUC of 0.858.

We have also retrained our model using a dataset that excludes general gene properties as listed in the 'Data Sets' section and found that a reduction in model's performance was evident but very small. The model trained on this dataset achieved an AUC of 0.835 and a sensitivity of 55% at a specificity of 89%. This small reduction in the predictability of the models indicates that essentiality and protein-protein interaction network properties are the most important features predicting cancer gene and that information carried by gene general properties can be in most part be represented by information carried by these properties. This can be rationalised, as longer genes (median transcript length =3737) tend to have the highest number of protein–protein interactions (*25*).

4.3.3 Comparison with other cancer driver genes prediction methods

According to a recent comprehensive review of cancer driver genes prediction models, currently the best performing machine learning model is driverMaps with AUC= 0.94 followed by HotNet2 with AUC=0.81 (27). When comparing our model performance using AUC to the other 12 reviewed cancer driver genes prediction models, our model would come second with AUC= 0.86. Our predictive model achieved better AUC measured performance when compared to the top-performing model using similar network based approach (HotNet2 with AUC=0.81) and better than the best function-based prediction model (MutPanning with AUC=0.62). The strong performance of our model based on AUC of the ROC graph indicates the importance of combining different and distinctive gene properties when building prediction models while avoiding reliance on the frequency approach that could mask important driver genes that were detected in fewer samples. Despite the apparent success and high AUC score reported by our model, this should be treated with some caution. The AUC value is based on the ROC curve which is constructed

by varying the threshold and then plotting the resulting sensitivities against the corresponding false positive rates. Several statistical methods are available to use to compare two AUC results and determine if the difference is significant (*26-28*). These methods require the ranking of the variables in its calculations (e.g., to calculate the variance or covariance of the AUC). The ranking of predicated cancer associated genes was not available from all the other 12 cancer driver genes prediction methods. Thus, we were not able to measure whether the difference between the AUC score of our method and the AUC scores of these methods is significant.

4.3.4 The Cancer genes association with WGD and Ohnologs

Enriching the model's training dataset with added properties that show correlation with oncogenes could enhance the model prediction ability and elevate further the accuracy of the model. One potential feature is knowing whether a gene is an ohnolog gene.

Paralogs retained from whole genome duplications (WGD) events have occurred in all vertebrates (two rounds of WGDs) some 500 MY ago are called 'ohnologs' after Susumu Ohno (*29*). Ohnologs have been shown to be prone to dominant deleterious mutations and frequently implicated in cancer and genetic diseases (*29*). We investigated the enrichment of ohnologs within cancer-associated genes. Ohnolog genes can be divided into three sets: strict, intermediate, and relaxed. These three sets are constructed using statistical confidence criteria (*29*) . We found that 44% of the total number of cancer-associated genes (as reported in COSMIC census) belongs to an ohnologs family (using strict & intermediate thresholds). Considering that 20% of all known human genes are ohnologs (strict & intermediate) and the ratio of cancer-associated genes makes less than 4% of all human genes, the enrichment of ohnolog genes with cancer-related genes is 2 times higher than expected. If only ohnologs that pass the strict threshold were considered, the fraction of cancer-related genes that are ohnologs is still high at 34%. This association between oncogenes and genes retained from the whole genome duplication events (ohnologs) could potentially added as a supplementary feature in our model (e.g., a feature indicating if the gene is an ohnologs). Enriching our training dataset with this feature could potentially increase the model accuracy further.

4.4 Novelty of results

Here I was able to I contribute to the on-going efforts highlighted in sections 1.3.2 and 1.4,4 in the literature review in predicting cancer associated genes. I showed that combining various properties of cancer genes, including evolutionary related measures such as selection pressure and measures of genetic variants, to train a machine-learning model to identify cancer related genes could result in superior model performance. One property that was not investigated before in relation to predicting cancer associated genes is the essentiality of the gene. We found that the cancer-associated genes exhibit a higher degree of essentiality compared to the scores calculated for all protein coding human genes. Our results could be interpreted in the context of the genes' involvement in particular biological activities. Genes classed as essential are often involved in cell, embryo, and organism growth. Similarly, proliferation is key for cancer cells. Therefore, the sets of genes that are essential and those that are involved in unregulated growth, as seen in cancer, tend to overlap. This finding provides further evidence for the importance of evolutionary aspects when studying cancer genes. Scientists might be able to further the understanding of cancer by incorporating properties linked to essentiality in their studies.

We trained a machine-learning model (a classifier) using a distinctive blend of gene properties to measure the likelihood of a protein coding gene to be cancer associated. Our dataset included general gene properties like gene % GC content and transcript count. Also added were protein-protein interactions properties and various measures indicating selection pressure and essentiality scores. Protein-protein interactions properties were confirmed to be very influential when assessing the likelihood of a gene to be cancer-associated. Our model also showed that the essentiality score Phi and Tajima's D Regulatory to have the largest effect. These findings may offer targets for further research.

Our model was able to produce a novel list of candidate genes predicted to be cancer-associated providing a good basis for scientists to prioritise these genes in their research.

4.5 References

1. S. A. Forbes *et al.*, The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10.11 (2008).
2. N. López-Bigas, C. A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108-3114 (2004).
3. N. Spataro, J. A. Rodríguez, A. Navarro, E. Bosch, Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum Mol Genet* **26**, 489-500 (2017).
4. X. Shi *et al.*, Comprehensive evaluation of computational methods for predicting cancer driver genes. *Briefings in Bioinformatics* **23**, bbab548 (2022).
5. I. Bartha, J. di Iulio, J. C. Venter, A. Telenti, Human gene essentiality. *Nat Rev Genet* **19**, 51-62 (2018).
6. D. Park, J. Park, S. G. Park, T. Park, S. S. Choi, Analysis of human disease genes in the context of gene essentiality. *Genomics* **92**, 414-418 (2008).
7. B. Georgi, B. F. Voight, M. Bućan, From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet* **9**, e1003484 (2013).
8. C. A. Cassa *et al.*, Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics* **49**, 806-810 (2017).
9. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
10. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics* **9**, e1003709 (2013).
11. O. J. Rackham, H. A. Shihab, M. R. Johnson, E. Petretto, EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res* **43**, e33 (2015).
12. K. E. Samocha *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).
13. J. Fadista, N. Oskolkov, O. Hansson, L. Groop, LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* **33**, 471-474 (2017).
14. I. Bartha *et al.*, The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. *PLOS Computational Biology* **11**, e1004647 (2015).
15. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
16. F. Cunningham *et al.*, Ensembl 2015. *Nucleic Acids Res* **43**, D662-669 (2015).
17. C. Stark *et al.*, The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704 (2011).
18. A. A. Hagberg, D. A. Schult, P. J. Swart, *Exploring Network Structure, Dynamics, and Function using NetworkX*. G. e. Varoquaux, T. Vaught, J. Millman, Eds., Proceedings of the 7th Python in Science Conference\ (Pasadena, CA USA\, 2008), pp. 11 - 15\.
19. M. Kabir, A. Barradas, G. T. Tzotzos, K. E. Hentges, A. J. Doig, Properties of genes essential for mouse development. *PLoS One* **12**, e0178273 (2017).

20. I. K. Jordan, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12**, 962-968 (2002).
21. H. B. Fraser, D. P. Wall, A. E. Hirsh, A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol* **3**, 11 (2003).
22. C. Pál, B. Papp, L. D. Hurst, Genomic function: Rate of evolution and gene dispensability. *Nature* **421**, 496-497; discussion 497-498 (2003).
23. D. Sun, S. Dalin, M. T. Hemann, D. A. Lauffenburger, B. Zhao, Differential selective pressure alters rate of drug resistance acquisition in heterogeneous tumor populations. *Scientific Reports* **6**, 36198 (2016).
24. D. P. Wall *et al.*, Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A* **102**, 5483-5488 (2005).
25. I. Lopes, G. Altab, P. Raina, J. P. de Magalhães, Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Front Genet* **12**, 559998 (2021).
26. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).
27. K. Molodianovitch, D. Faraggi, B. Reiser, Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biom J* **48**, 745-757 (2006).
28. A. Hart, Mann-Whitney test is not just a test of medians: differences in spread can be important. *Bmj* **323**, 391-393 (2001).
29. P. P. Singh, J. Arora, H. Isambert, Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput Biol* **11**, e1004394 (2015).

5. Chapter 5: Discussion

Cancer is a complex disease; research that worked on providing the genomic profile of the disease is still producing new findings. The number of genes implicated in carcinogenesis is constantly increasing along with the number of mutations that are instigating the cascade of steps necessary for the transformation of normal cells into carcinogenic cells. For example, the number of genes implicated in cancer increased by over 30% in the last 4 years as per the COSMIC genes census (*1*). Cancer as a disease has a distinct nature when compared to other genetic diseases. Whole genome sequencing of tumor samples showed that single nucleotide variants in tumor cells can be two to three orders of magnitude more abundant than variants found in adjacent normal cells (*2*). The biological effect of these mutations is often non-catastrophic, as for a tumor to form and progress, several mutations may need to be present. Characteristics known about deleterious mutations found in the human population and those reported for other genetic diseases are not likely to be similar to cancer-associated mutations and their genes.

Identifying how cancer-associated genes and mutations differ from non-cancer related mutations and genes would allow us a better understanding of the disease and could enhance the usage of advanced prediction analysis in finding mutations that trigger tumor initiation and progression (drivers). Studying mutation characteristics at the protein level is particularly useful in understanding their effect on the downstream protein products and reveals the link between the genetic variants and the effect on the biological processes involved in carcinogenesis.

Using machine learning techniques has a great potential in producing accurate predictive models that have the ability to indicate the likelihood of a gene or a mutation to be cancer-associated, highlighting at the same time the relationship between the properties used in the model and the prediction. Machine learning techniques are often superior to traditional statistical methods because they are more flexible and rely on fewer statistical assumptions. The only assumption being made is that the model training data is representative of the future scoring data.

This chapter summarises the overall findings of this thesis and discusses how the results are interpreted, their implications and how they can be utilised in the cancer research field, highlighting any limitation.

We started investigating the impact missense point cancer-associated mutations have on protein stability. Stability is one of the most important measures that can indicate the effect a deleterious mutation has on the protein (*3*). Several previous studies reported the destabilising effect of this type of mutation on the protein structure and the association of this effect with other genetic diseases (*4*). We were able to utilise the availability of protein structures for proteins produced by some of the most frequent oncogenes. We studied the consequences of the cancer associated missense amino acid replacements, known for each of these genes, on the stability of the protein products. We showed that unlike most of other non-cancer deleterious replacements, cancer-associated replacements exhibit on average a neutral to stabilising effect on the protein with some exceptions. Despite doing this analysis for highly frequently mutated genes found in tumor samples, there is still the drawback of not being able to process all oncoproteins due to a lack of resolved 3-dimensional protein structures. As more advanced artificial intelligence systems are being developed providing accurate predictions for the needed structural data, it will be possible to expand this work and confirm the distinct effect on stability many cancer-associated mutations exhibit. A recent example is the AlphaFold program (*5*) that became available publicly after our work was completed. AlphaFold is showing excellent prediction performance in solving protein structures.

Nonetheless our result in chapter 2, indicated a distinct impact on the stability of the protein by some amino acid replacements in cancer-associated genes, prompted us to investigate the spectrum of amino acid replacements found implicated in cancer and their properties. To further confirm the presence of distinct patterns, we identified favoured amino acids that feature at a high frequency in cancer mutations when compared against other control groups, such as expected probabilities based on the genetic code table or Blosum62 (*6*). This approach provided several advantages compared to determining the enrichment purely using mutation rate (the number of times the same mutation is found in different tumor samples). This method allowed us to avoid potential bias in the finding

resulting from unbalanced number of tumor samples belonging to different cancer types and identify the enriched ratios of amino acids in cancer-associated mutation through comparisons with their expected frequencies under no selection (genetic code), natural selection (Blosum62) and selection applied on deleterious mutations in other genetic diseases (ExAC). Tryptophan (Trp) and Cysteine (Cys) were the most frequently replaced amino acids in cancer-associated mutations compared to the expected frequency based on the genetic code. Trp and Cys featured also in the most frequent replacements when compared to Blosum62. Notably, Cys and Trp are the least likely amino acids to be replaced within Blosum62, which in addition to being less abundant could also corroborate their key position in conserved protein regions. We found that all aromatic amino acids have an enrichment > 1.2 when comparing replacement residues in cancer-associated mutations to their frequencies based on genetic code. Moreover, Phe and Tyr have an enrichment > 1.5 when comparing replacement residues in cancer-associated mutations to missense mutations from other genetic diseases. Thus, we conclude that aromatic amino acid group plus Cys are the most enriched amino acids as 'replacement residue' linked to cancer-associated mutations in comparison to all other control groups. This expands the previous view to include all aromatic amino acids in the highly enriched residues category in cancer (*7, 8*). We also confirmed that Arg is the most likely amino acid to mutate in cancer-associated mutations, in terms of absolute frequencies, despite Ser, Leu, Ala and Gly being substantially more abundant in proteins (*7-9*).

Supposing that these patterns pertain to the cancer genes molecular functions, we examined the cancer-associated genes ontology enrichment. By analysing the molecular functions of all 590 genes implicated in cancer in our study as per COSMIC Oct – 2017, we confirmed they are frequently associated with binding activity. We noticed a wide spread of involvement in 'Bindings' functions at Fold Enrichment >1.35 in all types of binding. In some binding functions such as 'damaged DNA' binding, the enrichment was > 9. Out of 590 cancer-associated genes analysed 80 were involved in RNA binding with fold enrichment =9.81 and 228 genes in DNA binding types with fold enrichment =3.24. We hypothesised a link between the oncogenes affinity to binding activity in cancer and the amino acid residues and replacements found enriched in cancer-associated mutations.

Trp and the other aromatic amino acids are often essential in interactions with non-protein ligands (frequently via stacking interactions). In particular, they are often critical for forming protein-nucleotide complexes realised through interactions between aromatic residues and the bases in the nucleotides. Aromatic stacking (involving Trp for example) is necessary to recognize binding sites on DNA or RNA. Moreover, this aromatic stacking is involved in the process of mismatch repair; strand separation, degradation and RNA cap binding (*10*). These interactions can explain the high enrichment of Trp in cancer mutations and indeed the high enrichment in general of the aromatic amino acids Phe, Tyr and His (semi aromatic) as the replacement residues.

Cys can be involved in forming disulphide bonds, particularly in extracellular proteins. These bonds are known to stabilize the protein structure (*11*). Also, disulphide bonds were shown play a key role in proteins and enzymes that stimulate cell proliferation (*12*), in particular, they were shown to affect receptors regulating cellular growth and proliferation altering their functions to be constitutively activated (*13, 14*). If a mutation occurs with Cys as the replacement residue, then there is a chance that an inter-molecular disulphide bond would form altering the receptors and enzymes regulating cellular growth and causing the constant 'turned on' state in some elements within the signalling pathways instituting the uncontrolled proliferation of the cell (one of cancer main hallmarks). Such consideration could explain some of the functional changes that led to tumor development that were possibly not understood before.

The high enrichment of Cys in cancer-associated replacement (in the replacement residue) could also potentially be explained in terms of the role Cys oxidation plays in intracellular signalling and cell growth. Gaining Cys may provide an opportunity for Cys oxidation (potentially by the elevated reactive oxygen generation rate detected in almost all cancers (*15*)). Cys oxidation can affect proteins, altering their functions and in some instances enabling signal transmission to downstream targets (*16*) . An example of Cys oxidation involvement in cell growth control is its ability to inactivate certain tyrosine phosphatases; thus phosphorylated-tyrosine signal persists until the oxidized enzyme is degreaded (*17*). We also recommend studying the impact of this modification during cancer cells development.

CpG dinucleotides in DNA are known to mutate at high rates, and so the high mutability of Arg in deleterious mutations in general was explained by the high number of CpG sequences (the highest among all amino acids) presented in the codons for Arg (8). However, Arg is specifically highly enriched in cancer-associated mutations at the 'original residue'. Our investigation found that Arg is also found to be frequent in binding sites and plays a key role in the stability of the protein; replacement of Arg is likely to have a detrimental effect on the function and structure of a DNA binding complex (*18, 19*). Thus, cancer-associated genes affinity with binding activities can also explain the high mutability of Arg.

Our proposed explanations for the high enrichment of certain amino acid residues provide a missing link in numerous studies that reported on these findings. It connects these patterns to biological processes that were proven to be vital for carcinogenesis. We believe this approach could allow scientists researching these processes to highlight their impact further and encourage other possible biological processes to be put forward as explanations to the enrichment of Cys and aromatic amino acid group in cancer.

When the whole replacement is considered as the entity for analysis, we showed that 17 amino acid replacements are highly enriched in cancer-associated mutations (ratio > 2). This extended list of highly enriched replacements found in cancer changes the perceived view that only a handful of cancer driver mutations are frequently found in tumors (*8*) emphasising the complex nature of the disease. Our analysis showed that these enriched amino acid replacements in cancer exhibit on average an increase in hydrophobicity and decrease in polarity in comparison to less enriched replacements. This result underlines the different characteristics of these replacements compared to other non-cancer-associated reported replacements and that labelling cancer-associated mutations by their impact on the downstream proteins (by measuring the change in the property values between the original residue and the replacement residue) could yield important patterns and allow us to build a model that can score any mutation based on their likelihood of being cancer-associated. We recommend adapting this characteristic based categorisation of cancer-associated replacements when studying carcinogenesis. This approach could prove to be more insightful when assessing new targeted therapies than solely attributing them to the

organs or tissues in which the tumor sample was from. This characteristic based categorisation could reveal presence of same or similar effects certain biological processes presented by these replacements across different cancer types.

We used the differences in physico-chemical properties value between the 'original residue' and the 'replacement residue' and the position of the replacement on the protein sequence to train a machine-learning model predicting if a replacement is cancer-associated for the *PTEN* gene. This approach can be used to score any amino acid replacement and can be deployed to every oncoprotein with enough number of reported replacements. Although the model showed an adequate performance (F1 score of 0.76), it would be more powerful to extend the data to include other attributes such as genomic data and other protein structure data (when available). One recent study did demonstrate the efficacy of protein structure features in predicting cancer-associated mutation (*20*). The authors built a machine-learning model that achieved accuracy of over 89%, demonstrating protein structure and some microenvironment features could be excellent (when available) descriptors when predicting cancer-associated replacements. One noteworthy finding in our model was the impact of the position of the amino acid replacement on the prediction scores. There is a significant increase in likelihood of an amino acid replacement to be cancer-associated if it is positioned earlier in the sequence of the protein. This could be an artefact of the size of oncoproteins. However, There is evidence that diseases associated SNPs do occur in special locations (pockets and voids) on the protein structure (*21*) and these early segment of protein sequence could be responsible for forming these regions. It could also be that the early segment of the oncoprotein primarily influences key functions (e.g., binding) that are vital in tumor initiation and progression. We recommend this to be further validated.

It could beneficial in future research to prioritise replacements found when sequencing tumor samples that have higher scores of being cancer associated as per our model results. Our model can be utilised as-is without limitation or necessity for the 3-dimensional protein structure to be resolved. Using the model's prediction scores produced by our model, mutations found in tumors could be ranked by their likelihood of being cancer-associated

allowing for better understanding of mutations driving the initiation and progression of tumors.

Of particular interest are those replacements that were predicted to be cancer-associated but were not yet classed as such in the original training dataset for the gene *PTEN*. The dataset used to train the model was extracted from COSMIC Oct -2017. To confirm the model's ability to predict cancer-associated mutations, we extracted the somatic missense replacements implicated in cancer from COSMIC 2022 for the gene *PTEN* and compared the list of our false positives to the updated list of cancer-associated mutations. We found that 47% of the replacements in our false positives list are now included in COSMIC. If we only check against replacements that scored > 0.8 then the percentage increases to over 60%. This further confirms the model ability to predict novel candidate cancer mutations. We recommend considering the false positive replacements provided in the Appendix A - Table A.3 when researching novel mutations as these could be experimentally confirmed later.

We investigated aspects of cancer-associated genes to see if they show distinctive characteristics when compared to other human genes. One property of interest was the essentiality of the gene, where the essentiality score given to a gene indicates the effect of loss of function in this gene on the viability of the human. As mutated cancer-associated genes generally do not compromise viability in a direct manner, it could be expected that it is unlikely for these genes to score high on the essentiality spectrum. However, we demonstrated that on average there is positive correlation between gene essentiality scores and cancer associated genes. We applied a range of methods that score the degree of essentiality. In particular, we applied LofTool and Missense Z-score where the calculation of essentiality scores is based on intolerance to variants in human population sequenced data, and Blomen KBM7 and Wang K562 where cell viability data is used. We found that the cancer-associated genes exhibit a higher degree of essentiality compared to the scores calculated for all protein coding human genes. This finding was true for both types of measurement of essentiality (intolerance to variants and cell line viability). We also showed that this elevated essentiality score is also found in the case of Tumor Suppressor (TS) genes, as a distinct group of genes, as it is for all cancer-associated genes when compared to other human protein coding genes. Our results could be interpreted in the context of the

genes' involvement in particular biological activities. Genes classed as essential are often involved in cell, embryo, and organism growth. Similarly, proliferation is key for cancer cells. Therefore, the sets of genes that are essential and those that are involved in unregulated growth, as seen in cancer, tend to overlap. This finding provides further evidence of the importance of evolutionary aspects when studying cancer genes. Scientists might be able to further the understanding of cancer by incorporating properties linked to essentiality in their studies.

Several previous studies looked at the relationship between evolutionary conservation and the degree of essentiality in genes across species (*22, 23*). Essential genes have been shown to be more conserved and to evolve more slowly to nonessential genes in human (*22, 23*). We hypothesise those cancer-associated genes that are highly essential could be more suitable candidates for targeted therapies potentially providing less likelihood of developing drug resistance due their increased conserved status. It has been shown that cancer drugs cause a change in the selection pressure when administrated leading to increased drug resistance (*24*) so if the essential nature of a gene could slow its ability to evolve drug resistance, compared to less essential genes, then these genes should prioritised for drug discoveries when possible.

This result prompted us to develop a machine-learning model that could predict cancer-associated genes using essentiality related and general genomic properties; we extended the range of gene properties in our dataset to include, in addition to the essentiality scores, properties strongly linked to (although do not directly measure of) the gene's essentiality. Essential genes are likely to encode hub proteins in protein–protein interaction networks, have smaller-sized introns, are abundant and are ubiquitously expressed in cells and tissues (*25*). It was shown too that the more essential the gene is, the smaller the number of reported missense mutations for this gene (*26*). Therefore, in addition to general gene properties like gene % GC content and transcript count, we added protein–protein interaction network properties, such as degree indicating the number of interactions, closeness and betweenness. We also added various measures indicating selection pressure, such as dN/dS and measures of genetic variants, such as Tajima's D based on exons and regulatory sequences and Fay and Wu's H based on exons and regulatory sequences.

We tested different model configurations, selecting the model with the best performance. The resulting classifier displays excellent performance in predicting whether a human protein-coding gene is cancer-related; it achieved 89% for the accuracy and the area under curve (AUC) was > 0.85. Our machine-learning model prediction scores provide a good base to prioritise the likelihood of a human protein coding genes to be a cancer gene. Of key importance in our results are those predictions that are false positives, i.e., those genes with high scores that have no published cancer association. Two possible explanations exist: either they represent a failure of the model to correctly classify the data or, alternatively, these gene are in fact cancer related but have not yet been characterised as such. These genes are therefore likely to encode future cancer targets.

Our machine-learning model identified the most important properties for the classification, ranking the properties by their impact on the prediction and revealing their influences on the genes found to be cancer associated. Protein-protein interactions properties such as degree and closeness are confirmed to be very influential when assessing the likelihood of a gene to be cancer-associated. This reflects that cancer-associated genes often code for protein found in 'hubs' within the protein-protein interaction networks. The ranking also showed essential score Phi and Tajima's D Regulatory to be among top impactful features (albeit to lesser extent). This confirms that on average these genes are more essential than other non-cancer genes and shows evidence for positive selection on these genes. These findings may offer targets for further research. Furthermore, our model is easy to implement by scientists investigating potential cancer-related genes using the open-source code provided in the Appendix C. Importantly, additional properties may be easily incorporated into our model revealing their influence on the prediction.

According to a recent comprehensive review of cancer driver genes prediction models, currently the best performing machine learning model is driverMaps with AUC= 0.94 followed by HotNet2 with AUC=0.81 (*27*). When comparing our model performance using AUC to the other 12 reviewed cancer driver genes prediction models, our model would come second with AUC= 0.86. Our predictive model achieved a better AUC measured performance when compared to the top-performing model using a similar network based

approach (HotNet2 with AUC=0.81) and better than the best function-based prediction model (MutPanning with AUC=0.62). This strong performance indicates the importance of combining different and distinctive gene properties when building prediction models. To train our model, we used a combination of protein-protein interaction network-based properties, essentiality scores and evolutionary based properties in addition to general genomic properties. We recommend combining properties from different approaches and using the most recent list of genes currently implicated in cancer to further enhance the performance of cancer-associated genes prediction models. We also recommend considering evolutionary related properties of studied genes when predicting the association with cancer. As demonstrated in our results in chapter 4, could provide a significant contribution to our ability to identify cancer-associated genes. Our model and the other top performing cancer-associated genes prediction models provide a good basis for scientists to start considering candidate genes predicted to be cancer-associated by these methods in their research. Cancer genes databases such as COSMIC could also incorporate these candidate cancer-associated genes (possibly in a separate tier) and make them available for researchers.

In this thesis, we worked on a key challenge in cancer research: the identification of cancer-related oncogenes and cancer-associated point mutations and their distinctive characteristics through the utilisation of the machine learning techniques. Accurate identification of genes and mutations potentially related to cancer would provide an opportunity to advance both personalised treatment of cancer and aid drug discovery by providing new targets. Identifying the distinctive characteristics of these genes and mutations furthers our understanding of this disease and using them to train and build computational models is showing great potential. This approach would be practical and provide results in timely fashion accelerating multiple aspects in cancer research.

5.1 References

1. Z. Sondka *et al.*, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
2. W. Lee *et al.*, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
3. N. Tokuriki, D. S. Tawfik, Stability effects of mutations and protein evolvability. *Carbohydradtes and glycoconjugates / Biophysical methods* **19**, 596-604 (2009).
4. P. Yue, Z. Li, J. Moult, Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. *Journal of Molecular Biology* **353**, 459-473 (2005).
5. E. Callaway, 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203+ (2020).
6. S. R. Eddy, Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **22**, 1035-1036 (2004).
7. P. Anoosha, R. Sakthivel, M. Michael Gromiha, Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *ScienceDirect* **1862**, 155-165 (2016).
8. Z. A. Szpiech *et al.*, Prominent features of the amino acid mutation landscape in cancer. *PLOS ONE* **12**, e0183273 (2017).
9. T. A. P. de Beer *et al.*, Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLOS Computational Biology* **9**, e1003382 (2013).
10. H. Nishi *et al.*, Cancer Missense Mutations Alter Binding Properties of Proteins and Their Interaction Networks. *PLOS ONE* **8**, e66273 (2013).
11. M. E. Ortiz-Soto, S. Reising, A. Schlosser, J. Seibel, Structural and functional role of disulphide bonds and substrate binding residues of the human beta-galactoside alpha-2,3-sialyltransferase 1 (hST3Gal1). *Scientific Reports* **9**, 17993 (2019).
12. A. L. Harris, S. Nicholson, in *Epidermal growth factor receptors in human breast cancer*. (Springer US, Boston, MA, 1988), pp. 93-118.
13. R. Mor-Cohen *et al.*, Unique Disulfide Bonds in Epidermal Growth Factor (EGF) Domains of β3 Affect Structure and Function of αIIbβ3 and αvβ3 Integrins in Different Manner. *Journal of Biological Chemistry* **287**, 8879-8891 (2012).
14. M. Popielarski, H. Ponamarczuk, M. Stasiak, C. Watała, M. Świątkowska, Modifications of disulfide bonds in breast cancer cell migration and invasiveness. *American journal of cancer research* **9**, 1554-1582 (2019).
15. G.-Y. Liou, P. Storz, Reactive oxygen species in cancer. *Free radical research* **44**, 479-496 (2010).
16. H. Miki, Y. Funato, Regulation of intracellular signalling through cysteine oxidation by reactive oxygen species. *Journal of biochemistry* **151**, 255-261 (2012).
17. E. K. Krasnowska *et al.*, N-acetyl-l-cysteine fosters inactivation and transfer to endolysosomes of c-Src. *Free Radical Biology and Medicine* **45**, 1566-1572 (2008).
18. R. Fan *et al.*, The effects of L-arginine on protein stability and DNA binding ability of SaeR, a transcription factor in Staphylococcus aureus. *ScienceDirect* **177**, 105765 (2021).

19. K. Yao *et al.*, The Arginine/Lysine-Rich Element within the DNA-Binding Domain Is Essential for Nuclear Localization and Function of the Intracellular Pathogen Resistance 1. *PloS one* **11**, e0162832-e0162832 (2016).
20. J.-J. Liu *et al.*, The structure-based cancer-related single amino acid variation prediction. *Scientific Reports* **11**, 13599 (2021).
21. N. O. Stitziel *et al.*, Structural Location of Disease-associated Single-nucleotide Polymorphisms. *Journal of Molecular Biology* **327**, 1021-1030 (2003).
22. I. K. Jordan, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12**, 962-968 (2002).
23. B. Georgi, B. F. Voight, M. Bućan, From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet* **9**, e1003484 (2013).
24. D. Sun, S. Dalin, M. T. Hemann, D. A. Lauffenburger, B. Zhao, Differential selective pressure alters rate of drug resistance acquisition in heterogeneous tumor populations. *Scientific Reports* **6**, 36198 (2016).
25. D. Park, J. Park, S. G. Park, T. Park, S. S. Choi, Analysis of human disease genes in the context of gene essentiality. *Genomics* **92**, 414-418 (2008).
26. I. Bartha, J. di Iulio, J. C. Venter, A. Telenti, Human gene essentiality. *Nat Rev Genet* **19**, 51-62 (2018).
27. X. Shi *et al.*, Comprehensive evaluation of computational methods for predicting cancer driver genes. *Briefings in Bioinformatics* **23**, bbab548 (2022).

Appendix A

Cancer-associated replacements prediction models:

Model Type	Model Performance
Gradient Boosted Trees Classifier	AUC: [0.70631, 0.670514], Area Under PR Curve: [0.77224, 0.743286], FVE Binomial: [0.09021, 0.065328], Gini Norm: [0.41262, 0.341028], Kolmogorov-Smirnov: [0.32233, 0.27244199999999996], LogLoss: [0.6123, 0.628756], Max MCC: [0.32928, 0.277004], RMSE: [0.46053, 0.468478], Rate@Top10%: [0.83019, 0.8275840000000001], Rate@Top5%: [0.88889, 0.88148], Rate@TopTenth%: [1.0, 0.8], labels: ['(0,-1)', '(.,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%', 'Rate@Top5%', 'Rate@TopTenth%']
RandomForest Classifier (Entropy)	AUC: [0.69823, 0.667972], Area Under PR Curve: [0.75349, 0.7327980000000001], FVE Binomial: [0.08541, 0.059401999999999996], Gini Norm: [0.39646, 0.33594399999999996], Kolmogorov-Smirnov: [0.30975, 0.273954], LogLoss: [0.61553, 0.63274], Max MCC: [0.32083, 0.28122199999999997], RMSE: [0.46029, 0.46963599999999994], Rate@Top10%: [0.84906, 0.7936420000000001], Rate@Top5%: [0.85185, 0.8152379999999999], Rate@TopTenth%: [1.0, 0.8], labels: ['(0,-1)', '(.,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%',
eXtreme Gradient Boosted Trees Classifier	AUC: [0.72917, 0.697958], Area Under PR Curve: [0.78893, 0.7693519999999999], FVE Binomial: [0.11706, 0.090182], Gini Norm: [0.45834, 0.395916], Kolmogorov-Smirnov: [0.3695, 0.319348], LogLoss: [0.59423, 0.612034], Max MCC: [0.37168, 0.322612], RMSE: [0.45165, 0.460532], Rate@Top10%: [0.88679, 0.883508], Rate@Top5%: [0.88889, 0.903704], Rate@TopTenth%: [1.0, 1.0], labels: ['(0,-1)', '(.,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-
Elastic-Net Classifier (L2 / Binomial Deviance)	AUC: [0.68496, 0.6466879999999999], Area Under PR Curve: [0.7355, 0.711044], FVE Binomial: [0.06836, 0.043074], Gini Norm: [0.36992, 0.293376], Kolmogorov-Smirnov: [0.33176, 0.2622], LogLoss: [0.627, 0.64372], Max MCC: [0.32845, 0.264946], RMSE: [0.46703, 0.475044], Rate@Top10%: [0.79245, 0.762262], Rate@Top5%: [0.77778, 0.76296], Rate@TopTenth%: [1.0, 0.8], labels: ['(0,-1)', '(.,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss',
Light Gradient Boosting on ElasticNet Predictions	AUC: [0.65781], Area Under PR Curve: [0.72615], FVE Binomial: [0.05246], Gini Norm: [0.31562], Kolmogorov-Smirnov: [0.2673], LogLoss: [0.63771], Max MCC: [0.2619], RMSE: [0.47265], Rate@Top10%: [0.83019], Rate@Top5%: [0.77778], Rate@TopTenth%: [1.0], labels: ['(0,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%', 'Rate@Top5%', 'Rate@TopTenth%']
Advanced AVG Blender	AUC: [0.72099, 0.6774739999999999], Area Under PR Curve: [0.77786, 0.7470320000000001], FVE Binomial: [0.09941, 0.069484], Gini Norm: [0.44198, 0.354948], Kolmogorov-Smirnov: [0.33648, 0.28780799999999995], LogLoss: [0.60611, 0.625958], Max MCC: [0.34583, 0.292926], RMSE: [0.45709, 0.46691399999999994], Rate@Top10%: [0.83019, 0.83019], Rate@Top5%: [0.88889, 0.88148], Rate@TopTenth%: [1.0, 1.0], labels: ['(0,-1)', '(.,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%',
Naive Bayes combiner classifier	AUC: [0.61626], Area Under PR Curve: [0.68894], FVE Binomial: [0.02706], Gini Norm: [0.23252], Kolmogorov-Smirnov: [0.22956], LogLoss: [0.6548], Max MCC: [0.2402], RMSE: [0.48055], Rate@Top10%: [0.69811], Rate@Top5%: [0.74074], Rate@TopTenth%: [1.0], labels: ['(0,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%', 'Rate@Top5%', 'Rate@TopTenth%']
Generalized Additive2 Model	AUC: [0.6603], Area Under PR Curve: [0.71606], FVE Binomial: [0.0485], Gini Norm: [0.3206], Kolmogorov-Smirnov: [0.29403], LogLoss: [0.64037], Max MCC: [0.29253], RMSE: [0.47365], Rate@Top10%: [0.7037], Rate@Top5%: [0.88889], Rate@TopTenth%: [1.0], labels: ['(0,-1)'], metrics: ['AUC', 'Area Under PR Curve', 'FVE Binomial', 'Gini Norm', 'Kolmogorov-Smirnov', 'LogLoss', 'Max MCC', 'RMSE', 'Rate@Top10%', 'Rate@Top5%', 'Rate@TopTenth%']

Table A.1 List of all different machine learning models built to predict the liklihood of a replacement to be cancer associated for the PTEN protein with their performance results (extreme Gradient Boosted Trees Classifier was selected as the preferred method)

Cancer-associated genes prediction models:

Model Name	Modeling Performance
RuleFit Classifier	AUC: [0.82562], FVE Binomial: [0.1289], Gini Norm: [0.65124], Kolmogorov-Smirnov: [0.53206], LogLoss: [0.11173], Max MCC: [0.23719], RMSE: [0.16585], Rate@Top10%: [0.13542], Rate@Top5%: [0.17361], Rate@TopTenth%: [0.0], labels: [u'(0,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
RandomForest Classifier (Gini)	AUC: [0.86065, 0.847272], FVE Binomial: [0.23378, 0.18973], Gini Norm: [0.7213, 0.6945439999999999], Kolmogorov-Smirnov: [0.59054, 0.5542], LogLoss: [0.09828, 0.103328], Max MCC: [0.33818, 0.29684999999999995], RMSE: [0.15441, 0.15600600000000003], Rate@Top10%: [0.17014, 0.150696], Rate@Top5%: [0.27083, 0.220832], Rate@TopTenth%: [0.66667, 0.733336], labels: [u'(0,-1)', u'(.,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
Keras Slim Residual Neural Network Classifier using Training Schedule (1 Layer: 64 Units)	AUC: [0.8524], FVE Binomial: [0.05679], Gini Norm: [0.7048], Kolmogorov-Smirnov: [0.58594], LogLoss: [0.12098], Max MCC: [0.29886], RMSE: [0.16094], Rate@Top10%: [0.16667], Rate@Top5%: [0.22222], Rate@TopTenth%: [0.33333], labels: [u'(0,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
eXtreme Gradient Boosted Trees Classifier with Early Stopping	AUC: [0.8642, 0.862538], FVE Binomial: [0.25164, 0.231666], Gini Norm: [0.7284, 0.7250759999999999], Kolmogorov-Smirnov: [0.60732, 0.5877100000000001], LogLoss: [0.09599, 0.097986], Max MCC: [0.36036, 0.30265200000000003], RMSE: [0.15374, 0.154798], Rate@Top10%: [0.18056, 0.160418], Rate@Top5%: [0.25694, 0.22221999999999995], Rate@TopTenth%: [0.33333, 0.6], labels: [u'(0,-1)', u'(.,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
AVG Blender	AUC: [0.85787, 0.8576460000000001], FVE Binomial: [0.24329, 0.22457000000000002], Gini Norm: [0.71574, 0.715292], Kolmogorov-Smirnov: [0.58074, 0.577866], LogLoss: [0.09706, 0.098894], Max MCC: [0.35001, 0.30664199999999997], RMSE: [0.15372, 0.154976], Rate@Top10%: [0.17014, 0.15486], Rate@Top5%: [0.27083, 0.23055600000000004], Rate@TopTenth%: [0.66667, 0.6000019999999999], labels: [u'(0,-1)', u'(.,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
Light Gradient Boosting on ElasticNet Predictions	AUC: [0.82197], FVE Binomial: [0.11823], Gini Norm: [0.64394], Kolmogorov-Smirnov: [0.54529], LogLoss: [0.1131], Max MCC: [0.28227], RMSE: [0.15852], Rate@Top10%: [0.15972], Rate@Top5%: [0.21528], Rate@TopTenth%: [0.66667], labels: [u'(0,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
Elastic-Net Classifier (mixing alpha=0.5 / Binomial Deviance)	AUC: [0.81906], FVE Binomial: [0.10206], Gini Norm: [0.63812], Kolmogorov-Smirnov: [0.53509], LogLoss: [0.11518], Max MCC: [0.28221], RMSE: [0.15957], Rate@Top10%: [0.15625], Rate@Top5%: [0.19444], Rate@TopTenth%: [0.0], labels: [u'(0,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
Gradient Boosted Trees Classifier	AUC: [0.85472], FVE Binomial: [0.19056], Gini Norm: [0.70944], Kolmogorov-Smirnov: [0.60883], LogLoss: [0.10382], Max MCC: [0.28079], RMSE: [0.15898], Rate@Top10%: [0.14931], Rate@Top5%: [0.22222], Rate@TopTenth%: [0.0], labels: [u'(0,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']
Generalized Additive2 Model	AUC: [0.86928, 0.85505], FVE Binomial: [0.25555, 0.21647], Gini Norm: [0.73856, 0.7101], Kolmogorov-Smirnov: [0.60161, 0.56743], LogLoss: [0.09549, 0.099926], Max MCC: [0.3905, 0.29954], RMSE: [0.15352, 0.15575799999999998], Rate@Top10%: [0.18056, 0.156944], Rate@Top5%: [0.28472, 0.223612], Rate@TopTenth%: [0.33333, 0.4], labels: [u'(0,-1)', u'(.,-1)'], metrics: [u'AUC', u'FVE Binomial', u'Gini Norm', u'Kolmogorov-Smirnov', u'LogLoss', u'Max MCC', u'RMSE', u'Rate@Top10%', u'Rate@Top5%', u'Rate@TopTenth%']

Table A.2 List of all different machine learning models built to predict the likelihood of a gene to be cancer associated with their performance results (extreme Gradient Boosted Trees Classifier was selected as the preferred method)

Potential cancer-associated replacements identified by our model for *PTEN* gene (PTEN protein):

Replacement	Likelihood to be cancer associated score
Y2199C	97.90%
P167S	91.82%
E18K	91.11%
G125E	90.50%

P78S	90.16%
F33S	89.92%
N421I	89.87%
E105K	89.69%
S65F	89.14%
G103E	88.82%
M131K	88.66%
P168L	88.43%
D104N	87.51%
D113N	87.51%
Y102C	87.51%
P117R	87.01%
L175P	87.01%
L180P	86.75%
L220P	86.75%
D57N	86.54%
P269S	85.82%
L90F	85.74%
F183L	85.67%
K38T	85.33%
L60F	85.25%
E206K	85.01%
E251K	85.01%
I14N	84.42%
F21Y	84.26%
E204K	84.05%
H4Y	83.54%
T94N	83.31%
H89Y	83.16%
L13F	83.06%
L84S	82.64%
G267D	82.54%
S188F	82.50%
P886H	82.26%
W11R	81.93%
Q169H	81.88%
E236K	81.77%
Y176C	81.74%
V116G	81.27%
K564I	81.22%
V139G	81.02%
S72C	81.01%
H156L	80.96%
L124F	80.94%
L111F	80.94%

D248N	80.93%
V184D	80.89%
R80C	80.62%
H62N	80.52%
P214L	80.44%
E357K	80.35%
K59N	80.25%
F311L	80.12%
G180D	80.09%
T91S	79.95%
T55N	79.95%
T46N	79.95%
P130A	79.80%
P286L	79.74%
A163V	79.62%
R80L	79.60%
C12S	79.59%
S191Y	79.54%
W222R	79.49%
R86C	79.40%
H127N	79.38%
A9S	79.26%
Y557F	79.25%
T46A	79.13%
G140R	79.08%
S79C	78.91%
C162Y	78.90%
G14R	78.85%
W1505C	78.68%
L32F	78.58%
P1635S	78.55%
P1826S	78.55%
P1020S	78.55%
P1452S	78.55%
C307F	78.51%
E438K	78.46%
E436K	78.46%
K140R	78.44%
T165M	78.34%
E1294K	78.19%
E1432K	78.19%
E981K	78.19%
I21M	78.17%
Y1367D	78.10%
G125R	77.97%

P354Q	77.95%
F1047L	77.94%
F1407L	77.94%
A81V	77.88%
C31R	77.88%
F538I	77.54%
K164N	77.49%
G145S	77.31%
S160R	77.30%
A29P	77.14%
P1191S	77.13%
P867S	77.13%
G886C	77.12%
A88T	76.86%
G77S	76.82%
A68T	76.75%
A119T	76.75%
G152S	76.72%
P330S	76.65%
G282E	76.62%
G374E	76.62%
A155T	76.54%
P618S	76.51%
T78N	76.46%
L1578P	76.44%
Q98E	76.39%
L8V	76.32%
P91T	76.30%
I164M	76.19%
D341Y	76.17%
L237R	76.15%
P938L	76.12%
P1262L	76.12%
P1348L	76.12%
L109V	76.08%
P1816S	76.04%
Y779D	76.02%
G1108E	76.00%
V63I	75.82%
G1187W	75.77%
S145R	75.71%
S190N	75.70%
S1200F	75.69%
S881F	75.69%
G1137E	75.30%

F1177S	75.17%
P44T	75.14%
E191V	75.05%
D360Y	75.04%
L197R	74.98%
E296Q	74.90%
P508S	74.88%
R86S	74.87%
Y860N	74.87%
S1554I	74.83%
R92Q	74.65%
P1081R	74.52%
S225L	74.49%
A68E	74.48%
H245Y	74.41%
E1105K	74.21%
E1242K	74.21%
E1002K	74.21%
Q74E	74.14%
E746K	73.96%
E1753K	73.96%
E1090K	73.96%
E1192K	73.96%
P520L	73.92%
P735L	73.92%
F313V	73.87%
A6V	73.85%
C645F	73.83%
F724L	73.67%
S824F	73.47%
S803F	73.47%
S839F	73.47%
S821F	73.47%
F631L	73.43%
G1062E	73.39%
G1775E	73.28%
F1265L	73.21%
S1320F	73.18%
S969F	73.18%
T1005K	73.14%
E383K	73.10%
I303S	73.10%
P2197L	72.97%
D261A	72.96%
C12Y	72.95%

P282L	72.85%
D863H	72.75%
G1465E	72.64%
S834I	72.55%
P1075L	72.51%
E894K	72.31%
S1418F	72.26%
A6T	72.22%
M858K	72.20%
D301N	72.19%
D826Y	72.15%
D599Y	72.15%
D739Y	72.15%
F430S	72.08%
L883R	71.98%
L1484R	71.98%
I103L	71.96%
E283G	71.83%
P772S	71.72%
P963L	71.64%
E644K	71.63%
Q48R	71.54%
V38M	71.33%
S1685F	71.17%
D695Y	71.17%
G7V	71.12%
L1218P	71.10%
S700I	71.03%
L1513Q	70.91%
A240E	70.90%
V176L	70.89%
S42P	70.85%
Y404C	70.85%
Y180H	70.83%
N87D	70.76%
D805Y	70.73%
D910N	70.69%
D1309N	70.69%
D1138N	70.69%
C1295S	70.56%
F1122L	70.40%
W334G	70.33%
E1019G	70.31%
A250V	70.30%
R147Q	70.28%

D1222H	70.21%
V336D	70.06%
S209Y	70.05%
P592S	70.03%
V173I	70.01%
G100A	70.00%
A1415D	69.90%
E251D	69.86%
S145A	69.86%
D360N	69.82%
R76H	69.82%
E520K	69.79%
E441K	69.79%
R143Q	69.74%
P1943L	69.72%
P1080L	69.65%
F1079V	69.64%
S778F	69.63%
S718F	69.63%
D113G	69.63%
D1698N	69.56%
L638R	69.55%
T173I	69.52%
V176I	69.36%
Q30R	69.33%
L2V	69.23%
D77G	69.18%
I135T	69.14%
E646K	69.09%
R80H	68.86%
L368F	68.65%
R86H	68.63%
H173R	68.48%
N50S	68.38%
G639E	68.18%
R312K	68.17%
E1656V	68.11%
Q207H	67.87%
Y1161N	67.85%
S359L	67.80%
D1065N	67.78%
D1308N	67.78%
G1147D	67.72%
V104M	67.65%
D589H	67.62%

M76V	67.37%
N1125Y	67.29%
P461L	67.21%
P423L	67.21%
I1403N	67.21%
G1150W	67.12%
P748L	67.07%
S1568Y	66.90%
G581D	66.82%
S158P	66.77%
S302N	66.66%
G180S	66.63%
S372L	66.62%
D599H	66.57%
S1647L	66.31%
R45T	66.31%
E786G	66.30%
E690G	66.30%
E542G	66.30%
G195R	66.26%
G206R	66.26%
P1882S	66.23%
V39F	66.20%
L473F	66.12%
D450V	66.03%
R913K	66.02%
Y1730C	66.00%
E1839Q	65.95%
E1926Q	65.86%
L1337F	65.76%
L1053F	65.76%
H213P	65.71%
D1042V	65.68%
D1168V	65.68%
Q1724K	65.65%
Y980C	65.62%
E736V	65.50%
R392K	65.35%
V116M	65.34%
A191V	65.25%
D450N	65.20%
D805N	65.20%
N344K	65.18%
H442Y	65.13%
S1872F	65.09%

T218A	65.08%
Q961P	65.06%
V576D	65.04%
Y714D	64.95%
T1312K	64.89%
R265L	64.79%
K256R	64.73%
A1027S	64.61%
V69A	64.59%
G236R	64.57%
G209R	64.57%
N73S	64.50%
Y428H	64.29%
H388Y	64.24%
M303I	64.03%
E1002Q	64.02%
D1556V	63.90%
S221T	63.89%
D1778N	63.87%
Y742C	63.69%
Y777S	63.67%
F1118Y	63.64%
T271R	63.46%
C2067G	63.46%
T220A	63.45%
R1370K	63.43%
G1932D	63.42%
A196T	63.28%
V159M	63.03%
H4R	63.02%
I23L	62.98%
E1927G	62.90%
N1136K	62.90%
R270L	62.89%
G451D	62.85%
S263R	62.83%
A1576S	62.81%
G394V	62.78%
E1722Q	62.66%
A335S	62.60%
N161S	62.57%
L1817F	62.50%
L1727F	62.50%
D1916H	62.48%
F1835Y	62.31%

M1610I	62.22%
Q771P	62.21%
P1348R	62.17%
N299K	61.94%
T216I	61.81%
T280I	61.81%
G919V	61.77%
S228C	61.72%
R332W	61.66%
E1017G	61.65%
L463S	61.65%
D1657N	61.63%
D1266N	61.63%
S1326L	61.57%
S903L	61.57%
N1148K	61.53%
K386N	61.52%
W882G	61.39%
R536M	61.30%
V179L	61.26%
R12Q	61.20%
S956N	61.14%
S1475N	61.14%
D248E	61.03%
R186S	61.02%
N392K	61.01%
H611Y	60.97%
M750I	60.87%
L1141R	60.84%
M241L	60.78%
R772K	60.67%
R608K	60.61%
R539K	60.61%
Y498C	60.39%
L814F	60.33%
L612F	60.33%
T1089S	60.25%
T1374S	60.25%
D759N	60.21%
D659N	60.21%
L843P	60.12%
R1402W	60.10%
R1210W	60.10%
E761Q	60.07%
R203G	60.06%

R205G	60.06%
E1346D	60.05%
E898D	60.05%
H442Q	60.04%
F354Y	59.97%
S256R	59.96%
A1305S	59.91%
S1694N	59.85%
M489I	59.72%
G496V	59.66%
G428V	59.66%
S1159L	59.62%
S1058L	59.62%
H222R	59.58%
H1318D	59.56%
L708I	59.54%
Y1058H	59.42%
N1904Y	59.41%
N1952Y	59.41%
E1753G	59.40%
R1012W	59.39%
R1608W	59.39%
V24A	59.38%
A1588S	59.31%
H550Y	59.28%
R203Q	59.22%
L84W	59.19%
S444N	59.18%
V1134E	59.16%
V1153G	59.04%
S422C	59.02%
T1659S	58.96%
T387N	58.94%
R821K	58.94%
A774D	58.91%
H1318Y	58.90%
D595N	58.81%
S261R	58.81%
S1164N	58.80%
S1195N	58.80%
W230L	58.76%
R409C	58.74%
E385D	58.67%
R1190W	58.64%
K1684M	58.58%

R1344W	58.51%
R323L	58.43%
H418Y	58.40%
Y701H	58.39%
I103V	58.37%
M7L	58.33%
P337T	58.10%
L351W	57.84%
C1146Y	57.79%
W1558R	57.67%
E367D	57.52%
I50V	57.48%
I21V	57.48%
R260Q	57.42%
W1363S	57.34%
H388N	57.30%
I362M	57.27%
D1556A	57.22%
D1065A	57.22%
V169M	57.20%
C324R	57.15%
P1074T	57.05%
T1050S	56.99%
R1188C	56.96%
R1318C	56.96%
R1535C	56.96%
Q427H	56.89%
W383R	56.89%
T737P	56.85%
N816K	56.84%
T307S	56.84%
S536L	56.77%
T547K	56.74%
Q489K	56.62%
R1582W	56.61%
K541T	56.58%
G458V	56.58%
R500W	56.53%
R617W	56.53%
K1269T	56.50%
K1077N	56.42%
P644R	56.41%
A640S	56.31%
T226I	56.26%
N2193K	56.21%

D534N	56.19%
Q1296H	56.11%
Y557H	56.00%
Y504H	56.00%
S1141N	55.96%
N211H	55.87%
R1446W	55.87%
K1236N	55.84%
I71V	55.80%
G1033R	55.75%
G1009R	55.75%
G1023R	55.75%
R241H	55.73%
G857V	55.69%
R769K	55.65%
R770K	55.65%
R1946K	55.62%
L554S	55.61%
A309T	55.57%
I83V	55.40%
G267S	55.39%
R323C	55.32%
R322C	55.32%
D218E	55.18%
N228S	55.18%
L1214V	55.16%
S410C	55.14%
T415S	55.09%
T786N	55.04%
T598N	55.04%
Q795L	55.04%
T565N	55.03%
T547N	55.03%
T522N	55.03%
R247H	55.00%
V3M	55.00%
S231A	54.97%
A283P	54.94%
R265H	54.93%
S268P	54.90%
R205Q	54.84%
R183Q	54.84%
N1934K	54.82%
A284T	54.72%
A1473E	54.70%

L329S	54.67%
P1037A	54.58%
D467V	54.56%
G355R	54.55%
P2035R	54.54%
T341M	54.43%
I441M	54.43%
L1825S	54.40%
R201Q	54.39%
D327V	54.39%
D1112V	54.39%
R1740W	54.37%
Q202R	54.36%
T248I	54.35%
G1893V	54.34%
E1382D	54.29%
Q406L	54.16%
E1051D	54.09%
D202G	54.07%
I164V	54.07%
G477V	54.04%
A1263V	53.98%
A1245V	53.98%
A1117V	53.98%
H809D	53.95%
G451R	53.93%
R499W	53.92%
C1003R	53.91%
G1983V	53.90%
G282R	53.90%
G289S	53.83%
R1067C	53.55%
R1541C	53.55%
R1665C	53.55%
R1351C	53.55%
R1027C	53.55%
G888R	53.53%
G1294R	53.53%
K541N	53.52%
G1226R	53.48%
G1328R	53.48%
G1011R	53.48%
G1150R	53.48%
T333M	53.47%
N1402Y	53.45%

D268E	53.43%
R558W	53.34%
Q489P	53.31%
H1580N	53.26%
S1270T	53.16%
S1095C	53.08%
R851L	53.05%
L754F	52.99%
L638F	52.99%
S849Y	52.99%
A284V	52.95%
R1123S	52.93%
S294R	52.93%
T931A	52.92%
T1249A	52.92%
R384G	52.91%
A868T	52.88%
A1235T	52.88%
L566S	52.85%
L705S	52.85%
G1239S	52.82%
G1299S	52.82%
R653W	52.82%
S1483C	52.77%
E319A	52.76%
R608W	52.76%
T1108S	52.72%
K882R	52.63%
R1117C	52.60%
R1157C	52.60%
R1384C	52.60%
R1322C	52.60%
L1731V	52.53%
R1046C	52.46%
R998C	52.46%
R1414C	52.46%
R1257C	52.46%
A987V	52.33%
A1342V	52.33%
R870L	52.33%
A1186T	52.30%
E2070Q	52.29%
Q730H	52.19%
L306V	52.12%
L724H	52.11%

P1348T	52.10%
I1605M	52.09%
H1886L	52.08%
C1146R	52.08%
T1443M	52.06%
M241V	52.04%
S577N	52.03%
P1671A	51.94%
R1895W	51.88%
A909V	51.87%
S342P	51.87%
G1187R	51.86%
P576T	51.72%
P550T	51.72%
W1409L	51.70%
Y663H	51.66%
G1294V	51.63%
R812C	51.62%
K661N	51.61%
G1199R	51.52%
G1369R	51.52%
Q538L	51.45%
T351A	51.44%
I1404M	51.42%
E2190K	51.35%
P1039T	51.25%
P1755T	51.25%
T993A	51.24%
T1170A	51.24%
S345T	51.20%
K1202N	51.15%
T1690N	51.03%
V201M	50.94%
A350G	50.92%
K1277Q	50.89%
P1135A	50.75%
C1144Y	50.73%
Q373E	50.68%
R815L	50.65%
A999V	50.65%
A1473V	50.65%
A1520V	50.65%
A1088V	50.65%
R817C	50.62%
R692C	50.62%

K468N	50.57%
K644N	50.57%
E746D	50.56%
S1227R	50.55%
D1657A	50.50%
D1664A	50.50%
S1141R	50.48%
S1198R	50.48%
S334T	50.31%
S348T	50.31%
S1380P	50.26%
R332Q	50.20%
S1049T	50.12%
P1191T	50.11%
K493N	50.04%
S1440P	50.02%
S1134P	50.02%
R366G	50.00%
R311G	50.00%
R328G	50.00%
R338G	50.00%

Table A.3: The likelihood score of each amino acid replacement for PTEN protein to be cancer–associated as predicted by our model

Potential cancer-associated genes candidates as identified by our model:

Gene Name	**Gene ID**
APC	ENSG00000134982
APP	ENSG00000142192
PSEN1	ENSG00000080815
ACTB	ENSG00000075624
INTS6	ENSG00000102786
LRP2	ENSG00000081479
LRP1	ENSG00000123384
CUL4A	ENSG00000139842
SRC	ENSG00000197122
SIN3A	ENSG00000169375
SMAD1	ENSG00000170365
IRS1	ENSG00000169047
HDAC4	ENSG00000068024
GSK3B	ENSG00000082701
CSNK2A1	ENSG00000101266
DAG1	ENSG00000173402

SPTBN1	ENSG00000115306
FYN	ENSG00000010810
LRPPRC	ENSG00000138095
SOS1	ENSG00000115904
SMURF2	ENSG00000108854
RANBP9	ENSG00000010017
KRT18	ENSG00000111057
SMARCA2	ENSG00000080503
USP7	ENSG00000187555
HDAC2	ENSG00000196591
DDB1	ENSG00000167986
PPARGC1A	ENSG00000109819
EP400	ENSG00000183495
TRAF3	ENSG00000131323
RXRA	ENSG00000186350
ARNTL	ENSG00000133794
TP73	ENSG00000078900
HTT	ENSG00000197386
BRE	ENSG00000158019
ITGB1	ENSG00000150093
PIAS1	ENSG00000033800
KMT2A	ENSG00000118058
PRKCB	ENSG00000166501
VAV1	ENSG00000141968
PAK2	ENSG00000180370
NCOA3	ENSG00000124151
TGFBR1	ENSG00000106799
MAPK8IP3	ENSG00000138834
RELA	ENSG00000173039
TLE1	ENSG00000196781
DNMT1	ENSG00000130816
UBE2D1	ENSG00000072401
POU2F1	ENSG00000143190
CUL1	ENSG00000055130
DYNC1H1	ENSG00000197102
ACTN1	ENSG00000072110
JARID1A,KDM5A	ENSG00000073614
RPS6KB1	ENSG00000108443
UBE3A	ENSG00000114062
KAT2B	ENSG00000114166
CHD8	ENSG00000100888
TRIM33	ENSG00000197323
SMURF1	ENSG00000198742
MAP3K7	ENSG00000135341
LMNB1	ENSG00000113368

DHX9	ENSG00000135829
NLK	ENSG00000087095
NFKB1	ENSG00000109320
RASA1	ENSG00000145715
PTK2	ENSG00000169398
BAZ1B	ENSG00000009954
SETDB1	ENSG00000143379
BMPR1A	ENSG00000107779
TERF2	ENSG00000132604
HDAC1	ENSG00000116478
NTRK2	ENSG00000148053
MAFG	ENSG00000197063
IKZF3	ENSG00000161405
MAP3K1	ENSG00000095015
RBPJ	ENSG00000168214
ILF3	ENSG00000129351
TRAF6	ENSG00000175104
NEDD4	ENSG00000069869
CARM1	ENSG00000142453
HDAC9	ENSG00000048052
RELB	ENSG00000104856
MEF2C	ENSG00000081189
PRKCI	ENSG00000163558
RNF2	ENSG00000121481
ITSN1	ENSG00000205726
PTPN6	ENSG00000111679
RBL2	ENSG00000103479
DLG1	ENSG00000075711
NFKBIA	ENSG00000100906
ETS1	ENSG00000134954
ZEB2	ENSG00000169554
MED1	ENSG00000125686
GSN	ENSG00000148180
KLF5	ENSG00000102554
ATR	ENSG00000175054
RNF4	ENSG00000063978
CDK8	ENSG00000132964
HGS	ENSG00000185359
EIF3H	ENSG00000147677
YWHAQ	ENSG00000134308
TSG101	ENSG00000074319
DOT1L	ENSG00000104885
CEBPB	ENSG00000172216
MAP1B	ENSG00000131711
NCOR2	ENSG00000196498

PTPRS	ENSG00000105426
BMPR2	ENSG00000204217
GRB10	ENSG00000106070
SKI	ENSG00000157933
CDK2	ENSG00000123374
CUL3	ENSG00000036257
SMARCC1	ENSG00000173473
KPNB1	ENSG00000108424
RAB27A	ENSG00000069974
TAF4	ENSG00000130699
JUP	ENSG00000173801
RNF111	ENSG00000157450
TRIM28	ENSG00000130726
SAFB	ENSG00000160633
SATB1	ENSG00000182568
YWHAG	ENSG00000170027
GRB2	ENSG00000177885
SYT1	ENSG00000067715
HDAC3	ENSG00000171720
CHUK	ENSG00000213341
BAT3,BAG6	ENSG00000204463
HIRA	ENSG00000100084
YAP1	ENSG00000137693
XRCC5	ENSG00000079246
NCKAP1	ENSG00000061676
IGF2BP1	ENSG00000159217
SP1	ENSG00000185591
HNRNPC	ENSG00000092199
MARK2	ENSG00000072518
H2AFY	ENSG00000113648
RB1CC1	ENSG00000023287
PCNA	ENSG00000132646
TCEA1	ENSG00000187735
CDC5L	ENSG00000096401
VCP	ENSG00000165280
HDAC7	ENSG00000061273
DAPK1	ENSG00000196730
RFX3	ENSG00000080298
USP25	ENSG00000155313
C11orf30	ENSG00000158636
CUL2	ENSG00000108094
VCAN	ENSG00000038427
TCOF1	ENSG00000070814
DVL3	ENSG00000161202
RYR1	ENSG00000196218

USP4	ENSG00000114316
GTF3C1	ENSG00000077235
CDC42	ENSG00000070831
TAB2	ENSG00000055208
KAT7	ENSG00000136504
NRIP1	ENSG00000180530
PRKCD	ENSG00000163932
FXR1	ENSG00000114416
SPTAN1	ENSG00000197694
EHMT1	ENSG00000181090
RERE	ENSG00000142599
UBQLN4	ENSG00000160803
RAE1	ENSG00000101146
CCNA2	ENSG00000145386
MAP3K3	ENSG00000198909
ACTR3	ENSG00000115091
POU2F2	ENSG00000028277
BARD1	ENSG00000138376
RBL1	ENSG00000080839
DLG2	ENSG00000150672
MAP3K4	ENSG00000085511
ADRBK1	ENSG00000173020
ZBTB7A	ENSG00000178951
GNA12	ENSG00000146535
DNM1	ENSG00000106976
PSMD4	ENSG00000159352
TOP2B	ENSG00000077097
TFAP2A	ENSG00000137203
FOS	ENSG00000170345
CRK	ENSG00000167193
NRF1	ENSG00000106459
BCL2L1	ENSG00000171552
CBFA2T2	ENSG00000078699
COPS6	ENSG00000168090
STX1A	ENSG00000106089
GRIA1	ENSG00000155511
SP3	ENSG00000172845
CDH2	ENSG00000170558
SMAD7	ENSG00000101665
GRIP1	ENSG00000155974
PLK1	ENSG00000166851
AMBRA1	ENSG00000110497
XRCC6	ENSG00000196419
UPF1	ENSG00000005007
TBK1	ENSG00000183735

PTK2B	ENSG00000120899
TCF4	ENSG00000196628
MAPK14	ENSG00000112062
SREBF2	ENSG00000198911
BPTF	ENSG00000171634
E2F4	ENSG00000205250
PLEC1,PLEC	ENSG00000178209
ITPR1	ENSG00000150995
UPF2	ENSG00000151461
SIN3B	ENSG00000127511
PRKCQ	ENSG00000065675
PRMT1	ENSG00000126457
VCL	ENSG00000035403
TOPBP1	ENSG00000163781
GNA13	ENSG00000120063
TERF1	ENSG00000147601
HSPA5	ENSG00000044574
ATF7IP	ENSG00000171681
LRP6	ENSG00000070018
INSM1	ENSG00000173404
MECOM	ENSG00000085276
RBM39	ENSG00000131051
MYH11	ENSG00000133392
POLR2A	ENSG00000181222
EPN1	ENSG00000063245
GAB1	ENSG00000109458
USF1	ENSG00000158773
BCAR1	ENSG00000050820
PAFAH1B1	ENSG00000007168
TCF8,ZEB1	ENSG00000148516
BIRC2	ENSG00000110330
RAP1A	ENSG00000116473
UBE2E2	ENSG00000182247
SH3GL2	ENSG00000107295
CDC25A	ENSG00000164045
PIAS4	ENSG00000105229
VAV2	ENSG00000160293
KRT8	ENSG00000170421
THBS1	ENSG00000137801
SSRP1	ENSG00000149136
CCNK	ENSG00000090061
USP32	ENSG00000170832
RPA1	ENSG00000132383
PDPK1	ENSG00000140992
37469	ENSG00000123908

STAT5A	ENSG00000126561
KHDRBS1	ENSG00000121774
ABCD3	ENSG00000117528
SREBF1	ENSG00000072310
MPP6	ENSG00000105926
CAND1	ENSG00000111530
ACTG1	ENSG00000184009
PPP4C	ENSG00000149923
SIRT1	ENSG00000096717
CANX	ENSG00000127022
WWOX	ENSG00000186153
KDM1A	ENSG00000004487
CDKN1A	ENSG00000124762
ARID1B	ENSG00000049618
NR2F1	ENSG00000175745
TFDP1	ENSG00000198176
GRIA2	ENSG00000120251
HCK	ENSG00000101336
YBX1	ENSG00000065978
RIPK1	ENSG00000137275
MARK3	ENSG00000075413
MAPK7	ENSG00000166484
EPC1	ENSG00000120616
SNAP23	ENSG00000092531
SMARCC2	ENSG00000139613
RAPGEF1	ENSG00000107263
PIK3CA	ENSG00000121879
NIPBL	ENSG00000164190
GIT1	ENSG00000108262
MAPK8	ENSG00000107643
TRIP13	ENSG00000071539
MBD2	ENSG00000134046
UBE2N	ENSG00000177889
TRAF2	ENSG00000127191
MAP3K5	ENSG00000197442
NR2C2	ENSG00000177463
STXBP1	ENSG00000136854
UBE2E3	ENSG00000170035
SRF	ENSG00000112658
MGRN1	ENSG00000102858
NCOA6	ENSG00000198646
KIF11	ENSG00000138160
PHF21A	ENSG00000135365
PPARD	ENSG00000112033
HMGB1	ENSG00000189403

ATF4	ENSG00000128272
SRSF1	ENSG00000136450
UBE4B	ENSG00000130939
ATG3	ENSG00000144848
UCHL5	ENSG00000116750
RORA	ENSG00000069667
CTBP1	ENSG00000159692
HR	ENSG00000168453
G3BP2	ENSG00000138757
RBBP8	ENSG00000101773
CCT3	ENSG00000163468
PSMA1	ENSG00000129084
NFIX	ENSG00000008441
ASH2L	ENSG00000129691
MAP4K4	ENSG00000071054
PRKG1	ENSG00000185532
DISC1	ENSG00000162946
CSK	ENSG00000103653
JADE1	ENSG00000077684
HSPA4	ENSG00000170606
USP15	ENSG00000135655
CNOT2	ENSG00000111596
FAF1	ENSG00000185104
CDYL	ENSG00000153046
RAD50	ENSG00000113522
SPRY2	ENSG00000136158
DSP	ENSG00000096696
MYH10	ENSG00000133026
DDX17	ENSG00000100201
STAT2	ENSG00000170581
COPS2	ENSG00000166200
ATF2	ENSG00000115966
DAB2IP	ENSG00000136848
PLCG2	ENSG00000197943
MIB1	ENSG00000101752
RTN4	ENSG00000115310
WASL	ENSG00000106299
KAT2A	ENSG00000108773
WDR82	ENSG00000164091
NFYA	ENSG00000001167
TFAP2B	ENSG00000008196
KAT5	ENSG00000172977
ROCK1	ENSG00000067900
PPP1CB	ENSG00000213639
FANCM	ENSG00000187790

GIGYF2	ENSG00000204120
ESRRG	ENSG00000196482
PHOX2A	ENSG00000165462
LDB1	ENSG00000198728
RFXANK	ENSG00000064490
SRRM2	ENSG00000167978
MTA1	ENSG00000182979
ZNF423	ENSG00000102935
STUB1	ENSG00000103266
COPS5	ENSG00000121022
DLG4	ENSG00000132535
BECN1	ENSG00000126581
SNRPN	ENSG00000128739
BMI1	ENSG00000168283
SKIL	ENSG00000136603
MYC	ENSG00000136997
MAP3K8	ENSG00000107968
THRA	ENSG00000126351
MSX2	ENSG00000120149
RPS6KA2	ENSG00000071242
YY1	ENSG00000100811
NSF	ENSG00000073969
SUMO1	ENSG00000116030
PTPN2	ENSG00000175354
BIRC6	ENSG00000115760
CTBP2	ENSG00000175029
PKN2	ENSG00000065243
EPOR	ENSG00000187266
TRIO	ENSG00000038382
AP2M1	ENSG00000161203
NUP62	ENSG00000213024
ABLIM1	ENSG00000099204
MAP2K7	ENSG00000076984
CHD1	ENSG00000153922
RUNX3	ENSG00000020633
EPHA3	ENSG00000044524
CD247	ENSG00000198821
DAB2	ENSG00000153071
SMARCA5	ENSG00000153147
E2F3	ENSG00000112242
TLN1	ENSG00000137076
CD44	ENSG00000026508
RIPK2	ENSG00000104312
APLP2	ENSG00000084234
DCC	ENSG00000187323

REV3L	ENSG00000009413
PAG1	ENSG00000076641
SCNN1A	ENSG00000111319
RPTOR	ENSG00000141564
FANCL	ENSG00000115392
PRMT5	ENSG00000100462
CSNK2B	ENSG00000204435
GNAI2	ENSG00000114353
TWIST1	ENSG00000122691
DHX15	ENSG00000109606
AP1M1	ENSG00000072958
PXN	ENSG00000089159
EPC2	ENSG00000135999
KDM5B	ENSG00000117139
TSC22D1	ENSG00000102804
SQSTM1	ENSG00000161011
CRMP1	ENSG00000072832
FKBP4	ENSG00000004478
MED13	ENSG00000108510
CRKL	ENSG00000099942
FBXW11	ENSG00000072803
ZNF24	ENSG00000172466
GRIK3	ENSG00000163873
TRIP4	ENSG00000103671
ACTN2	ENSG00000077522
CFLAR	ENSG00000003402
PLEKHA5	ENSG00000052126
DYRK1A	ENSG00000157540
CTNNA1	ENSG00000044115
PPFIA2	ENSG00000139220
SOCS3	ENSG00000184557
SLX4	ENSG00000188827
SYNCRIP	ENSG00000135316
FOXA2	ENSG00000125798
GNB1	ENSG00000078369
ISL1	ENSG00000016082
TNF	ENSG00000232810
MAPK8IP1	ENSG00000121653
CDC37	ENSG00000105401
SNX6	ENSG00000129515
ATG5	ENSG00000057663
PAXIP1	ENSG00000157212
DST	ENSG00000151914
STK25	ENSG00000115694
HDAC5	ENSG00000108840

G3BP1	ENSG00000145907
FZR1	ENSG00000105325
EIF6	ENSG00000242372
CDC20	ENSG00000117399
HSP90B1	ENSG00000166598
ACLY	ENSG00000131473
PRKD1	ENSG00000184304
MARK4	ENSG00000007047
POU4F1	ENSG00000152192
CDK5RAP2	ENSG00000136861
PARP1	ENSG00000143799
ATF6	ENSG00000118217
GATAD2A	ENSG00000167491
FN1	ENSG00000115414
PIK3C2A	ENSG00000011405
SNAP25	ENSG00000132639
EPHB2	ENSG00000133216
AP3B1	ENSG00000132842
SFN	ENSG00000175793
TNKS	ENSG00000173273
UBC	ENSG00000150991
HIC1	ENSG00000177374
PRKRA	ENSG00000180228
PPM1B	ENSG00000138032
WDR48	ENSG00000114742
HELLS	ENSG00000119969
RREB1	ENSG00000124782
CSMD1	ENSG00000183117
MRE11A	ENSG00000020922
ARPC3	ENSG00000111229
PROX1	ENSG00000117707
UBE2E1	ENSG00000170142
XRCC4	ENSG00000152422
HSPD1	ENSG00000144381
TRIM32	ENSG00000119401
BIN1	ENSG00000136717
OTUB1	ENSG00000167770
SIAH2	ENSG00000181788
SCN5A	ENSG00000183873
MCC	ENSG00000171444
DAB1	ENSG00000173406
SCNN1B	ENSG00000168447
NR2F2	ENSG00000185551
ARIH2	ENSG00000177479
CUL5	ENSG00000166266

GATA4	ENSG00000136574
RHEB	ENSG00000106615
AP2B1	ENSG00000006125
PCBD2	ENSG00000132570
MYBL2	ENSG00000101057
RIMS1	ENSG00000079841
ID2	ENSG00000115738
EXOC4	ENSG00000131558
SYVN1	ENSG00000162298
TRIM29	ENSG00000137699
SF1	ENSG00000168066
CHEK1	ENSG00000149554
GLI2	ENSG00000074047
PDCD10	ENSG00000114209
BMP7	ENSG00000101144
MEF2A	ENSG00000068305
CRYAB	ENSG00000109846
NGFR	ENSG00000064300
TRIM2	ENSG00000109654
PKP2	ENSG00000057294
FEZ1	ENSG00000149557
ELAVL1	ENSG00000066044
AHR	ENSG00000106546
ERBB2IP	ENSG00000112851
CDK5	ENSG00000164885
NR5A1	ENSG00000136931
EIF4B	ENSG00000063046
PRRC2A	ENSG00000204469
DOCK1	ENSG00000150760
SF3B2	ENSG00000087365
MAPK10	ENSG00000109339
ARF6	ENSG00000165527
PDCD6IP	ENSG00000170248
MAML1	ENSG00000161021
MTA2	ENSG00000149480
HIPK2	ENSG00000064393
DVL2	ENSG00000004975
RNF41	ENSG00000181852
PRKCE	ENSG00000171132
BTRC	ENSG00000166167
IRAK4	ENSG00000198001
PARD3	ENSG00000148498
CLU	ENSG00000120885
HNRNPM	ENSG00000099783
ATN1	ENSG00000111676

HERC2	ENSG00000128731
DLGAP1	ENSG00000170579
CAV1	ENSG00000105974
ADD1	ENSG00000087274
USP49	ENSG00000164663
TUBB	ENSG00000196230
RAI1	ENSG00000108557
PPARGC1B	ENSG00000155846
SHC1	ENSG00000160691
CHD3	ENSG00000170004
POLD1	ENSG00000062822
NUP205	ENSG00000155561
PRKCA	ENSG00000154229
RFX1	ENSG00000132005
ACVR1B	ENSG00000135503
ACTA,ACTA1	ENSG00000143632
PFN1	ENSG00000108518
ITGA5	ENSG00000161638
NCK2	ENSG00000071051
TUBA1A	ENSG00000167552
IKZF4	ENSG00000123411
CTTN	ENSG00000085733
MST1R	ENSG00000164078
MNAT1	ENSG00000020426
PPP3CA	ENSG00000138814
TARDBP	ENSG00000120948
E2F1	ENSG00000101412
FRS2	ENSG00000166225
MACF1	ENSG00000127603
HES1	ENSG00000114315
TYK2	ENSG00000105397
VDAC1	ENSG00000213585
WDTC1	ENSG00000142784
RGS20	ENSG00000147509
SGK1	ENSG00000118515
PPP5C	ENSG00000011485
ARHGEF2	ENSG00000116584
AIMP2	ENSG00000106305
ACTN4	ENSG00000130402
AKAP13	ENSG00000170776
RBBP6	ENSG00000122257
PSMC4	ENSG00000013275
MBD3	ENSG00000071655
CHAF1A	ENSG00000167670
TEC	ENSG00000135605

TAF5	ENSG00000148835
SLC25A12	ENSG00000115840
MCPH1	ENSG00000147316
PTPN12	ENSG00000127947
FOXP1	ENSG00000114861
UBE2I	ENSG00000103275
RFWD2	ENSG00000143207
HNRNPA1	ENSG00000135486
GTF2IRD1	ENSG00000006704
EPS8	ENSG00000151491
GABPA	ENSG00000154727
TBP	ENSG00000112592
TAF6	ENSG00000106290
JARID2	ENSG00000008083
ANK3	ENSG00000151150
ITCH	ENSG00000078747
CRADD	ENSG00000169372
RAPGEF2	ENSG00000109756
INPPL1	ENSG00000165458
PRPF6	ENSG00000101161
WWP1	ENSG00000123124
PSMD7	ENSG00000103035
INTS1	ENSG00000164880
ZHX1	ENSG00000165156
MED13L	ENSG00000123066
GTF2H1	ENSG00000110768
ERC2	ENSG00000187672
COL4A1	ENSG00000187498
YWHAZ	ENSG00000164924
SETD8	ENSG00000183955
TERF2IP	ENSG00000166848
SLC12A2	ENSG00000064651
SOX6	ENSG00000110693
MCM7	ENSG00000166508
XPO6	ENSG00000169180
ARID4B	ENSG00000054267
ARHGAP32	ENSG00000134909
COL4A3BP	ENSG00000113163
HAND2	ENSG00000164107
SYNE1	ENSG00000131018
LCP2	ENSG00000043462
MCL1	ENSG00000143384
SRCAP	ENSG00000080603
LIMS1	ENSG00000169756
SF3A1	ENSG00000099995

CENPE	ENSG00000138778
CBX4	ENSG00000141582
PHB	ENSG00000167085
PPP1R8	ENSG00000117751
TNKS2	ENSG00000107854
PLAGL1	ENSG00000118495
PTPRF	ENSG00000142949
MAP3K12	ENSG00000139625
SAP130	ENSG00000136715
NECAB2	ENSG00000103154
AP1G1	ENSG00000166747
PHLPP1	ENSG00000081913
GATA6	ENSG00000141448
FADD	ENSG00000168040
PHB2	ENSG00000215021
AEBP2	ENSG00000139154
CXXC1	ENSG00000154832
LAMA4	ENSG00000112769
HSPB1	ENSG00000106211
RBPMS	ENSG00000157110
FBL	ENSG00000105202
DLG5	ENSG00000151208
SMAD6	ENSG00000137834
SETD7	ENSG00000145391
NFE2	ENSG00000123405
UBE2K	ENSG00000078140
FBLN2	ENSG00000163520
TTN	ENSG00000155657
MCM2	ENSG00000073111
SRPK2	ENSG00000135250
UBQLN1	ENSG00000135018
MDFI	ENSG00000112559
ZNRF1	ENSG00000186187
C7orf55-LUC7L2	ENSG00000146963
MAP3K7IP1,TAB1	ENSG00000100324
PAK7	ENSG00000101349
HNRNPU	ENSG00000153187
IQGAP1	ENSG00000140575
LBR	ENSG00000143815
CFL1	ENSG00000172757
PIAS2	ENSG00000078043
ZFR	ENSG00000056097
POLH	ENSG00000170734
KIF3A	ENSG00000131437
CSNK2A2	ENSG00000070770

ERN1	ENSG00000178607
NEDD9	ENSG00000111859
XBP1	ENSG00000100219
ANAPC2	ENSG00000176248
PRPF40A	ENSG00000196504
SUMO3	ENSG00000184900
ITGB2	ENSG00000160255
RRAS2	ENSG00000133818
ZC3H13	ENSG00000123200
SERPING1	ENSG00000149131
PTGS2	ENSG00000073756
PRKCZ	ENSG00000067606
SMARCE1	ENSG00000073584
RNF185	ENSG00000138942
MAPKAP1	ENSG00000119487
DTNBP1	ENSG00000047579
CDK1	ENSG00000170312
DAP3	ENSG00000132676
HAP1	ENSG00000173805
NCBP1	ENSG00000136937
NOS1	ENSG00000089250
ZMYM2	ENSG00000121741
BANP	ENSG00000172530
TJP1	ENSG00000104067
FKBP8	ENSG00000105701
CDC42BPB	ENSG00000198752
ESRRA	ENSG00000173153
GRID2	ENSG00000152208
BTBD2	ENSG00000133243
ZMYND11	ENSG00000015171
IPO5	ENSG00000065150
ESR2	ENSG00000140009
PDS5A	ENSG00000121892
NR4A3	ENSG00000119508
MCM6	ENSG00000076003
37104	ENSG00000092847
MED15	ENSG00000099917
FHL3	ENSG00000183386
VIM	ENSG00000026025
EEF1A1	ENSG00000156508
ITGA2B	ENSG00000005961
RBX1	ENSG00000100387
IGF1	ENSG00000017427
TRAF1	ENSG00000056558
EFNA5	ENSG00000184349

TBX21	ENSG00000073861
MAPKAPK2	ENSG00000162889
COPS3	ENSG00000141030
PACSIN1	ENSG00000124507
ANKS1B	ENSG00000185046
UBE2D3	ENSG00000109332
FOXK1	ENSG00000164916
TCERG1	ENSG00000113649
USP3	ENSG00000140455
MGA	ENSG00000174197
SNRNP200	ENSG00000144028
RBM4	ENSG00000173933
NR4A2	ENSG00000153234
COPG1	ENSG00000181789
CTR9	ENSG00000198730
EPHB1	ENSG00000154928
CAPRIN1	ENSG00000135387
MAP2K5	ENSG00000137764
RAD9A	ENSG00000172613
CNOT1	ENSG00000125107
RACGAP1	ENSG00000161800
MAGI1	ENSG00000151276
ASF1A	ENSG00000111875
GNAO1	ENSG00000087258
DTNB	ENSG00000138101
MKLN1	ENSG00000128585
UBR1	ENSG00000159459
RARB	ENSG00000077092
CNTN1	ENSG00000018236
GCN1L1	ENSG00000089154
NUP153	ENSG00000124789
TP53BP1	ENSG00000067369
PTPRJ	ENSG00000149177
MYOCD	ENSG00000141052
RFC3	ENSG00000133119
NELFB	ENSG00000188986
HEXIM1	ENSG00000186834
ARRB1	ENSG00000137486
KHSRP	ENSG00000088247
TGIF1,TGIF	ENSG00000177426
PPFIA1	ENSG00000131626
BAG3	ENSG00000151929
FASLG	ENSG00000117560
HNRNPD	ENSG00000138668
DDX23	ENSG00000174243

MED23	ENSG00000112282
DNAJA3	ENSG00000103423
EPB41L2	ENSG00000079819
PAK1	ENSG00000149269
GUCY1A2	ENSG00000152402
IL1R1	ENSG00000115594
FHL2	ENSG00000115641
IGSF21	ENSG00000117154
VPRBP	ENSG00000145041
USP53	ENSG00000145390
MBIP	ENSG00000151332
STRN4	ENSG00000090372
KLC2	ENSG00000174996
APPBP2	ENSG00000062725
OTUD4	ENSG00000164164
SRRT	ENSG00000087087
GRIK2	ENSG00000164418
ZNF638	ENSG00000075292
MORF4L1	ENSG00000185787
MGMT	ENSG00000170430
ODC1	ENSG00000115758
TUBB2A	ENSG00000137267
RNF11	ENSG00000123091
SOCS6	ENSG00000170677
CDKN1C	ENSG00000129757
UBE2D2	ENSG00000131508
SAP30L	ENSG00000164576
EVL	ENSG00000196405
ATXN1	ENSG00000124788
INO80	ENSG00000128908
DRD2	ENSG00000149295
SSBP2	ENSG00000145687
COPS8	ENSG00000198612
PPP2R5C	ENSG00000078304
RPS19	ENSG00000105372
NUMB	ENSG00000133961
PKM	ENSG00000067225
RND3	ENSG00000115963
GNAI1	ENSG00000127955
GAB2	ENSG00000033327
PRLR	ENSG00000113494
SF3B3	ENSG00000189091
PIK3C2B	ENSG00000133056
AURKA	ENSG00000087586
FLT1	ENSG00000102755

IL2RB	ENSG00000100385
MAD1L1	ENSG00000002822
ESRRB	ENSG00000119715
ETS2	ENSG00000157557
EEF1A2	ENSG00000101210
TPI1	ENSG00000111669
ING4	ENSG00000111653
F2	ENSG00000180210
EHMT2	ENSG00000204371
NFATC1	ENSG00000131196
ATXN7	ENSG00000163635
MAPK9	ENSG00000050748
CBX7	ENSG00000100307
KIFAP3	ENSG00000075945
SLC9A3R1	ENSG00000109062
CTNND2	ENSG00000169862
ZBTB17	ENSG00000116809
HDLBP	ENSG00000115677
ZMIZ1	ENSG00000108175
STC2	ENSG00000113739
ILK	ENSG00000166333
FXR2	ENSG00000129245
PRKAR2A	ENSG00000114302
MCM4	ENSG00000104738
NEDD4L	ENSG00000049759
SNCB	ENSG00000074317
COMMD1	ENSG00000173163
ELF3	ENSG00000163435
WIPF1	ENSG00000115935
GEMIN4	ENSG00000179409
SNAPIN	ENSG00000143553
TK1	ENSG00000167900
SMAD5	ENSG00000113658
RBBP4	ENSG00000162521
ZBTB7B	ENSG00000160685
RAD54L2	ENSG00000164080
STK39	ENSG00000198648
SYNJ1	ENSG00000159082
APLP1	ENSG00000105290
RNF216	ENSG00000011275
STARD13	ENSG00000133121
PSMC3	ENSG00000165916
CD19	ENSG00000177455
AP2A2	ENSG00000183020
ANAPC7	ENSG00000196510

NEK6	ENSG00000119408
KDM2B	ENSG00000089094
PSEN2	ENSG00000143801
ARRB2	ENSG00000141480
PSMD11	ENSG00000108671
VEGFA	ENSG00000112715
KCNA1	ENSG00000111262
CAPNS1	ENSG00000126247
PARD3B	ENSG00000116117
ONECUT1	ENSG00000169856
SNAI1	ENSG00000124216
GAK	ENSG00000178950
MEPCE	ENSG00000146834
RABEP1	ENSG00000029725
MYOG	ENSG00000122180
PCBP2	ENSG00000197111
TEK	ENSG00000120156
MED24	ENSG00000008838
RAD23B	ENSG00000119318
ATG16L1	ENSG00000085978
DTL	ENSG00000143476
STK24	ENSG00000102572
VAMP2	ENSG00000220205
BRD1	ENSG00000100425
KAT8	ENSG00000103510
SEL1L	ENSG00000071537
PPP1R12A	ENSG00000058272
PHLDA3	ENSG00000174307
KPNA6	ENSG00000025800
CDC25B	ENSG00000101224
PPP2CA	ENSG00000113575
C1QBP	ENSG00000108561
SUPT3H	ENSG00000196284
SORT1	ENSG00000134243
UBE2U	ENSG00000177414
PRPF31	ENSG00000105618
BLMH	ENSG00000108578
MARK1	ENSG00000116141
PRKAR1B	ENSG00000188191
PPM1G	ENSG00000115241
FOSL1	ENSG00000175592
MAD2L1	ENSG00000164109
KPNA1	ENSG00000114030
NDEL1	ENSG00000166579
GRN	ENSG00000030582

RICTOR	ENSG00000164327
TIAM1	ENSG00000156299
IGF2BP2	ENSG00000073792
MAP2	ENSG00000078018
ARHGEF7	ENSG00000102606
GNB2L1	ENSG00000204628
IGF2BP3	ENSG00000136231
PSMD1	ENSG00000173692
SERPINA1	ENSG00000197249
HGF	ENSG00000019991
FKBP1A	ENSG00000088832
HSPG2	ENSG00000142798
ZBTB9	ENSG00000213588
USF2	ENSG00000105698
MYO7A	ENSG00000137474
IRF5	ENSG00000128604
FBXW8	ENSG00000174989
YWHAB	ENSG00000166913
RPLP1	ENSG00000137818
SOS2	ENSG00000100485
S100A8	ENSG00000143546
NCDN	ENSG00000020129
WASF2	ENSG00000158195
RAD51	ENSG00000051180
TRA2B	ENSG00000136527
SLC3A2	ENSG00000168003
SPECC1L	ENSG00000100014
BAK1	ENSG00000030110
NPAS2	ENSG00000170485
AP1B1	ENSG00000100280
SIX1	ENSG00000126778
HSD11B2	ENSG00000176387
SIX3	ENSG00000138083
NR3C2	ENSG00000151623
CSF1	ENSG00000184371
FBP1	ENSG00000165140
CCT8	ENSG00000156261
STAMBP	ENSG00000124356
EFEMP2	ENSG00000172638
TNNC1	ENSG00000114854
ACVR2A	ENSG00000121989
E4F1	ENSG00000167967
GRB14	ENSG00000115290
APBA1	ENSG00000107282
RBFOX2	ENSG00000100320

LAT	ENSG00000213658
CCDC85B	ENSG00000175602
ARHGAP17	ENSG00000140750
PPP2R2C	ENSG00000074211
UBE2Z	ENSG00000159202
FAS	ENSG00000026103
NOTCH3	ENSG00000074181
TDG	ENSG00000139372
USP1	ENSG00000162607
IRF8	ENSG00000140968
HIVEP2	ENSG00000010818
UBR4	ENSG00000127481
LYST	ENSG00000143669
TUBG1	ENSG00000131462
MYCBP2	ENSG00000005810
GRAP2	ENSG00000100351
CIT	ENSG00000122966
WDR5	ENSG00000196363
PABPC1	ENSG00000070756
PPP1R10	ENSG00000204569
CD28	ENSG00000178562
GRK5	ENSG00000198873
TNRC6B	ENSG00000100354
TAF1B	ENSG00000115750
PPP2R5D	ENSG00000112640
CCDC101	ENSG00000176476
REST	ENSG00000084093
CLNS1A	ENSG00000074201
KDM6B	ENSG00000132510
FGF2	ENSG00000138685
LSM4	ENSG00000130520
APBB1	ENSG00000166313
NPEPPS	ENSG00000141279
MAP3K11	ENSG00000173327
UBA5	ENSG00000081307
CSF1R	ENSG00000182578
RALBP1	ENSG00000017797
TTBK2	ENSG00000128881
LINGO1	ENSG00000169783
CADPS	ENSG00000163618
GFAP	ENSG00000131095
POLB	ENSG00000070501
PTPRA	ENSG00000132670
UCHL1	ENSG00000154277
RRN3	ENSG00000085721

DMAP1	ENSG00000178028
UIMC1	ENSG00000087206
MNT	ENSG00000070444
EGLN1	ENSG00000135766
FTH1	ENSG00000167996
SIRT6	ENSG00000077463
FANCI	ENSG00000140525
ATP2B2	ENSG00000157087
NR1H4	ENSG00000012504
SPATA2	ENSG00000158480
PRSS23	ENSG00000150687
PTPRZ1	ENSG00000106278
TEAD1	ENSG00000187079
SIAH1	ENSG00000196470
DDX1	ENSG00000079785
RCOR1	ENSG00000089902
MEF2D	ENSG00000116604
PHYHIP	ENSG00000168490
USP42	ENSG00000106346
NDN	ENSG00000182636
RUVBL2	ENSG00000183207
ADRM1	ENSG00000130706
CABIN1	ENSG00000099991
CHFR	ENSG00000072609
ABI2	ENSG00000138443
RNF165	ENSG00000141622
POGZ	ENSG00000143442
IKZF2	ENSG00000030419
HCN1	ENSG00000164588
FES	ENSG00000182511
CDH10	ENSG00000040731
BUB3	ENSG00000154473
SPI1	ENSG00000066336
SUPT5H	ENSG00000196235
ORC2	ENSG00000115942
KCNA3	ENSG00000177272
ZZZ3	ENSG00000036549
PSMC6	ENSG00000100519
LNX1	ENSG00000072201
RPN2	ENSG00000118705
JUNB	ENSG00000171223
NPHP1	ENSG00000144061
IL1RAP	ENSG00000196083
EFTUD2	ENSG00000108883
PHC2	ENSG00000134686

PIK3C3	ENSG00000078142
ANKS1A	ENSG00000064999
CBX3	ENSG00000122565
SUPT16H	ENSG00000092201
USP14	ENSG00000101557
UBE2H	ENSG00000186591
CTCFL	ENSG00000124092
AMIGO1	ENSG00000181754
ABI3	ENSG00000108798
MMP2	ENSG00000087245
HNRNPK	ENSG00000165119
ATG4B	ENSG00000168397
NCAN	ENSG00000130287
TCEA2	ENSG00000171703
SH3BP2	ENSG00000087266
USP22	ENSG00000124422
CD2	ENSG00000116824
DBNL	ENSG00000136279
INTS5	ENSG00000185085
SETD1A	ENSG00000099381
TRPC4AP	ENSG00000100991
SMC3	ENSG00000108055
ILF2	ENSG00000143621
HADHA	ENSG00000084754
KBTBD7	ENSG00000120696
ASCC3	ENSG00000112249
HIPK1	ENSG00000163349
HSF1	ENSG00000185122
TLK1	ENSG00000198586
KCNJ3	ENSG00000162989
CNTFR	ENSG00000122756
YLPM1	ENSG00000119596
USP12	ENSG00000152484
SNRPD3	ENSG00000100028
MMP14	ENSG00000157227
CCNB1	ENSG00000134057
APAF1	ENSG00000120868
VCPIP1	ENSG00000175073
DPPA2	ENSG00000163530
PAPOLA	ENSG00000090060
DLGAP4	ENSG00000080845
SSB	ENSG00000138385
RAD17	ENSG00000152942
BAD	ENSG00000002330
INPP5D	ENSG00000168918

CEP72	ENSG00000112877
XRCC1	ENSG00000073050
RPS3A	ENSG00000145425
MEIS1	ENSG00000143995
SLC9A1	ENSG00000090020
TANK	ENSG00000136560
BACH2	ENSG00000112182
ACTC1,ACTC	ENSG00000159251
RNF10	ENSG00000022840
HNRNPUL1	ENSG00000105323
CTTNBP2	ENSG00000077063
MTPN	ENSG00000105887
DPY30	ENSG00000162961
SHANK1	ENSG00000161681
PIAS3	ENSG00000131788
PAK4	ENSG00000130669
MAP3K10	ENSG00000130758
AMFR	ENSG00000159461
LRP8	ENSG00000157193
CLCN3	ENSG00000109572
RAB1A	ENSG00000138069
GLUL	ENSG00000135821
SKP1	ENSG00000113558
PTGES3	ENSG00000110958
PSME3	ENSG00000131467
MED16	ENSG00000175221
ING3	ENSG00000071243

Table A.4: The list of candidate genes ranked by their likelihood to be cancer–associated as predicted by the model

Appendix B

The code (in Python programming language) is stand-alone module that can be executed to make predictions based on our machine-learning model built to predict cancer-associated replacements for the protein PTEN using Physico-chemical properties of the amino acid residues:

```
import calendar
from datetime import datetime
from collections import namedtuple
import re
import sys
import time
import os
```

```
import numpy as np
import pandas as pd

PY3 = sys.version_info[0] == 3
if PY3:
   string_types = str,
   text_type = str
   long_type = int
else:
   string_types = basestring,
   text_type = unicode
   long_type = long

def predict(row):
   round_ASAD = np.float32(row[u'ASAD'])
   round_El = np.float32(row[u'El'])
   round_Et = np.float32(row[u'Et'])
   round_F = np.float32(row[u'F'])
   round_GhD = np.float32(row[u'GhD'])
   round_Hgm = np.float32(row[u'Hgm'])
   round_Hnc = np.float32(row[u'Hnc'])
   round_Ht = np.float32(row[u'Ht'])
   round_Location = np.float32(row[u'Location'])
   round_Mu = np.float32(row[u'Mu'])
   round_Ns = np.float32(row[u'Ns'])
   round_Pb = np.float32(row[u'Pb'])
   round_Pc = np.float32(row[u'Pc'])
   round_Pf_s = np.float32(row[u'Pf_s'])
   round_Pt = np.float32(row[u'Pt'])
   round_Ra = np.float32(row[u'Ra'])
   round_Rf = np.float32(row[u'Rf'])
   round_aC = np.float32(row[u'aC'])
   round_dG = np.float32(row[u'dG'])
   round_dGh = np.float32(row[u'dGh'])
   round_dH = np.float32(row[u'dH'])
   round_pK__ = np.float32(row[u'pK\''])
   return sum([
      0.4980069,
       -0.030802308502971359472 * (round_Location > 154.5 and
                           round_Pb <= 0.4950000047683716 and
                           round_Ra > -0.11499999463558197 and
                           round_aC <= -0.009999999776482582),
        0.042897671247850641119 * (round_pK__ <= 0.9199999570846558 and
                           round_El <= -0.2850000262260437 and
                           round_ASAD > -0.2900000214576721 and
                           round_Pf_s > -0.9049999713897705),
```

0.046376835827466024453 * (round_Location <= 258.5),
-0.0089180017527997167137 * (round_Et <= 0.12999999523162842 and
round_Pc <= 0.48500001430511475 and
round_Pf_s > -0.9049999713897705),
0.010488690939436063829 * (round_Location <= 416.5 and
round_pK__ > -0.5450000166893005 and
round_Pb <= 0.4950000047683716),
0.065823940039791375978 * (round_Pc > -0.7649999856948853 and
round_ASAD > 0.3450000286102295 and
-0.9049999713897705 < round_Pf_s <= 0.019999999552965164),
-0.12814740185713269227 * (round_Pc <= 0.6650000214576721 and
round_ASAD <= -0.19499999284744263 and
round_dGh <= 0.004999999888241291 and
round_dG > -0.2850000262260437),
0.017950279696725548323 * (-0.2750000059604645 < round_ASAD <=
0.25999999046325684 and
round_Pf_s > 0.019999999552965164),
-0.006761195667625109 9355 * (round_Location <= 2204.5 and
round_Pt <= 0.1550000011920929 and
round_F > -0.1599999964237213 and
round_dG > -0.6349999904632568),
0.040932779830598259307 * (round_El <= -0.07500000298023224 and
round_Ns > -0.19499999284744263),
0.0044318155064351667793 * (round_Location > 154.5 and
round_El > 0.08500000089406967 and
round_Pb <= 0.4950000047683716 and
round_aC > -0.009999999776482582),
0.037460725728313236382 * (round_F <= -0.5699999928474426 and
round_dH <= 0.7200000286102295),
0.010111500425799288885 * (round_Location > 2195.5 and
round_dH > -0.6349999904632568),
-0.063290952377108716798 * (round_Location <= 2204.5 and
round_F > -0.6100000143051147 and
round_Ns <= -0.47499996423721313 and
round_Pf_s <= 0.16499999165534973),
0.1424292742484439267 * (round_Ht <= 0.1550000011920929 and
round_pK__ <= 0.9199999570846558 and
round_Rf <= -0.02499999850988388 and
round_Pb > -0.6449999809265137),
-0.14147827129531662105 * (round_Et <= 0.044999998062849045 and
-0.4350000023841858 < round_ASAD <= -0.2900000214576721),
-0.048133563780341961924 * (281.5 < round_Location <= 2204.5 and
round_pK__ <= 0.9199999570846558 and
round_Rf <= 0.4950000047683716),
0.028140576211365345843 * (round_Location <= 279.0 and
round_ASAD > -0.19499999284744263),
0.030988185279260817284 * (round_Rf > -0.48000001907348633 and

round_Pb > 0.22499999403953552 and
round_Pt <= -0.5149999856948853),
-0.011191993168808238995 * (round_Location <= 1988.5 and
round_Rf > 0.16499999165534973 and
round_Pb <= 0.42500001192092896 and
round_dGh <= 0.6650000214576721),
0.025234167909347456765 * (round_pK__ <= 0.9199999570846558 and
round_Rf > -0.02499999850988388 and
round_Et <= -0.004999999888241291 and
round_Pb > 0.029999999329447746),
0.007243956293122907 2029 * (round_Location <= 279.0 and
round_Pb > 0.22499999403953552),
0.09790673330750181147 * (round_El > -0.07500000298023224 and
round_Pb > -0.07500000298023224 and
round_Hgm <= 0.20499999821186066 and
round_dG > 0.4449999928474426),
0.0042433454460802776803 * (round_Pb > 0.4950000047683716 and
round_dG > 0.019999999552965164),
0.049099205879184981693 * (round_Rf <= -0.33500000834465027 and
round_Hgm > -0.5649999976158142),
0.008451012591256135 1313 * (round_F > 0.10500000417232513 and
round_dGh > 0.14500001072883606 and
round_dG <= 0.3199999928474426 and
round_Pf_s <= 0.16499999165534973),
0.009996497362710493606 * (round_Location <= 416.5 and
round_Rf > -0.8350000381469727 and
round_Pb > -0.6449999809265137),
-0.039037863454884545733 * (282.5 < round_Location <= 2204.5 and
round_aC > -0.6100000143051147 and
round_dGh <= 0.42000001668930054),
0.0225966234597955902 * (round_Location <= 416.5),
-0.089558222847778604092 * (1992.5 < round_Location <= 2192.5 and
round_Pf_s > -0.9049999713897705),
0.12237007610425067183 * (round_pK__ <= 0.49000000953674316 and
round_Ra > -0.9149999618530273 and
round_Ns > -0.6100000143051147 and
round_Pf_s <= 0.9049999713897705),
-0.005185792929359487757795 * (154.5 < round_Location <= 2204.5 and
round_Pc <= 0.6499999761581421 and
round_Pf_s <= 0.9049999713897705),
0.010789485809534872865 * (round_Location <= 377.5),
0.013917897260441604301 * (round_Location <= 1896.5 and
round_Et <= 0.1599999964237213 and
round_Pb <= 0.42500001192092896 and
round_Pf_s > 0.16499999165534973),
-0.011369883981885960458 * (round_Location <= 2204.5 and
round_Pb <= 0.42500001192092896 and

round_Pt <= 0.23499999940395355 and
round_dG > -0.04999999701976776),
0.034270415470539571101 * (round_Rf > -0.48500001430511475 and
round_F > -0.800000011920929 and
round_aC > -0.5349999666213989 and
round_Pf_s <= -0.019999999552965164),
-0.010046176632626116834 * (416.5 < round_Location <= 2195.5),
0.006327373734122383656908 * (round_Location <= 2001.5 and
round_Pt <= 0.1550000011920929 and
round_Ns > -0.47499996423721313),
0.01798394555249511334 * (round_Rf > -0.9449999928474426 and
round_aC <= -0.6100000143051147 and
round_ASAD > -0.23499999940395355),
0.068794197016393637822 * (round_Rf <= -0.3700000047683716 and
round_Hgm > -0.5649999976158142 and
round_GhD > -0.9950000047683716 and
round_Pf_s > -0.9049999713897705),
-0.13528399324496856448 * (154.5 < round_Location <= 2204.5 and
round_Pc <= 0.6499999761581421 and
round_Pf_s > -0.9049999713897705),
0.10809391439486315534 * (round_Rf <= -0.3700000047683716 and
round_Pf_s > 0.1899999976158142),
0.011045291581269784525 * (round_Hgm <= 0.20499999821186066 and
round_GhD > 0.08500000089406967 and
round_dG <= 0.7749999761581421),
-0.0060701138138203047934 * (416.5 < round_Location <= 2195.5 and
round_Hnc <= 0.1599999964237213),
-0.028284000600343469495 * (round_Rf > -0.33500000834465027 and
round_El > -0.07500000298023224 and
round_Pb <= 0.22499999403953552 and
round_dG <= 0.7749999761581421),
0.009522087525250418071 9 * (round_El <= 0.019999999552965164 and
round_aC <= -0.6100000143051147),
-0.009925595835525755084 * (round_Location <= 2204.5 and
0.07500000298023224 < round_Rf <= 0.7999999523162842 and
round_ASAD > -0.23499999940395355),
0.077223008851680999265 * (round_Pb > -0.4699999988079071 and
round_Pt <= 0.1550000011920929 and
round_ASAD > -0.1599999964237213 and
round_dH > 0.014999999664723873),
0.040973694470871284412 * (round_Ht > 0.11500000208616257 and
round_El <= -0.07500000298023224 and
round_Pb <= 0.4950000047683716),
0.014449552380014107911 * (round_aC <= -0.6100000143051147 and
round_ASAD > -0.30000001192092896),
0.006926899982391369024 7 * (round_pK__ > -0.5450000166893005 and
round_Hnc > -0.3999999761581421 and

round_ASAD > -0.23499999940395355 and
round_Pf_s > -0.9049999713897705),
-0.006988346910078486844 1 * (round_Location <= 2204.5 and
round_Ra > -0.8799999952316284 and
round_Ns <= -0.5199999809265137 and
round_Pf_s <= 0.17000000178813934),
0.007759675463091347119 6 * (round_Rf <= 0.16499999165534973 and
round_Pb <= 0.4950000047683716 and
round_Pt > -0.8100000023841858 and
round_F > -0.6100000143051147),
0.000341819984989362739 19 * (round_pK__ <= 0.9199999570846558 and
round_Rf <= -0.02499999850988388 and
round_Hnc > -0.3149999976158142 and
round_Pb > -0.6449999809265137),
-0.043529818912104306 65 * (round_pK__ > -0.10500000417232513 and
round_El > -0.4599999785423279 and
round_aC > -0.6100000143051147 and
round_dG > -0.6349999904632568),
-0.035101250537342622 93 * (round_F > -0.5699999928474426 and
round_dGh > -0.3700000047683716 and
round_dG <= 0.7749999761581421 and
round_dH > -0.014999999664723873),
-0.004167908988450206 7836 * (394.0 < round_Location <= 2204.5 and
round_aC > -0.6050000190734863 and
round_dG > -0.6349999904632568),
0.033190250729287856 801 * (round_Pb > 0.42500001192092896 and
round_dG > -0.03999999910593033),
0.005405078197908565 1963 * (round_Rf > -0.23499998450279236 and
round_Pb > 0.24500000476837158 and
round_Hgm <= 0.5649999976158142),
0.099557301263992037 388 * (round_Location <= 1992.5 and
round_pK__ <= 0.9199999570846558 and
round_Ra > -0.29500001668930054 and
round_Pf_s <= 0.9049999713897705),
0.001335826565963667 6687 * (round_Location <= 1992.5 and
round_pK__ <= 0.9199999570846558 and
round_Rf <= -0.02499999850988388),
0.011107770396153940 698 * (round_Pb > 0.22499999403953552 and
round_Hgm <= 0.5649999976158142 and
round_dGh > -0.6349999904632568 and
round_dG <= 0.7749999761581421),
-0.020351499235991343 112 * (round_Mu > -0.29500001668930054 and
round_Et > -0.5299999713897705 and
round_Ns <= -0.375 and
round_dGh <= -0.05000000074505806),
-0.000799418945013998243 25 * (round_Rf <= 0.33500000834465027 and
round_El > -0.07500000298023224 and

round_Hgm <= 0.20499999821186066 and
round_dG <= 0.7749999761581421),
0.038949967269597665642 * (round_Rf <= -0.48000001907348633 and
round_Pb <= 0.22499999403953552),
-0.088995470545923466288 * (281.5 < round_Location <= 2204.5 and
round_ASAD > -0.4350000023841858 and
round_dGh <= 0.6650000214576721),
0.0088683904707671963596 * (round_Location <= 1885.0 and
round_Hgm <= 0.054999999701976776 and
round_Pf_s > 0.16499999165534973),
0.026765527618360561435 * (round_Location > 2198.5),
0.039436429950153853441 * (round_Et <= -0.5299999713897705 and
round_aC > -0.6100000143051147 and
round_dH > -0.2849999964237213),
0.035212846696491106879 * (round_El <= -0.07500000298023224 and
round_dGh > -0.10000000149011612 and
round_dG <= 0.7749999761581421),
-0.031068735593921953386 * (282.5 < round_Location <= 2204.5 and
round_aC > -0.6100000143051147 and
round_dG <= 0.7749999761581421),
0.024891376636083796525 * (round_pK__ <= 0.9199999570846558 and
round_Pb <= 0.45499998331069946 and
round_GhD > 0.0949999988079071 and
round_dH <= -0.04500000178813934),
0.013371250298207711452 * (round_Location <= 1992.5 and
round_pK__ > -0.5450000166893005 and
round_Pc > -0.7100000381469727 and
round_F > -0.6100000143051147),
-0.006542982017505727040 9 * (377.5 < round_Location <= 2204.5 and
round_pK__ > -0.7250000238418579 and
round_Rf <= 0.4950000047683716),
1.069641997122267929 * (round_Location > 2204.5),
0.013953543892781777869 * (round_Location <= 280.5 and
round_Rf > -0.48000001907348633 and
round_Pb > 0.22499999403953552),
0.010287356142774180603 * (round_El <= -0.07500000298023224 and
round_Hgm <= 0.20499999821186066 and
round_ASAD > -0.2850000262260437 and
round_dG <= 0.7749999761581421),
-0.002516653526546320639 9 * (round_Location <= 1992.5 and
round_Pb <= 0.4950000047683716 and
round_Ra > 0.08500000089406967 and
round_Hgm > 0.39499998092651367),
-0.000521007995220688284 28 * (154.5 < round_Location <= 2195.5 and
round_El > -0.07500000298023224 and
round_Pb > -0.6449999809265137),
-0.056923511123344788798 * (1992.5 < round_Location <= 2195.5 and

round_Rf <= 0.8700000047683716),
0.08565220537901134523 2 * (round_Rf <= -0.48000001907348633),
0.014301444746502533362 * (round_Location <= 1989.0 and
round_Rf > -0.25 and
round_Pb > 0.22499999403953552),
-0.047338014981209056153 * (154.5 < round_Location <= 2204.5 and
round_Et > -0.15000000596046448 and
round_F > -0.5699999928474426),
-0.008524798749305218664 7 * (round_Location <= 1986.5 and
round_Rf > -0.33500000834465027 and
round_Pb <= 0.4950000047683716 and
round_dG > -0.6349999904632568),
-0.20455289112817254264 * (1992.5 < round_Location <= 2204.5),
-0.022589588473050835338 * (round_Location <= 1989.0 and
-0.33500000834465027 < round_Rf <= -0.25 and
round_dG > -0.5199999809265137),
0.018830236278526858024 * (round_pK__ <= 0.26500001549720764 and
round_Mu <= 0.1850000023841858 and
round_Hgm > 0.20499999821186066),
-0.058098542791413770869 * (round_Mu <= 0.38999998569488525 and
round_dGh <= 0.6650000214576721 and
round_dG > -0.6349999904632568 and
round_Pf_s <= 0.019999999552965164),
0.029005188341247021416 * (round_Ra > 0.41999998688697815 and
round_dGh > 0.42000001668930054 and
round_Pf_s > -0.9049999713897705),
0.001518378588913777901 8 * (round_F <= -0.6100000143051147),
0.19263970542116704014 * (round_Location <= 154.5),
-0.000385870438944194214 63 * (round_Et > 0.2150000035762787 and
round_Pc <= 0.48500001430511475 and
round_dGh <= 0.5950000286102295 and
round_Pf_s > -0.9049999713897705),
0.009655527830533370062 2 * (round_Location <= 1885.0 and
round_El > -0.2900000214576721 and
round_Ns > -0.47499996423721313 and
round_Pf_s <= 0.16499999165534973),
0.008213762429026202904 8 * (round_El > -0.07500000298023224 and
round_Hgm > -0.09000000357627869),
0.004815378258631749196 2 * (round_Location > 2195.5),
0.017851259894991602928 * (round_Location <= 2204.5 and
round_Rf <= 0.07500000298023224 and
round_ASAD > -0.23499999940395355),
0.076251239162951206518 * (round_Et <= 0.2150000035762787 and
round_Pb <= 0.4950000047683716 and
round_Pt > -0.8100000023841858 and
round_Ra > 0.125),
0.017198841857197066235 * (round_Rf > 0.4950000047683716 and

round_Pb > 0.44999998807907104 and
round_dG > -0.7099999785423279),
-0.020340021109107105091 * (154.5 < round_Location <= 2204.5 and
round_pK__ <= 0.9199999570846558 and
round_Pf_s > -0.9049999713897705),
-0.027837306550162799895 * (157.5 < round_Location <= 2204.5 and
round_Rf > -0.48000001907348633 and
round_Pb <= 0.22499999403953552),
-0.034295791418998688993 * (round_pK__ > -0.05999999865889549 and
round_Rf > -0.48000001907348633 and
round_Pb <= 0.22499999403953552 and
round_dG > -0.6349999904632568),
0.027011535354938012721 * (round_Rf > 0.0 and
round_Hgm > -0.14000000059604645 and
round_dG <= 0.4449999928474426),
0.013282405637632511627 * (round_Mu <= 0.17499999701976776 and
round_Et > -0.5299999713897705 and
round_aC > -0.6100000143051147 and
round_dGh > 0.03500000014901161),
0.017807046800771823836 * (round_dG <= 0.054999999701976776 and
round_Pf_s > 0.16499999165534973),
-0.0045265526214365754687 * (round_Rf > -0.04500000178813934 and
round_Hgm <= 0.20499999821186066 and
round_GhD <= 0.08500000089406967 and
round_dG <= 0.7749999761581421),
0.031127545432432054962 * (-0.9199999570846558 < round_pK__ <=
0.9199999570846558 and
round_Rf > -0.3700000047683716 and
round_ASAD > -0.2900000214576721),
0.054203531691976995777 * (round_El <= 0.2199999988079071 and
round_Hgm > 0.20499999821186066 and
round_dG <= 0.7749999761581421),
-0.01819028973713624972 * (round_Location > 154.5 and
round_El > -0.07000000029802322 and
round_F > -0.5699999928474426 and
round_dH <= -0.014999999664723873),
-0.053709419863356297475 * (274.5 < round_Location <= 2204.5 and
round_ASAD <= -0.19499999284744263 and
round_dG > -0.2850000262260437),
-0.027995627466905193687 * (round_Rf <= 0.8700000047683716 and
round_Pt <= 0.1550000011920929 and
round_F > -0.5699999928474426 and
round_dG <= 0.7749999761581421),
-0.066174678051963795045 * (round_Et > -0.14500001072883606 and
round_Pb <= 0.39499998092651367 and
round_F > -0.5699999928474426 and
round_dG > -0.04500000178813934),

-0.13690034456909391802 * (1878.5 < round_Location <= 2195.5),
-0.033542540857916261499 * (round_Pb <= 0.22499999403953552 and
round_Ra <= 0.4449999928474426 and
round_dG <= 0.7749999761581421 and
round_Pf_s <= 0.1899999976158142),
0.049377582365736583103 * (round_Rf <= -0.3700000047683716 and
round_Mu > -0.15000000596046448 and
round_GhD > -0.9950000047683716 and
round_dG > -0.6349999904632568),
0.20817070252243766171 * (round_Location <= 282.5 and
round_pK__ <= 0.9199999570846558),
-0.0086144169931692635839 * (round_Pc <= 0.48500001430511475 and
round_dG <= 0.20499999821186066 and
-0.9049999713897705 < round_Pf_s <= 0.10500000417232513),
0.12896496243464933285 * (round_Rf > -0.9449999928474426 and
round_El > 0.08500000089406967 and
round_aC > 0.06499999761581421 and
round_ASAD > -0.23499999940395355),
-0.018573264286127166151 * (round_dG > -0.6349999904632568 and
round_dH <= 0.2549999952316284 and
-0.9049999713897705 < round_Pf_s <= 0.1550000011920929),
-0.12628998419694242861 * (round_Location <= 2204.5 and
round_Pb <= 0.4950000047683716 and
round_Hgm > -0.33500000834465027 and
round_dH > -0.48500001430511475),
-0.072250280355614801553 * (157.5 < round_Location <= 2204.5 and
round_aC > -0.6100000143051147 and
round_dG > -0.6349999904632568),
-0.054527218345746371331 * (280.5 < round_Location <= 2204.5 and
round_ASAD <= -0.19499999284744263 and
round_dG > -0.2850000262260437),
0.0012151662524716440143 * (0.12999999523162842 < round_Et <=
0.2150000035762787 and
round_Pc <= 0.48500001430511475 and
round_Pf_s > -0.9049999713897705),
0.0030058209141644370507 * (round_Location <= 1992.5 and
round_Pb <= 0.4950000047683716 and
0.20499999821186066 < round_Hgm <= 0.39499998092651367),
-0.042273652150263026084 * (round_Ht <= 0.7200000286102295 and
round_El > -0.07500000298023224 and
round_Pb <= 0.08500000089406967 and
round_dG <= 0.7749999761581421),
0.027673013766360339549 * (round_Pt <= -0.009999999776482582 and
round_F <= -0.5699999928474426),
0.027891051105625650625 * (round_El > 0.4599999785423279 and
round_Pt <= -0.14000000059604645),
0.0086383011037628777695 * (round_Location <= 2007.5 and

```
                            round_El <= -0.07500000298023224 and
                            round_dGh <= 0.5849999785423279),
        -0.009735134786261629 9106 * (2007.5 < round_Location <= 2198.5)    ])

def get_type_conversion():
    return {}
INDICATOR_COLS = []

IMPUTE_VALUES = {
    u'ASAD': 0.030000,
    u'El': 0.040000,
    u'Et': 0.040000,
    u'F': -0.060000,
    u'GhD': 0.010000,
    u'Hgm': 0.060000,
    u'Hnc': 0.000000,
    u'Ht': 0.000000,
    u'Location': 623.500000,
    u'Mu': 0.050000,
    u'Ns': 0.030000,
    u'Pb': 0.110000,
    u'Pc': -0.020000,
    u'Pf_s': 0.000000,
    u'Pt': 0.000000,
    u'Ra': 0.020000,
    u'Rf': 0.020000,
    u'aC': -0.030000,
    u'dG': 0.000000,
    u'dGh': 0.020000,
    u'dH': 0.010000,
    u'pK\'': 0.000000,}


def bag_of_words(text):
    """ set of whole words  in a block of text """
    if type(text) == float:
        return set()

    return set(word.lower() for word in
            re.findall(r'\w+', text, re.UNICODE | re.IGNORECASE))


def parse_date(x, date_format):
    """ convert date strings to numeric values. """
    try:
        # float values no longer pass isinstance(x, np.float64)
        if isinstance(x, (np.float64, float)):
```

```
            x = long_type(x)
        if '%f' in date_format and date_format.startswith('v2'):
            temp = str(x)
            if re.search('[\+-][0-9]+$', temp):
                temp = re.sub('[\+-][0-9]+$', '', temp)

            date_format = date_format[2:]
            dt = datetime.strptime(temp, date_format)
            sec = calendar.timegm(dt.timetuple())
            return sec * 1000 + dt.microsecond // 1000
        elif '%M' in date_format:
            temp = str(x)
            if re.search('[\+-][0-9]+$', temp):
                temp = re.sub('[\+-][0-9]+$', '', temp)

            return calendar.timegm(datetime.strptime(temp, date_format).timetuple())
        else:
            return datetime.strptime(str(x), date_format).toordinal()
    except:
        return float('nan')


def parse_percentage(s):
    """ remove percent sign so percentage variables can be converted to numeric """
    if isinstance(s, float):
        return s
    if isinstance(s, int):
        return float(s)
    try:
        return float(s.replace('%', ''))
    except:
        return float('nan')

def parse_nonstandard_na(s):
    """ if a column contains numbers and a unique non-numeric,
        then the non-numeric is considered to be N/A
    """
    try:
        ret = float(s)
        if np.isinf(ret):
            return float('nan')
        return ret
    except:
        return float('nan')

def parse_length(s):
    """ convert feet and inches as string to inches as numeric """
```

```
    try:
        if '"' in s and "'" in s:
            sp = s.split("'")
            return float(sp[0]) * 12 + float(sp[1].replace('"', ''))
        else:
            if "'" in s:
                return float(s.replace("'", '')) * 12
            else:
                return float(s.replace('"', ''))
    except:
        return float('nan')

def parse_currency(s):
    """ strip currency characters and commas from currency columns """
    if not isinstance(s, text_type):
        return float('nan')
    s = re.sub(u'[\$\u20AC\u00A3\uFFE1\u00A5\uFFE5]|(EUR)', '', s)
    s = s.replace(',', '')
    try:
        return float(s)
    except:
        return float('nan')


def parse_currency_replace_cents_period(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(",", "")
        val = float(val)
    except ValueError:
        val = float('nan')
    return val


def parse_currency_replace_cents_comma(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
```

```
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(".", "")
        val = val.replace(",", ".")
        val = float(val)
    except ValueError:
        val = float('nan')
    return val


def parse_currency_replace_no_cents(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(",", "")
        val = val.replace(".", "")
        val = float(val)
    except ValueError:
        val = float('nan')
    return val

def parse_numeric_types(ds):
    """ convert strings with numeric types (date, currency, etc.)
        to actual numeric values """
    TYPE_CONVERSION = get_type_conversion()
    for col in ds.columns:
        if col in TYPE_CONVERSION:
            convert_func = TYPE_CONVERSION[col]['convert_func']
            convert_args = TYPE_CONVERSION[col]['convert_args']
            ds[col] = ds[col].apply(convert_func, args=convert_args)
    return ds

def sanitize_name(name):
    safe = name.strip().replace("-", "_").replace("$", "_").replace(".", "_")
    safe = safe.replace("{", "_").replace("}", "_")
    safe = safe.replace('"', '_')
```

```
    safe = safe.replace("\n", "_")
    safe = safe.replace("\r", "_")
    return safe

def rename_columns(ds):
    new_names = {}
    existing_names = set()
    blank_index = 0
    for old_col in ds.columns:
        col = sanitize_name(old_col)
        if col == '':
            col = 'Unnamed: %d' % blank_index
            blank_index += 1
        if col in existing_names:
            raise ValueError('Duplication detected. Column with name=['
                    + old_col + '] was preprocessed to['
                    + col + '] that already exists')
        existing_names.add(col)
        new_names[old_col] = col
    ds.rename(columns=new_names, inplace=True)
    return ds

def add_missing_indicators(ds):
    for col in INDICATOR_COLS:
        ds[col + '-mi'] = ds[col].isnull().astype(int)
    return ds

def impute_values(ds):
    for col in ds:
        if col in IMPUTE_VALUES:
            ds.loc[ds[col].isnull(), col] = IMPUTE_VALUES[col]
    return ds

BIG_LEVELS = {
}


SMALL_NULLS = {
}


VAR_TYPES = {
    u'ASAD': 'N',
    u'El': 'N',
    u'Et': 'N',
    u'F': 'N',
    u'GhD': 'N',
```

```
    u'Hgm': 'N',
    u'Hnc': 'N',
    u'Ht': 'N',
    u'Location': 'N',
    u'Mu': 'N',
    u'Ns': 'N',
    u'Pb': 'N',
    u'Pc': 'N',
    u'Pf_s': 'N',
    u'Pt': 'N',
    u'Ra': 'N',
    u'Rf': 'N',
    u'aC': 'N',
    u'dG': 'N',
    u'dGh': 'N',
    u'dH': 'N',
    u'pK\'': 'N',
}


def combine_small_levels(ds):
    for col in ds:
        if BIG_LEVELS.get(col, None) is not None:
            mask = np.logical_and(~ds[col].isin(BIG_LEVELS[col]), ds[col].notnull())
            if np.any(mask):
                ds.loc[mask, col] = 'small_count'
        if SMALL_NULLS.get(col):
            mask = ds[col].isnull()
            if np.any(mask):
                ds.loc[mask, col] = 'small_count'
        if VAR_TYPES.get(col) == 'C' or VAR_TYPES.get(col) == 'T':
            mask = ds[col].isnull()
            if np.any(mask):
                if ds[col].dtype == float:
                    ds[col] = ds[col].astype(object)
                ds.loc[mask, col] = 'nan'
    return ds

# N/A strings in addition to the ones used by Pandas read_csv()
NA_VALUES = ['null', 'na', 'n/a', '#N/A', 'N/A', '?', '.', '', 'Inf', 'INF', 'inf', '-inf', '-Inf', '-INF', ' ',
'None', 'NaN', '-nan', 'NULL', 'NA', '-1.#IND', '1.#IND', '-1.#QNAN', '1.#QNAN', '#NA', '#N/A
N/A', '-NaN', 'nan']

# True/False strings in addition to the ones used by Pandas read_csv()
TRUE_VALUES = ['TRUE', 'True', 'true']
FALSE_VALUES = ['FALSE', 'False', 'false']
```

```
DEFAULT_ENCODING = 'utf8'

REQUIRED_COLUMNS =
[u"ASAD",u"El",u"Et",u"F",u"GhD",u"Hgm",u"Hnc",u"Ht",u"Location",u"Mu",u"Ns",u"Pb",u"
Pc",u"Pf_s",u"Pt",u"Ra",u"Rf",u"aC",u"dG",u"dGh",u"dH",u"pK'"]


def validate_columns(column_list):
   if set(REQUIRED_COLUMNS) <= set(column_list):
      return True
   else :
      raise ValueError("Required columns missing: %s" %
                  (set(REQUIRED_COLUMNS) - set(column_list)))

def convert_bool(ds):
   TYPE_CONVERSION = get_type_conversion()
   for col in ds.columns:
      if VAR_TYPES.get(col) == 'C' and ds[col].dtype in (int, float):
         mask = ds[col].notnull()
         ds[col] = ds[col].astype(object)
         ds.loc[mask, col] = ds.loc[mask, col].astype(text_type)
      elif VAR_TYPES.get(col) == 'N' and ds[col].dtype == bool:
         ds[col] = ds[col].astype(float)
      elif ds[col].dtype == bool:
         ds[col] = ds[col].astype(text_type)
      elif ds[col].dtype == object:
         if VAR_TYPES.get(col) == 'N' and col not in TYPE_CONVERSION:
            mask = ds[col].apply(lambda x: x in TRUE_VALUES)
            if np.any(mask):
               ds.loc[mask, col] = 1
            mask = ds[col].apply(lambda x: x in FALSE_VALUES)
            if np.any(mask):
               ds.loc[mask, col] = 0
            ds[col] = ds[col].astype(float)
         elif TYPE_CONVERSION.get(col) is None:
            mask = ds[col].notnull()
            ds.loc[mask, col] = ds.loc[mask, col].astype(text_type)
   return ds

def get_dtypes():
   return {a: object for a, b in VAR_TYPES.items() if b == 'C'}

def predict_dataframe(ds):
   return ds.apply(predict, axis=1)

def run_dataframe(ds):
   ds = rename_columns(ds)
```

```
    ds = convert_bool(ds)
    validate_columns(ds.columns)
    ds = parse_numeric_types(ds)
    ds = add_missing_indicators(ds)
    ds = impute_values(ds)
    ds = combine_small_levels(ds)
    prediction = 1/(1 + np.exp(-predict_dataframe(ds)))
    return prediction


def run(dataset_path, output_path, encoding=None):
    if encoding is None:
        encoding = DEFAULT_ENCODING

    ds = pd.read_csv(dataset_path, na_values=NA_VALUES, low_memory=False,
                     dtype=get_dtypes(), encoding=encoding)

    prediction = run_dataframe(ds)
    prediction_file = output_path
    prediction.name = 'Prediction'
    prediction.to_csv(prediction_file, header=True, index_label='Index')


def _construct_parser():
    import argparse

    parser = argparse.ArgumentParser(description='Make offline predictions with DataRobot
Prime')

    parser.add_argument(
        '--encoding',
        type=str,
        help=('the encoding of the dataset you are going to make predictions with. '
              'DataRobot Prime defaults to UTF-8 if not otherwise specified. See the '
              '"Codecs" column of the Python-supported standards chart '
              '(https://docs.python.org/2/library/codecs.html#standard-encodings) '
              'for possible alternative entries.'),
        metavar='<encoding>'
    )
    parser.add_argument(
        'input_path',
        type=str,
        help=('a .csv file (your dataset); columns must correspond to the '
              'feature set used to generate the DataRobot Prime model.'),
        metavar='<data_file>'
    )
    parser.add_argument(
```

```
        'output_path',
        type=str,
        help='the filename where DataRobot writes the results.',
        metavar='<output_file>'
    )

    return parser


def _parse_command(args):
    parser = _construct_parser()
    parsed_args = parser.parse_args(args[1:])

    if parsed_args.encoding is None:
        sys.stderr.write('Warning: For input data encodings other than UTF-8, '
                         'search "Prime examples" in the DataRobot Users Guide at
https://app.eu.datarobot.com/docs/users-guide/index.html')
        parsed_args.encoding = DEFAULT_ENCODING

    return parsed_args


if __name__ == '__main__':
    args = _parse_command(sys.argv)
    run(args.input_path, args.output_path, encoding=args.encoding)
```

Appendix C

The code (in Python programming language) could be run as a stand alone code to execute predictions based on the machine-learning model we built to predict Cancer-associated genes using on protein-protein interaction networks, essentiality scores and evolutionary properties:

```
import calendar
from datetime import datetime
from collections import namedtuple
import re
import sys
import time
import os

import numpy as np
import pandas as pd

PY3 = sys.version_info[0] == 3
```

```
if PY3:
   string_types = str,
   text_type = str
   long_type = int
else:
   string_types = basestring,
   text_type = unicode
   long_type = long

def predict(row):
   Group = row[u'Group']
   round_Average_Transcript_length = np.float32(row[u'Average Transcript length'])
   round_Blomen_KBM7 = np.float32(row[u'Blomen KBM7'])
   round_Blomen_KBM7_mi = np.float32(row[u'Blomen KBM7-mi'])
   round_Closeness = np.float32(row[u'Closeness'])
   round_Degree = np.float32(row[u'Degree'])
   round_Degree_mi = np.float32(row[u'Degree-mi'])
   round_End = np.float32(row[u'End'])
   round_Exon_Count = np.float32(row[u'Exon Count'])
   round_Gene_Length_bp = np.float32(row[u'Gene Length bp'])
   round_LofTool = np.float32(row[u'LofTool'])
   round_LofTool_mi = np.float32(row[u'LofTool-mi'])
   round_Phi = np.float32(row[u'Phi'])
   round_Phi_mi = np.float32(row[u'Phi-mi'])
   round_StdDev_Transcript_length = np.float32(row[u'StdDev Transcript length'])
   round_Tajima__s_D_regulatory = np.float32(row[u'Tajima\'s D regulatory'])
   round_Tajima__s_D_regulatory_mi = np.float32(row[u'Tajima\'s D regulatory-mi'])
   round_Transcript_count = np.float32(row[u'Transcript count'])
   round_dN_dS_Chimp = np.float32(row[u'dN/dS Chimp'])
   round_dN_dS_Chimp_mi = np.float32(row[u'dN/dS Chimp-mi'])
   round_missense_Z = np.float32(row[u'missense_Z'])
   round_missense_Z_mi = np.float32(row[u'missense_Z-mi'])
   round_s_het = np.float32(row[u's_het'])
   round_s_het_mi = np.float32(row[u's_het-mi'])
   return sum([
      -2.6863578,
        0.018202720175283508552 * (not Group == u'CM' and
                        not Group == u'MNC' and
                        round_Tajima__s_D_regulatory <= 0.4104999899864197 and
                        round_StdDev_Transcript_length > 1954.456787109375),
       -0.0054378635982185617379 * (round_dN_dS_Chimp_mi <= 0.5 and
                        round_Blomen_KBM7 <= -0.2516007423400879 and
                        round_LofTool <= 0.6634999513626099 and
                        round_LofTool_mi <= 0.5),
      -1.5583491431925192728E-11 * (round_End),
        0.048499210437713213828 * (round_Closeness <= 0.33500000834465027 and
                        3.2976694107055664 < round_missense_Z <= 4.050085067749023),
```

```
        0.09697593112305330 4983 * (round_Transcript_count > 20.5),
       -0.01738394816876158 7799 * (round_Degree <= 3.5 and
                    round_s_het > 0.023678744211792946),
      -0.008118824808685710 5895 * (not Group == u'CM' and
                    round_Tajima__s_D_regulatory <= 0.17550000548362732 and
                    round_Transcript_count <= 10.5 and
                    round_Gene_Length_bp > 2814.5),
       -0.023776888724096147121 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
                    round_missense_Z <= 3.2976694107055664 and
                    round_missense_Z_mi <= 0.5 and
                    round_LofTool <= 0.6634999513626099),
        0.015303739394473513113 * (round_LofTool > 0.6634999513626099 and
                    round_Gene_Length_bp > 104704.0 and
                    round_StdDev_Transcript_length > 2574.27734375 and
                    round_Exon_Count <= 156.5),
        0.032356712896822563408 * (round_Degree_mi <= 0.5 and
                    0.874500036239624 < round_LofTool <= 0.9921150207519531),
       -0.061064130313854887711 * (not Group == u'MNC' and
                    round_dN_dS_Chimp_mi > 0.5 and
                    round_missense_Z <= 3.2976694107055664 and
                    round_StdDev_Transcript_length <= 987.5745849609375),
        0.031074456293337455320 * (round_Phi > 0.00015248148702085018 and
                    round_Blomen_KBM7 > -0.5508977174758911 and
                    round_missense_Z <= 4.0604472160339355 and
                    round_Exon_Count > 87.5),
        0.027000278790342342738 * (round_Degree <= 12.5 and
                    round_s_het > 0.017613736912608147 and
                    round_Gene_Length_bp > 38001.5 and
                    round_Exon_Count <= 221.5),
      -0.007486176152675803 1915 * (round_Tajima__s_D_regulatory <=
0.4165000021457672 and
                    round_LofTool > 0.6634999513626099 and
                    round_Transcript_count <= 21.5 and
                    round_StdDev_Transcript_length <= 2558.05078125),
        0.033691518811985399218 * (round_Blomen_KBM7 <= -0.14686328172683716 and
                    round_LofTool > 0.962399959564209 and
                    round_StdDev_Transcript_length <= 2391.52490234375),
       -0.016812786172447410221 * (round_Degree_mi > 0.5 and
                    round_Phi_mi <= 0.5 and
                    round_Blomen_KBM7 > -0.27386000752449036 and
                    round_Blomen_KBM7_mi <= 0.5),
       -0.058684241640450440 78 * (round_Degree > 29.5 and
                    round_StdDev_Transcript_length <= 897.559326171875),
        0.014660818041939896808 * (round_Degree > 17.5 and
                    round_Degree_mi <= 0.5 and
                    round_LofTool > 0.9311000108718872),
```

0.0031378779289614714028 * (round_Degree > 5.5 and
round_Tajima__s_D_regulatory <= 0.44699999690055847 and
round_Blomen_KBM7 <= -0.2522552013397217 and
round_Average_Transcript_length > 2566.631103515625),
0.049576634648417938767 * (round_Closeness > 0.3149999976158142 and
round_Blomen_KBM7 > -0.5158457159996033 and
round_Gene_Length_bp <= 53140.5),
0.033125285580605914881 * (round_Degree > 56.5 and
round_Closeness > 0.2549999952316284 and
round_Gene_Length_bp > 41216.0),
0.015426516511084873567 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > 0.4235000014305115 and
round_s_het <= 0.014262920245528221),
-0.054664675004081307585 * (round_Closeness > 0.3149999976158142 and
round_s_het <= 0.025150161236524582),
0.064971575636202566484 * (round_Degree > 12.5 and
round_Blomen_KBM7 <= -0.17924460768699646 and
round_Gene_Length_bp > 38001.5),
0.008076281326937726282 * (round_End <= 100419720.0 and
round_Tajima__s_D_regulatory > -1.2885000705718994 and
round_Blomen_KBM7 <= -0.18547898530960083 and
round_missense_Z <= 3.2976694107055664),
-0.0025333307179157250021 * (not Group == u'CM' and
round_Tajima__s_D_regulatory <= -1.2894999980926514 and
round_Blomen_KBM7 <= -0.5147985219955444 and
round_Transcript_count <= 26.5),
0.020805097212741045093 * (not Group == u'NDNE' and
round_Degree_mi <= 0.5 and
0.659500002861023 < round_LofTool <= 0.9907699823379517),
0.0013239217591041027491 * (round_Closeness > 0.3149999976158142 and
round_Phi <= 0.0015477617271244526),
0.19073466683054091098 * (round_Closeness > 0.3149999976158142 and
round_Blomen_KBM7 > -0.5528146028518677 and
round_Exon_Count <= 224.5),
0.010314188977423621382 * (round_End > 127204736.0 and
round_Tajima__s_D_regulatory <= 0.4599999785423279 and
round_Transcript_count > 10.5 and
round_Gene_Length_bp > 2814.5),
-0.069447321340016812674 * (round_Degree_mi > 0.5),
0.016914894808412236221 * (Group == u'MNC' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5146373510360718),
0.027726420391460625953 * (round_Degree <= 3.5 and
round_StdDev_Transcript_length <= 1049.45654296875),
-0.0021865984701677676667 * (not Group == u'MNC' and
round_Degree <= 62.5 and
round_Degree_mi <= 0.5 and

round_Blomen_KBM7 <= -0.2516775131225586),
0.0053561320466416163441 * (Group == u'MNC' and
round_dN_dS_Chimp_mi <= 0.5 and
round_s_het > 0.028055116534233093 and
round_StdDev_Transcript_length > 168.6014404296875),
0.11425430800029053036 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Closeness > 0.3149999976158142 and
round_Gene_Length_bp > 41501.0),
-0.0032500288883767361817 * (round_Tajima__s_D_regulatory <=
1.4165000915527344 and
round_Blomen_KBM7 > -0.5153244733810425 and
round_StdDev_Transcript_length <= 166.56736755371094),
0.028508029551613675578 * (not Group == u'MNC' and
round_missense_Z <= 2.642360210418701 and
round_s_het <= 0.014316117390990257),
-0.0088550958230132672394 * (round_Closeness <= 0.33500000834465027 and
round_dN_dS_Chimp_mi > 0.5 and
round_Tajima__s_D_regulatory <= 0.4104999899864197 and
round_Exon_Count <= 97.5),
0.051816337733702685919 * (not Group == u'NDNE' and
round_Degree > 12.5 and
round_StdDev_Transcript_length > 636.7213134765625 and
round_Exon_Count <= 253.5),
-0.00060243449056097606292 * (round_Tajima__s_D_regulatory),
0.068946378006664615912 * (round_Transcript_count > 26.5),
-0.059757348541248728191 * (round_Degree_mi <= 0.5 and
round_Closeness > 0.3050000071525574 and
round_missense_Z_mi <= 0.5 and
round_Gene_Length_bp > 2814.5),
0.005517493171136504323 5 * (round_Degree > 3.5 and
round_Closeness > 0.3149999976158142 and
round_StdDev_Transcript_length > 626.7237548828125),
0.14400903577366730435 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Closeness <= 0.3149999976158142 and
round_Gene_Length_bp > 41501.0),
0.004987097945538400412 * (round_Blomen_KBM7 > -0.5154723525047302 and
round_missense_Z <= 2.6123709678649902 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Exon_Count <= 190.5),
-0.010315157710437320229 * (round_Degree <= 18.5 and
0.659500002861023 < round_LofTool <= 0.9602000117301941),
0.0073511514944097225768 * (round_Degree <= 35.5 and
round_Degree_mi <= 0.5 and
round_LofTool <= 0.874500036239624 and
round_StdDev_Transcript_length > 714.5440063476562),

-0.010425093114536482589 * (round_End > 88799552.0 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Phi > 0.9999969005584717 and
round_StdDev_Transcript_length > 2430.197265625),
0.018201114582529324265 * (not Group == u'CM' and
Group == u'MNC' and
round_Blomen_KBM7 > -0.5147985219955444 and
round_Gene_Length_bp > 37999.5),
0.0099794866530598415333 * (not Group == u'CM' and
not Group == u'MNC' and
round_Tajima__s_D_regulatory <= 0.4104999899864197 and
round_StdDev_Transcript_length <= 1954.456787109375),
-0.020922796241062330269 * (round_Degree_mi > 0.5 and
round_Average_Transcript_length <= 2606.02392578125),
0.10794282229118463967 * (round_missense_Z <= 3.2976694107055664 and
round_Average_Transcript_length > 2056.5712890625 and
round_Exon_Count > 240.5),
0.023013888862715946998 * (not Group == u'CM' and
round_Closeness > 0.3149999976158142 and
round_Blomen_KBM7 <= -0.5403047800064087 and
round_Transcript_count > 10.5),
-0.18056075313492864209 * (not Group == u'NDNE' and
round_Degree <= 3.5),
-0.0039957720980683267623 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_LofTool <= 0.874500036239624 and
round_Exon_Count <= 171.5),
-0.56961177766960269242 * (Group == u'NDNE' and
round_Degree <= 3.5 and
round_StdDev_Transcript_length <= 714.263427734375),
-0.010926747552256092094 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.32499998807907104 and
round_LofTool <= 0.9435499906539917 and
round_Gene_Length_bp > 54517.0),
0.0046976857155946452269 * (not Group == u'NDNE' and
round_Degree > 12.5 and
round_Gene_Length_bp <= 53079.0),
0.021939335798790569193 * (round_Phi),
-0.022286463855471616569 * (round_Tajima__s_D_regulatory >
0.4235000014305115 and
round_Blomen_KBM7 > -0.21081802248954773 and
round_s_het > 0.015080630779266357 and
round_StdDev_Transcript_length > 580.992431640625),
0.0027537702678576865545 * (round_Tajima__s_D_regulatory <=
0.4104999899864197 and
round_Phi > 0.0002437000221107155 and
round_missense_Z <= 3.308867931365967),

-0.0027288874703135009708 * (round_End > 127204736.0 and
round_Tajima__s_D_regulatory <= 0.4599999785423279 and
round_Transcript_count <= 10.5 and
round_Gene_Length_bp > 2814.5),
0.08109139290614296447 * (round_missense_Z_mi),
0.014012303940575399075 * (round_missense_Z <= 3.2976694107055664 and
round_LofTool > 0.9603500366210938),
0.052541363893365403137 * (not Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_Closeness > 0.3149999976158142 and
round_Average_Transcript_length > 2570.535888671875),
0.1551699454009141943 * (round_Tajima__s_D_regulatory <=
0.4235000014305115 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_s_het > 0.015603477135300636 and
round_Transcript_count <= 21.5),
-0.0080792105837217097208 * (round_Tajima__s_D_regulatory >
0.4104999899864197 and
round_Average_Transcript_length <= 1876.067626953125),
0.021504577862045057279 * (round_End > 125209176.0 and
round_Degree > 12.5 and
round_dN_dS_Chimp_mi <= 0.5),
-0.015295315530416180028 * (round_Degree <= 56.5 and
round_Closeness > 0.2549999952316284 and
round_s_het <= 0.031048648059368134 and
round_Gene_Length_bp > 41216.0),
-0.0078502999183946500783 * (round_Degree <= 3.5 and
round_s_het > 0.02518850564956665 and
round_Transcript_count <= 15.5),
0.024239284046736895434 * (not Group == u'MNC' and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z <= 3.2976694107055664 and
round_Average_Transcript_length <= 2056.5712890625),
0.03864183015657941811 * (round_Degree > 4.5 and
round_Blomen_KBM7 <= -0.15796872973442078 and
round_StdDev_Transcript_length > 1227.5733642578125),
0.012003822160604299754 * (not Group == u'MNC' and
round_missense_Z > 3.2976694107055664),
0.01265667266342007137 * (3.5 < round_Degree <= 70.0 and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 156.5),
-0.025979339010908503865 * (Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Phi > 0.0014898625668138266 and
round_StdDev_Transcript_length > 636.7213134765625),
0.088736011639851314348 * (not Group == u'CM' and
round_End > 110010712.0 and

round_Degree_mi <= 0.5 and
round_missense_Z <= 3.996763229370117),
-0.042684495370579833562 * (round_Tajima__s_D_regulatory <=
1.4184999465942383 and
round_Transcript_count <= 2.5 and
round_Gene_Length_bp > 9942.0 and
round_Average_Transcript_length > 2037.56787109375),
0.010718660893401628365 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 <= -0.5181520581245422 and
round_Transcript_count <= 26.5 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.0044768437453202277257 * (round_Closeness > 0.2549999952316284 and
round_Blomen_KBM7_mi <= 0.5 and
round_s_het_mi <= 0.5 and
round_StdDev_Transcript_length <= 2625.484375),
0.012631834697413829582 * (2814.5 < round_Gene_Length_bp <= 9338.5 and
round_StdDev_Transcript_length > 580.992431640625),
0.026831902960513024509 * (round_Closeness > 0.33500000834465027 and
3.2976694107055664 < round_missense_Z <= 4.050085067749023),
0.003391817256333500 5285 * (round_Blomen_KBM7 > -0.5146373510360718 and
round_Transcript_count > 10.5 and
round_Average_Transcript_length > 2037.5650634765625),
0.0070956564876167357164 * (round_End <= 124517040.0 and
round_Tajima__s_D_regulatory > 0.5564999580383301 and
round_Blomen_KBM7 <= -0.2610846161842346 and
round_Gene_Length_bp > 2814.5),
0.05358215945498377708 * (round_End <= 124883040.0 and
-1.2874999046325684 < round_Tajima__s_D_regulatory <=
1.4165000915527344 and
round_Transcript_count <= 20.5),
0.010564605410865832158 * (round_Degree > 45.0 and
round_dN_dS_Chimp_mi > 0.5 and
round_Transcript_count <= 26.5 and
round_Gene_Length_bp > 2402.0),
0.064211045541165565065 * (round_Degree <= 4.5 and
round_Degree_mi <= 0.5 and
round_StdDev_Transcript_length > 1301.4395751953125),
-0.099504943567778156299 * (Group == u'NDNE' and
round_Degree > 3.5 and
round_missense_Z <= 2.3079466819763184),
0.16847444975170958181 * (round_Closeness <= 0.3149999976158142 and
2.642360210418701 < round_missense_Z <= 4.0604472160339355),
0.04243059279242206161 * (round_Closeness <= 0.3050000071525574 and
round_Blomen_KBM7 > -0.5158457159996033 and
round_Gene_Length_bp <= 53140.5),
0.047538526872885802921 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and

round_Phi > 0.8754478693008423),
0.01377767556161369443 * (round_Tajima__s_D_regulatory <=
1.4184999465942383 and
round_Transcript_count > 2.5 and
round_Gene_Length_bp > 9942.0 and
round_Average_Transcript_length > 2037.56787109375),
-0.01544246978038381346 * (Group == u'NDNE' and
round_Degree > 3.5),
-0.031626465358424796226 * (round_Degree_mi > 0.5 and
round_Phi <= 0.1322648823261261),
0.023920451949571132355 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_LofTool > 0.9829000234603882),
0.026861723913622247845 * (round_Closeness > 0.3149999976158142 and
round_Tajima__s_D_regulatory > 0.4104999899864197),
0.0036581146817508628649 * (round_End > 100419720.0 and
round_Tajima__s_D_regulatory <= -0.234499990940094 and
round_missense_Z <= 2.6417436599731445),
-0.037779318023490326972 * (round_Closeness > 0.3149999976158142 and
round_Phi <= 3.313508932478726e-05 and
round_LofTool > 0.04450000077486038),
0.021565100631107423507 * (round_missense_Z <= 3.2976694107055664 and
2821.0 < round_Gene_Length_bp <= 9941.5 and
round_Average_Transcript_length <= 2071.857421875),
-0.018406415026041244437 * (round_Closeness <= 0.3149999976158142 and
round_Phi <= 0.0015477617271244526),
0.044321413351531856184 * (round_End > 100419720.0 and
round_Tajima__s_D_regulatory <= 0.5145000219345093 and
round_missense_Z <= 3.996763229370117 and
round_StdDev_Transcript_length > 166.57864379882812),
-0.040937321521144612313 * (0.3050000071525574 < round_Closeness <=
0.3149999976158142 and
round_Tajima__s_D_regulatory > 0.4104999899864197),
-0.01205681028131431326 * (round_Degree > 4.5 and
round_Gene_Length_bp <= 39298.0),
0.01376055986142680175 * (round_dN_dS_Chimp <= 0.02499999850988388 and
round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool <= 0.9922449588775635 and
round_Average_Transcript_length > 2037.5650634765625),
0.0012393689458476812339 * (round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z <= 2.571293354034424 and
168.6014404296875 < round_StdDev_Transcript_length <=
1450.0491943359375),
-0.0077058749532958803127 * (round_Tajima__s_D_regulatory <=
1.0544999837875366 and
round_Blomen_KBM7 > -0.5153281092643738 and
round_missense_Z_mi <= 0.5 and

round_Transcript_count <= 20.5),
0.035821139812792640589 * (round_Tajima__s_D_regulatory <= -1.2855000495910645 and
round_LofTool > 0.8144999742507935 and
round_StdDev_Transcript_length > 3607.19140625),
0.011287896175946629182 * (Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Phi > 0.8754478693008423),
-0.086681950407499250288 * (not Group == u'CM' and
not Group == u'MNC' and
round_Degree <= 60.5 and
round_missense_Z > 4.031624794006348),
-0.055560469246041271907 * (round_Phi <= 0.12447576969861984 and
round_StdDev_Transcript_length <= 1098.5848388671875),
0.056832869755874752815 * (round_End <= 100419720.0 and
round_Blomen_KBM7 <= -0.11097116768360138 and
round_missense_Z > 3.2974047660827637),
0.0042022538063266699077 * (round_Closeness <= 0.7749999761581421 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Tajima__s_D_regulatory <= 0.34049999713897705 and
round_Blomen_KBM7 > -0.5153244733810425),
0.010232660118857737214 * (not Group == u'NDNE' and
round_End <= 100419720.0 and
round_Phi <= 0.998741626739502 and
round_s_het > 0.017704255878925323),
-0.014904888524450881845 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7 > -0.514461874961853 and
round_StdDev_Transcript_length <= 1944.22802734375),
-0.12345581397405855362 * (round_Closeness <= 0.3149999976158142 and
round_Phi <= 3.313508932478726e-05 and
round_LofTool > 0.04450000077486038),
-0.0082097542006331434422 * (Group == u'NDNE' and
round_End <= 100419720.0 and
round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 <= -0.25102293491363525),
0.006659739649994870845 * (round_End <= 100419720.0 and
round_Degree <= 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Exon_Count <= 221.5),
-0.026228336463867144013 * (not Group == u'CM' and
round_End > 127204736.0 and
round_Phi > 0.9981463551521301 and
round_s_het > 0.1433388888835907),
0.007283276126256562482 7 * (round_End > 100419720.0 and
round_missense_Z > 2.6417436599731445),
-0.018198365192935522794 * (4.5 < round_Degree <= 31.5 and

round_Phi <= 0.919446587562561 and
round_Gene_Length_bp > 39298.0),
-0.001446911779157069536 5 * (0.659500002861023 < round_LofTool <=
0.9921150207519531 and
round_StdDev_Transcript_length > 589.7366333007812 and
round_Exon_Count <= 174.5),
0.04649175361763767389 * (round_LofTool > 0.6634999513626099 and
round_Exon_Count > 156.5),
-0.040594818357408524179 * (Group == u'NDNE' and
round_s_het > 0.016313210129737854 and
round_Gene_Length_bp <= 38001.5),
-0.0677060657236692314 82 * (round_dN_dS_Chimp_mi > 0.5 and
round_Blomen_KBM7 > -0.5146373510360718 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 2388.5126953125),
0.019417071016638930842 * (round_End <= 120516032.0 and
round_Closeness <= 0.33500000834465027 and
round_Tajima__s_D_regulatory <= 0.4104999899864197 and
round_Exon_Count > 97.5),
0.022064341778839532265 * (round_Transcript_count > 20.5 and
round_StdDev_Transcript_length <= 2418.21435546875),
0.0044133194936464073543 * (not Group == u'NDNE' and
round_Tajima__s_D_regulatory <= 0.5570000410079956 and
round_LofTool > 0.8105000257492065),
-0.14398676711490873692 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory <= -0.4115000069141388 and
8050.5 < round_Gene_Length_bp <= 53048.5),
0.040024524551637907788 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Transcript_count > 15.5),
-0.015835246306718037124 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7 <= -0.514461874961853 and
round_StdDev_Transcript_length <= 1944.22802734375),
0.010716122739572278219 * (round_Degree > 60.5 and
round_Blomen_KBM7 > -0.5153281092643738),
-0.003779888350437880448 2 * (round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z > 4.052361965179443 and
round_LofTool <= 0.9922449588775635 and
round_Exon_Count <= 108.5),
-0.001568140439094732347 5 * (round_dN_dS_Chimp_mi > 0.5 and
round_Average_Transcript_length <= 2037.5650634765625),
0.038785018347274782813 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_LofTool <= 0.6514999866485596),
-0.003997751487252501276 2 * (Group == u'NDNE' and
round_dN_dS_Chimp_mi <= 0.5 and

round_Transcript_count <= 26.5 and
round_StdDev_Transcript_length <= 2388.5126953125),
0.038112358760383956147 * (Group == u'NDNE' and
round_Degree_mi > 0.5),
0.026371943455055456978 * (not Group == u'MNC' and
round_Phi > 3.313508932478726e-05 and
3.2976694107055664 < round_missense_Z <= 4.052361965179443),
0.049444141026331135669 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > 0.4235000014305115 and
round_s_het > 0.014262920245528221),
0.013937161271230649046 * (not Group == u'NDNE' and
round_Closeness > 0.3149999976158142 and
round_missense_Z <= 1.43977952003479),
-0.020864630641113317278 * (round_Degree <= 12.5 and
round_missense_Z > 4.052361965179443 and
round_StdDev_Transcript_length <= 2616.1044921875),
0.010891575439352343263 * (not Group == u'CM' and
round_Phi > 3.318415838293731e-05 and
round_Blomen_KBM7 > -0.5403047800064087 and
round_Transcript_count > 10.5),
-0.020197315811600891067 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
round_Blomen_KBM7 > -0.12487166374921799 and
round_s_het > 0.02633928880095482 and
round_Exon_Count <= 53.0),
-0.025658439979292017169 * (round_Closeness <= 0.3149999976158142 and
round_s_het <= 0.025150161236524582 and
round_Average_Transcript_length <= 2057.064453125),
-0.042894291214758038799 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Gene_Length_bp > 9941.5 and
round_Average_Transcript_length <= 2037.5650634765625),
0.0114569281039716038 * (not Group == u'NDNE' and
round_Closeness <= 0.32499998807907104 and
round_Tajima__s_D_regulatory <= 0.5564999580383301 and
round_Gene_Length_bp <= 47758.5),
0.067766934758570276931 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_LofTool > 0.9787000417709351),
0.042775949189413867146 * (not Group == u'CM' and
round_End <= 127204736.0 and
round_Degree > 61.5),
-0.033595605379461213058 * (round_missense_Z > 4.052361965179443 and
round_LofTool <= 0.9807000160217285),
0.017127149696538990914 * (round_Gene_Length_bp > 37833.5 and
round_StdDev_Transcript_length <= 1956.23095703125),
0.011030940167582190675 * (round_Degree <= 12.5 and

round_Blomen_KBM7 <= -0.1940726488828659 and
round_s_het > 0.025150161236524582 and
round_Transcript_count <= 20.5),
-0.25855416246864998397 * (round_Phi <= 0.00015248148702085018 and
round_LofTool <= 0.6634999513626099 and
round_Exon_Count <= 157.5),
0.018976143304323109251 * (round_End > 100419720.0 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5072991847991943 and
round_LofTool <= 0.994350016117096),
0.0010472512424781869628 * (round_LofTool <= 0.6655000448226929 and
round_s_het > 0.01709270477294922 and
round_Exon_Count <= 87.5),
-0.018696452857061698211 * (not Group == u'CM' and
round_Degree <= 64.0 and
round_dN_dS_Chimp_mi > 0.5 and
round_StdDev_Transcript_length <= 1942.072509765625),
0.011962027921206701275 * (not Group == u'NDNE' and
12.5 < round_Degree <= 63.5 and
round_Degree_mi <= 0.5),
0.15322978417035446053 * (round_Degree <= 4.5 and
round_LofTool <= 0.9868500232696533 and
round_StdDev_Transcript_length > 1049.45654296875),
0.046437531520907876503 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_LofTool <= 0.9787000417709351 and
round_StdDev_Transcript_length > 1005.2410888671875),
-0.039805408645227642606 * (round_Degree <= 35.5 and
round_Blomen_KBM7 > -0.2521226406097412 and
round_s_het <= 0.01751864142715931 and
round_Gene_Length_bp > 2814.5),
-0.070420047677563976651 * (Group == u'NDNE' and
round_Degree > 3.5 and
round_missense_Z > 2.3079466819763184),
0.0025535846861198395648 * (Group == u'NDNE' and
round_Degree <= 35.5 and
round_Tajima__s_D_regulatory <= 1.0544999837875366 and
round_missense_Z <= 4.052361965179443),
0.092506701425168813557 * (round_Tajima__s_D_regulatory <=
0.4235000014305115 and
round_Blomen_KBM7 > -0.2521226406097412),
0.028016343518460877504 * (round_Degree > 4.5 and
round_missense_Z > 1.43977952003479 and
round_Transcript_count <= 21.5),
0.013589678385553336654 * (round_End <= 100419720.0 and
round_Degree > 12.5 and
round_LofTool <= 0.812000036239624 and

round_Exon_Count <= 221.5),
-0.010867009106563865761 * (round_Degree <= 60.5 and
round_Tajima__s_D_regulatory > -0.7795000076293945 and
round_Blomen_KBM7 > -0.5153281092643738 and
round_Exon_Count > 154.5),
-0.0036726609659116712936 * (round_Degree_mi <= 0.5 and
round_missense_Z <= 2.552614450454712 and
round_LofTool > 0.6644999980926514 and
round_Average_Transcript_length <= 2570.535888671875),
0.013832036762706318919 * (round_Closeness <= 0.7749999761581421 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Tajima__s_D_regulatory <= 0.34049999713897705 and
round_Blomen_KBM7 <= -0.5153244733810425),
0.053044796037296365609 * (round_Degree > 12.5 and
round_LofTool <= 0.9921150207519531 and
round_s_het > 0.025150161236524582 and
round_Exon_Count > 224.5),
-0.013287956853066012 * (not Group == u'MNC' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5146373510360718 and
round_Gene_Length_bp <= 37999.5),
0.10324470994342001273 * (round_LofTool > 0.9921150207519531),
0.005792180707143307855667 * (not Group == u'CM' and
round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 <= -0.5403047800064087 and
round_Transcript_count > 10.5),
0.092157718501607158168 * (round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.15531033277511597 and
round_missense_Z > 2.721888303756714 and
round_Exon_Count > 262.0),
-0.1305724926220721005 * (round_Degree_mi > 0.5 and
round_StdDev_Transcript_length <= 1060.87158203125),
0.016193134371026433188 * (round_Degree <= 35.5 and
round_Degree_mi <= 0.5 and
round_LofTool <= 0.874500036239624 and
round_StdDev_Transcript_length <= 714.5440063476562),
0.035774440858628839268 * (Group == u'CM' and
round_Tajima__s_D_regulatory > 0.4165000021457672),
-0.0069014727149561538172 * (round_missense_Z <= 4.054958820343018 and
round_Transcript_count <= 10.5 and
round_Gene_Length_bp > 10637.0 and
round_Average_Transcript_length > 2037.5650634765625),
-0.026771080458371804972 * (round_Closeness <= 0.2549999952316284 and
round_Phi > 0.8162848949432373 and
round_LofTool <= 0.9833999872207642),
-0.022253783562079281627 * (round_Closeness <= 0.26499998569488525 and
round_LofTool <= 0.8105000257492065 and

round_Gene_Length_bp <= 37966.5),
-0.002468892445794360068 * (not Group == u'MNC' and
-1.2874999046325684 < round_Tajima__s_D_regulatory <=
1.2934999465942383 and
round_Blomen_KBM7 > -0.4999815821647644),
0.042899309309499849060 26 * (not Group == u'NDNE' and
round_missense_Z > 1.43977952003479 and
round_StdDev_Transcript_length > 586.8510131835938 and
round_Exon_Count <= 229.5),
0.017509022546737609133 * (round_End > 100419720.0),
-0.066444044833209647827 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
round_Blomen_KBM7 <= -0.12487166374921799 and
3.134337902069092 < round_missense_Z <= 3.2974047660827637),
-0.00219063763450522489 * (not Group == u'NDNE' and
round_Degree <= 4.5 and
round_Phi > 0.8875079154968262),
0.11064953698387736125 * (Group == u'CM' and
round_End > 110010712.0 and
round_Degree_mi <= 0.5),
-0.059270604135849759564 * (Group == u'NDNE' and
round_missense_Z <= 2.7258143424987793 and
round_Exon_Count <= 171.5),
-0.018291629266993809227 * (round_Degree <= 12.5 and
round_Blomen_KBM7 > -0.12487166374921799),
0.035226494727457799416 * (round_Phi > 0.12447576969861984 and
93.5 < round_Exon_Count <= 265.5),
0.050722711654871383002 * (round_Degree > 12.5 and
round_Closeness > 0.33500000834465027 and
round_Tajima__s_D_regulatory <= 0.4235000014305115 and
round_LofTool <= 0.9922800064086914),
-0.051814515969269794859 * (round_Closeness <= 0.3149999976158142 and
round_Phi <= 0.12447576969861984 and
round_Exon_Count <= 83.5),
-0.004095557220962523122 * (round_Degree > 12.5 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z > 4.031624794006348 and
round_LofTool <= 0.9499499797821045),
-0.061167712425044540314 * (round_Degree_mi > 0.5 and
round_Gene_Length_bp <= 2510.5),
-0.0182849435306143282 * (Group == u'MNC' and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length <= 987.5745849609375),
-0.32389234107127440332 * (Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi <= 0.14362311363220215),
-0.14541641777476235764 * (Group == u'MNC' and

round_Gene_Length_bp <= 53048.5),
0.008732082275532983967 1 * (round_Degree <= 12.5 and
round_Degree_mi <= 0.5 and
round_missense_Z > 2.120723009109497 and
round_StdDev_Transcript_length > 1011.6593627929688),
0.017852506892799022836 * (round_Closeness > 0.2549999952316284 and
3.334120750427246 < round_missense_Z <= 4.052361965179443 and
round_s_het > 0.014458265155553818),
-0.038423746283818283054 * (round_Degree <= 3.5 and
round_Degree_mi > 0.5 and
round_s_het <= 0.023678744211792946),
-0.004412961621996456911 * (not Group == u'CM' and
round_Tajima__s_D_regulatory <= 0.9564999938011169 and
round_Blomen_KBM7 > -0.5147985219955444 and
round_Gene_Length_bp <= 37999.5),
0.050996412382406750008 * (not Group == u'MNC' and
round_missense_Z <= 2.642360210418701 and
round_s_het > 0.014316117390990257 and
round_Exon_Count <= 165.5),
0.017338694268998987996 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Closeness > 0.3149999976158142),
-0.023156005411269445921 * (not Group == u'CM' and
round_End <= 127204736.0 and
round_Degree <= 61.5 and
round_missense_Z <= 4.031624794006348),
0.002830861869401016511 1 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5146373510360718 and
round_LofTool <= 0.9922449588775635 and
round_Transcript_count <= 20.5),
-0.040546423472875826877 * (Group == u'NDNE' and
round_missense_Z <= 2.7258143424987793 and
round_Exon_Count > 171.5),
-0.015113303914765858355 * (not Group == u'CM' and
round_End > 127204736.0 and
round_Phi > 0.9974162578582764 and
round_StdDev_Transcript_length <= 1345.761474609375),
-0.10222460154756490835 * (round_Closeness <= 0.2549999952316284 and
round_Phi > 0.0002437000221107155 and
round_missense_Z <= 4.060447216033935 5 and
round_StdDev_Transcript_length > 615.416259765625),
0.012362969768704433135 * (not Group == u'MNC' and
round_Degree > 34.5 and
round_Degree_mi <= 0.5 and
round_missense_Z <= 4.0604472160339355),
-0.006450323123439150923 99 * (round_Tajima__s_D_regulatory <=
0.43549999594688416 and

round_missense_Z <= 3.2976694107055664 and
round_LofTool <= 0.9603500366210938 and
round_Exon_Count <= 164.5),
-0.0022205334822635804277 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > -1.2894999980926514 and
round_Blomen_KBM7 <= -0.5147985219955444 and
round_Transcript_count <= 26.5),
-0.25240843520512323828 * (round_Phi > 0.00015248148702085018 and
round_Blomen_KBM7 > -0.15492522716522217 and
round_LofTool <= 0.6634999513626099 and
round_Exon_Count <= 157.5),
0.009547672207145457140 6 * (not Group == u'NDNE' and
round_Closeness <= 0.32499998807907104 and
round_Tajima__s_D_regulatory <= 0.5564999580383301 and
round_Gene_Length_bp > 47758.5),
0.025785796597482577019 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_LofTool > 0.9787000417709351),
0.1378685779507847764 * (not Group == u'NDNE' and
round_Closeness > 0.3149999976158142 and
round_Phi > 0.8768634796142578),
0.0024299681294777109031 * (round_End <= 128392288.0 and
round_Gene_Length_bp > 9338.5 and
round_StdDev_Transcript_length > 1956.23095703125),
0.043814273445441413724 * (round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.14571721851825714 and
round_LofTool > 0.9435499906539917),
0.0033544123090482034534 * (round_missense_Z <= 2.6123709678649902 and
round_missense_Z_mi <= 0.5 and
round_Gene_Length_bp > 9941.5 and
round_StdDev_Transcript_length > 166.57949829101562),
0.027254350066886999515 * (round_Closeness > 0.33500000834465027 and
round_Tajima__s_D_regulatory <= 0.5564999580383301),
-0.0021195491504552806984 * (round_Degree <= 35.5 and
round_Gene_Length_bp > 9906.5 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Average_Transcript_length <= 1956.6083984375),
0.048251921050298546279 * (round_Degree > 12.5 and
round_Degree_mi <= 0.5 and
round_Closeness <= 0.3149999976158142 and
round_StdDev_Transcript_length > 1285.35791015625),
0.022409297933030165179 * (round_Degree > 3.5 and
round_Closeness <= 0.3149999976158142 and
round_LofTool <= 0.6775000095367432 and
round_Exon_Count <= 151.5),
0.045353884427582646932 * (round_Gene_Length_bp <= 9906.5 and
round_StdDev_Transcript_length > 166.57864379882812 and

round_Average_Transcript_length > 1577.136474609375),
-0.012720933738028348745 * (round_Phi <= 0.12447576969861984 and
round_StdDev_Transcript_length <= 1452.739013671875),
0.0020671485816952722345 * (round_Tajima__s_D_regulatory <=
1.0544999837875366 and
round_Phi <= 0.9990512132644653 and
round_Blomen_KBM7 > -0.5153281092643738 and
round_Transcript_count > 20.5),
-0.0072592442110520818271 * (round_Degree_mi > 0.5 and
round_Average_Transcript_length <= 2170.1923828125),
0.0016667596308927788533 * (3.308867931365967 < round_missense_Z <=
4.052361965179443 and
round_Exon_Count > 97.5),
0.01867224397834999633 * (round_Blomen_KBM7 <= -0.2516775131225586 and
3.308867931365967 < round_missense_Z <= 4.0613203048706055 and
round_s_het > 0.015855055302381516),
0.061536203645746870294 * (round_dN_dS_Chimp <= 0.5950000286102295 and
round_LofTool > 0.9922449588775635),
-0.01865222650192009321 * (not Group == u'CM' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7_mi > 0.5 and
round_StdDev_Transcript_length <= 2769.32568359375),
-0.031063061425573603586 * (round_dN_dS_Chimp_mi <= 0.5 and
round_StdDev_Transcript_length <= 168.6014404296875),
0.016728971448681104889 * (round_Degree > 10.5 and
round_Degree_mi <= 0.5 and
round_LofTool <= 0.9311000108718872 and
round_s_het > 0.025092266499996185),
-0.30582887886172710479 * (round_Degree <= 4.5 and
round_Phi <= 0.14253592491149902 and
round_StdDev_Transcript_length > 616.5709228515625),
0.017548642979559044702 * (round_End <= 100419720.0 and
round_missense_Z <= 4.031624794006348 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Exon_Count <= 221.5),
0.033533794559069317331 * (round_End <= 100419720.0 and
round_Degree > 12.5 and
round_LofTool > 0.812000036239624 and
round_Exon_Count <= 221.5),
0.0065947414569974410758 * (not Group == u'MNC' and
round_End > 100419720.0 and
round_Tajima__s_D_regulatory > -1.2855000495910645 and
round_StdDev_Transcript_length <= 635.1746826171875),
-0.0056566784579737171626 * (round_Degree > 4.5 and
round_Closeness <= 0.314999976158142 and
round_Transcript_count > 24.5 and
round_StdDev_Transcript_length > 897.559326171875),

0.081239321752520798903 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
3.2976694107055664 < round_missense_Z <= 3.8540451526641846),
0.030930760440212312634 * (round_Degree > 35.5 and
round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.1796257197856903 and
round_LofTool <= 0.874500036239624),
-0.0074955400533961748233 * (round_Closeness > 0.2549999952316284 and
round_Phi > 0.0015477617271244526 and
round_StdDev_Transcript_length <= 2427.245361328125 and
round_Exon_Count <= 96.5),
0.0060206295646320932488 * (round_dN_dS_Chimp_mi <= 0.5 and
round_s_het > 0.01754925772547722 and
round_Transcript_count <= 26.5 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.022479192865490577047 * (Group == u'NDNE' and
round_Degree <= 8.5 and
round_Phi <= 0.18718643486499786 and
round_StdDev_Transcript_length > 714.5440063476562),
-0.004902590801046578968 * (not Group == u'MNC' and
round_s_het <= 0.015855055302381516),
-0.017831132034095014544 * (Group == u'NDNE' and
round_End <= 100419720.0 and
round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 > -0.25102293491363525),
0.013623595360248288294 * (round_Closeness > 0.2549999952316284 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_StdDev_Transcript_length <= 2430.197265625 and
round_Exon_Count <= 246.0),
0.018865549838809763522 * (round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Average_Transcript_length <= 2071.857421875),
0.014732213904117162293 * (round_Degree > 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_s_het > 0.015838049352169037 and
round_StdDev_Transcript_length > 168.6014404296875),
-0.016714848080446102069 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > 0.9564999938011169 and
round_Blomen_KBM7 > -0.5147985219955444 and
round_Gene_Length_bp <= 37999.5),
-0.015963145907007737778 * (round_StdDev_Transcript_length <=
590.2589111328125),
-0.0060448945213366633844 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_s_het > 0.02767527475953102),
0.027704609106936072677 * (not Group == u'NDNE' and

round_missense_Z <= 1.43977952003479 and
round_StdDev_Transcript_length > 586.8510131835938 and
round_Exon_Count <= 229.5),
-0.051098898316382507234 * (round_Closeness > 0.2549999952316284 and
3.8498473167419434 < round_missense_Z <= 4.052361965179443),
-0.058386046266751971678 * (round_Blomen_KBM7 > -0.1469864696264267),
0.0060639684811802183756 * (round_Closeness <= 0.35500001907348633 and
round_dN_dS_Chimp <= 0.5849999785423279 and
round_missense_Z <= 4.052361965179443 and
round_Gene_Length_bp > 2814.5),
0.05233613286777938356 * (Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi > 0.1670895218849182),
-0.17848306160053159508 * (round_Degree <= 4.5 and
round_StdDev_Transcript_length <= 1049.45654296875),
-0.00011540571561985079577 * (round_Tajima__s_D_regulatory <= -
1.2855000495910645 and
round_Blomen_KBM7 > -0.5154076814651489 and
round_LofTool <= 0.8144999742507935),
-0.010416764352577573272 * (round_missense_Z <= 3.325303077697754 and
round_LofTool <= 0.9922800064086914 and
round_StdDev_Transcript_length > 2377.90625),
-0.05612691108153916586 * (Group == u'NDNE' and
round_missense_Z > 1.1351158618927002 and
round_Transcript_count <= 10.5),
0.090625893742273977427 * (round_Degree > 3.5 and
round_Closeness > 0.3149999976158142 and
round_Gene_Length_bp <= 41501.0),
0.019908066071534915448 * (round_LofTool > 0.9920099973678589 and
round_Exon_Count > 86.5),
-0.014119188513395540888 * (not Group == u'MNC' and
round_Phi <= 0.0015477617271244526),
0.0046295616520464280552 * (round_Degree > 3.5 and
round_Exon_Count > 156.5),
0.017556588763314336793 * (round_Degree <= 12.5 and
round_Blomen_KBM7 <= -0.1114160418510437 and
round_StdDev_Transcript_length <= 2377.90625 and
round_Average_Transcript_length > 2169.3193359375),
0.0187314523846858344 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_Phi > 0.00028779974672943354 and
round_StdDev_Transcript_length > 1207.362548828125),
0.021330433143297765353 * (round_Degree <= 64.5 and
round_Degree_mi <= 0.5 and
round_Closeness > 0.3149999976158142 and
round_Exon_Count <= 156.5),

0.021452094564182736663 * (round_Tajima__s_D_regulatory > -1.2874999046325684 and
round_Blomen_KBM7 <= -0.12487166374921799 and
round_missense_Z <= 3.134337902069092),
-0.032839560232847293808 * (round_missense_Z > 3.04426908493042 and
round_StdDev_Transcript_length <= 167.33212280273438),
0.012377636161393038711 * (round_End <= 100419720.0 and
round_Blomen_KBM7 <= -0.5154723525047302 and
round_missense_Z <= 3.2974047660827637),
0.00011644535307463677531 * (round_Degree),
0.011337114232236933375 * (round_Degree_mi <= 0.5 and
round_missense_Z > 1.43977952003479 and
round_LofTool <= 0.9886499643325806 and
round_Gene_Length_bp > 39341.0),
-0.012961434512462111784 * (not Group == u'NDNE' and
round_Degree <= 12.5 and
round_Degree_mi <= 0.5 and
round_missense_Z <= 2.3431456089019775),
-0.020474012020399504769 * (round_End <= 120572272.0 and
round_Closeness <= 0.32499998807907104 and
round_Tajima__s_D_regulatory <= 0.4235000014305115 and
round_Gene_Length_bp > 54517.0),
-0.010177296319135669886 * (round_Degree <= 10.5 and
round_Degree_mi <= 0.5 and
round_Phi <= 0.13315337896347046 and
round_LofTool <= 0.9311000108718872),
-0.0047629233659651970187 * (round_Tajima__s_D_regulatory > 0.4104999899864197 and
round_Transcript_count <= 10.5 and
round_Average_Transcript_length > 1876.067626953125),
0.0068520762408969563759 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi <= 0.919446587562561 and
round_StdDev_Transcript_length <= 1039.699462890625),
-0.028130566254400735798 * (Group == u'NDNE' and
round_Degree <= 4.5 and
round_StdDev_Transcript_length <= 621.470947265625),
-0.024145074195087025404 * (round_Degree <= 12.5 and
round_Degree_mi <= 0.5 and
round_Phi <= 0.12362469732761383 and
round_StdDev_Transcript_length <= 1011.6593627929688),
0.0015991642978274947951 * (round_Degree <= 21.5 and
round_Degree_mi <= 0.5 and
round_Exon_Count > 156.5),
0.088647162920468577929 * (Group == u'NDNE' and
round_Degree > 3.5 and
round_Gene_Length_bp > 41501.0),

0.026665282263039657984 * (round_Degree > 4.5 and
round_missense_Z <= 4.0604472160339355 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length > 1967.8238525390625),
-0.004844570427417022213 9 * (round_Closeness <= 0.2549999952316284 and
round_missense_Z <= 4.052361965179443 and
round_Gene_Length_bp <= 53173.0),
-0.014231203632158318309 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory > -1.2874999046325684 and
round_Blomen_KBM7 <= -0.12487166374921799 and
round_Gene_Length_bp <= 343581.5),
-0.018978210144039470153 * (Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_StdDev_Transcript_length > 636.7213134765625),
0.041662196873145339315 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.3050000071525574 and
round_missense_Z_mi <= 0.5 and
round_Gene_Length_bp > 2814.5),
0.016988971806798404407 * (not Group == u'MNC' and
round_Degree > 62.5 and
round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.2516775131225586),
-0.023466563573408403404 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Gene_Length_bp <= 9941.5 and
round_Average_Transcript_length <= 2037.5650634765625),
-0.021454619164280947646 * (round_Degree <= 12.5 and
round_missense_Z > 4.052361965179443 and
round_Gene_Length_bp <= 110374.0 and
round_Average_Transcript_length <= 2180.86572265625),
0.022697035856952464672 * (round_End <= 100419720.0 and
round_Exon_Count > 221.5),
0.008145665732536789757 6 * (not Group == u'CM' and
not Group == u'MNC' and
round_Blomen_KBM7 > -0.5148604512214661 and
round_Transcript_count > 10.5),
0.029906081927261407571 * (round_missense_Z <= 4.054958820343018 and
round_Transcript_count <= 10.5 and
round_Gene_Length_bp <= 10637.0 and
round_Average_Transcript_length > 2037.5650634765625),
0.047761530637367016761 * (round_End <= 88799552.0 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_StdDev_Transcript_length > 2430.197265625),
0.038357225265835417916 * (round_Degree > 12.5 and
round_Blomen_KBM7 <= -0.25064072012901306),
0.021452043305304778487 * (not Group == u'CM' and
round_End > 127204736.0 and

round_Phi <= 0.9981463551521301 and
round_Exon_Count > 69.5),
-0.048336036395590260828 * (round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 > -0.25219112634658813 and
round_Transcript_count <= 19.5),
0.052244759431406627426 * (round_Tajima__s_D_regulatory <= -
1.2874999046325684 and
round_Blomen_KBM7 > -0.41121259331703186),
-0.014142570663065691036 * (round_Blomen_KBM7 <= -0.2516775131225586 and
round_missense_Z > 4.0613203048706055 and
round_s_het > 0.015855055302381516 and
round_Exon_Count <= 224.5),
0.040131506231853936173 * (round_End <= 110010712.0 and
round_Degree > 61.5 and
round_Degree_mi <= 0.5),
-0.052525036398370784918 * (not Group == u'MNC' and
round_Closeness > 0.2549999952316284 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 987.5745849609375),
-0.03444322178607048951 * (round_Degree <= 12.5 and
round_Blomen_KBM7 > -0.1114160418510437),
0.0080778520322766257655 * (round_Phi > 0.9981780648231506 and
round_missense_Z > 2.6123709678649902 and
round_missense_Z_mi <= 0.5 and
round_Average_Transcript_length > 2957.866943359375),
-0.0018453559825937011947 * (not Group == u'MNC' and
round_Phi > 3.313508932478726e-05 and
round_missense_Z <= 3.2976694107055664),
0.088310649006325236954 * (Group == u'CM'),
-0.3321493231824571013 * (round_Degree_mi > 0.5 and
round_StdDev_Transcript_length > 1060.87158203125),
0.089806398288230407378 * (round_Phi <= 3.313508932478726e-05 and
round_missense_Z <= 4.052361965179443 and
round_LofTool <= 0.04450000077486038),
-0.056141049902893203072 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi <= 0.9640257358551025),
-0.035048251621686281332 * (round_Degree <= 12.5 and
round_Closeness > 0.3050000071525574 and
round_dN_dS_Chimp_mi <= 0.5 and
round_StdDev_Transcript_length > 168.6014404296875),
0.0049429908706841016106 * (not Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_Tajima__s_D_regulatory <= 1.0544999837875366 and
round_missense_Z <= 4.052361965179443),
0.025594410691442658068 * (round_End <= 125209176.0 and
round_Degree > 12.5 and

round_dN_dS_Chimp_mi <= 0.5 and
round_StdDev_Transcript_length > 3690.060546875),
0.020970553574886488524 * (round_Closeness > 0.3149999976158142 and
round_Phi > 0.0015477617271244526 and
round_StdDev_Transcript_length > 626.7237548828125 and
round_Exon_Count <= 283.5),
-0.0072708255207664862149 * (round_Degree_mi <= 0.5 and
round_LofTool <= 0.659500002861023 and
round_StdDev_Transcript_length <= 809.498046875),
-0.0010033902253807347 46 * (round_dN_dS_Chimp > 0.5950000286102295 and
round_Transcript_count <= 26.5 and
round_Average_Transcript_length > 1853.1895751953125),
0.0078267880577386761409 * (not Group == u'CM' and
not Group == u'MNC' and
round_Degree <= 70.0 and
round_LofTool > 0.8695000410079956),
0.30178102419431956926 * (round_Closeness > 0.3149999976158142 and
round_Exon_Count > 224.5),
-0.041635600428811765 65 * (round_Closeness <= 0.2549999952316284 and
round_LofTool <= 0.8535000085830688 and
round_StdDev_Transcript_length <= 617.4835815429688),
0.0030910877541061327484 * (round_Phi <= 0.1378774642944336 and
round_LofTool <= 0.659500002861023 and
round_Average_Transcript_length > 2515.3544921875),
0.029649301367085705017 * (round_dN_dS_Chimp_mi > 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 166.57864379882812),
0.031982851277675909685 * (round_Degree <= 12.5 and
round_Tajima__s_D_regulatory <= 0.4235000014305115 and
round_LofTool <= 0.9922800064086914 and
round_Gene_Length_bp > 37966.5),
-0.017070967066707309207 * (round_dN_dS_Chimp > 0.02499999850988388 and
round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool <= 0.9922449588775635 and
round_Average_Transcript_length > 2037.5650634765625),
0.029724330771197321477 * (round_missense_Z <= 3.1514573097229004 and
round_LofTool <= 0.9922449588775635 and
round_Average_Transcript_length > 2071.857421875),
0.036175061572276949462 * (round_Degree_mi <= 0.5 and
round_LofTool > 0.9886499643325806 and
round_Gene_Length_bp > 39341.0),
0.012711184921391694216 * (round_Phi <= 0.998741626739502 and
round_Gene_Length_bp > 9906.5 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Average_Transcript_length > 1956.6083984375),
0.036358077235948783879 * (not Group == u'NDNE' and
round_Degree > 3.5 and

round_Closeness <= 0.3149999976158142 and
round_s_het > 0.03478143364191055),
-0.062385456839495603831 * (round_s_het <= 0.016313210129737854 and
round_Gene_Length_bp <= 38001.5),
0.098082437827070073633 * (round_Closeness > 0.3149999976158142 and
round_Transcript_count <= 19.5),
0.011020545071693770359 * (round_Degree > 12.5 and
4.052361965179443 < round_missense_Z <= 7.099285125732422 and
round_Gene_Length_bp > 2814.5),
0.06749773290419343319 * (not Group == u'NDNE' and
round_Degree > 62.5),
-0.01639995939674137107 * (round_Degree > 4.5 and
round_Blomen_KBM7 > -0.15807728469371796 and
round_StdDev_Transcript_length <= 1275.7037353515625),
-0.010270465248197970312 * (round_Phi <= 3.313508932478726e-05 and
round_missense_Z <= 4.052361965179443),
-0.014236579328625676225 * (round_s_het <= 0.025150161236524582 and
round_Transcript_count > 13.5 and
round_Average_Transcript_length > 1873.0999755859375),
0.081211064552060438504 * (round_End > 85205200.0 and
round_Gene_Length_bp > 9930.5 and
round_StdDev_Transcript_length > 2430.197265625),
0.008996796161046419821919 * (round_Degree <= 12.5 and
round_Degree_mi <= 0.5 and
round_missense_Z <= 2.120723009109497 and
round_StdDev_Transcript_length > 1011.6593627929688),
-0.12108937018609394753 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory <= -0.4115000069141388 and
round_Gene_Length_bp <= 8050.5),
0.081999324716289165305 * (round_Closeness > 0.32499998807907104 and
round_Tajima__s_D_regulatory <= -1.2874999046325684 and
round_Blomen_KBM7 <= -0.41121259331703186),
-0.04692241815911364633 * (round_Closeness <= 0.3149999976158142 and
round_s_het <= 0.025150161236524582 and
round_Average_Transcript_length > 2057.064453125),
0.043127172504397952302 * (round_End > 100419720.0 and
round_Tajima__s_D_regulatory > 0.5145000219345093 and
round_missense_Z <= 3.996763229370117 and
round_StdDev_Transcript_length > 166.57864379882812),
-0.02368886292594349699 * (not Group == u'CM' and
round_End > 100419720.0 and
round_Degree <= 12.5 and
round_Blomen_KBM7 > -0.514844536781311),
0.012990241102245232707 * (round_Closeness > 0.2549999952316284 and
round_Phi > 0.4316735863685608),
-0.004636481984994052056 6 * (round_Tajima__s_D_regulatory_mi),
0.005152980455141433803 5 * (round_Degree > 3.5 and

round_dN_dS_Chimp <= 0.5950000286102295 and
round_Tajima__s_D_regulatory <= 0.18650001287460327 and
round_Gene_Length_bp <= 343581.5),
0.0438080419095184595 * (not Group == u'NDNE' and
round_Degree > 12.5 and
round_Closeness > 0.32499998807907104 and
round_Gene_Length_bp > 53079.0),
-0.024672830644793142252 * (not Group == u'NDNE' and
round_s_het > 0.016313210129737854 and
round_Gene_Length_bp <= 38001.5),
0.02030768844142512991 * (round_missense_Z <= 4.052361965179443 and
round_LofTool <= 0.04450000077486038 and
round_s_het <= 0.014458265155553818),
0.021488441370476205755 * (not Group == u'MNC' and
round_Closeness > 0.3149999976158142 and
round_StdDev_Transcript_length > 899.939453125),
-0.0377281103161658804 * (Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi <= 0.919446587562561),
-0.047992281795421115609 * (round_Gene_Length_bp <= 2599.0 and
round_StdDev_Transcript_length <= 166.57864379882812),
-0.0030180344075642685266 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Tajima__s_D_regulatory > 0.12399999797344208 and
round_StdDev_Transcript_length > 1944.22802734375),
-0.053645815472417320013 * (round_StdDev_Transcript_length <=
169.24288940429688 and
round_Average_Transcript_length <= 2318.75),
0.005484710865094958622 * (round_Degree > 3.5 and
round_Closeness > 0.3149999976158142 and
round_s_het > 0.02518850564956665 and
round_Transcript_count <= 15.5),
0.01851729216318677082 * (not Group == u'NDNE' and
round_Degree <= 4.5),
-0.009235200960148456234 * (Group == u'MNC' and
round_Tajima__s_D_regulatory > -1.2874999046325684),
0.0048887605175608542241 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Closeness <= 0.3149999976158142 and
round_Phi > 0.8785387277603149),
0.017492794284271317301 * (round_missense_Z <= 4.122845649719238 and
2814.5 < round_Gene_Length_bp <= 9338.5 and
round_StdDev_Transcript_length <= 580.992431640625),
0.0021128405691882255757 * (Group == u'NDNE' and
round_s_het > 0.017306189984083176 and
round_StdDev_Transcript_length > 621.5349731445312 and
round_Exon_Count > 171.0),

-0.028339346103928231974 * (not Group == u'NDNE' and
round_End <= 100419720.0 and
round_Phi > 0.998741626739502 and
round_Transcript_count <= 8.5),
2.8544422492723128641E-05 * (round_missense_Z <= 3.2976694107055664 and
round_Transcript_count > 25.5 and
round_Gene_Length_bp > 9941.5 and
round_Average_Transcript_length <= 2071.857421875),
-0.007224328319621695776 4 * (round_Degree <= 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool > 0.6634999513626099),
-0.022293006054387030229 * (round_LofTool <= 0.6634999513626099 and
round_s_het > 0.01709270477294922 and
round_Exon_Count <= 88.5),
0.022977805449827696377 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Transcript_count <= 15.5),
-0.017991002487537172821 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > 0.17550000548362732 and
round_Transcript_count <= 10.5 and
round_Gene_Length_bp > 2814.5),
0.13358397646296524264 * (round_Degree_mi <= 0.5 and
round_LofTool > 0.9907699823379517),
-0.049028952313931860318 * (round_Phi <= 0.12447576969861984 and
83.5 < round_Exon_Count <= 242.5),
0.0021989595530416358206 * (round_missense_Z),
0.021122374812430694951 * (round_Blomen_KBM7 <= -0.157728910446167 and
round_missense_Z > 3.2971019744873047 and
round_Exon_Count > 96.5),
0.040975473907031947918 * (round_LofTool > 0.9921150207519531 and
round_s_het > 0.025150161236524582),
-0.029039626470798590024 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi <= 0.8768634796142578),
-0.009311298603218606 1125 * (round_Gene_Length_bp > 9978.0 and
round_StdDev_Transcript_length <= 2377.90625 and
round_Exon_Count <= 87.5),
-0.010142192141983820061 * (round_End <= 100419720.0 and
round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Exon_Count <= 221.5),
-0.073112609476678830367 * (round_Blomen_KBM7_mi),
-0.009755204867212227334 8 * (round_Phi <= 0.1378774642944336 and
round_LofTool <= 0.659500002861023 and
round_Average_Transcript_length <= 2515.3544921875),
0.022243743099449538733 * (round_dN_dS_Chimp <= 0.5950000286102295 and
3.2976694107055664 < round_missense_Z <= 4.052361965179443 and

round_LofTool <= 0.9922449588775635),
-0.023271586599581565308 * (round_s_het),
-0.01302008707847174436 * (round_Degree_mi <= 0.5 and
round_missense_Z > 2.552614450454712 and
round_Average_Transcript_length <= 2570.535888671875),
-0.0089483655310949161005 * (round_End <= 100419720.0 and
round_missense_Z > 4.031624794006348 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Exon_Count <= 221.5),
-0.007611925555641918185 14 * (9338.5 < round_Gene_Length_bp <= 37833.5 and
round_StdDev_Transcript_length <= 1956.23095703125),
-0.1477711271932482251 * (round_Degree_mi > 0.5 and
round_Tajima__s_D_regulatory <= 0.47749999165534973),
-0.15850568711767087926 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory > -0.4115000069141388 and
9988.0 < round_Gene_Length_bp <= 53048.5),
-0.027338158540968542087 * (not Group == u'NDNE' and
round_Degree <= 12.5 and
round_missense_Z <= 3.2976694107055664 and
round_LofTool <= 0.9905250072479248),
0.10791704802889216797 * (round_Degree > 35.5 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7 <= -0.12487166374921799 and
round_missense_Z <= 4.052361965179443),
0.038052681275016403406 * (round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.15531033277511597 and
round_missense_Z > 2.721888303756714 and
round_Exon_Count <= 262.0),
0.032420599714973505345 * (round_Degree > 12.5 and
round_s_het > 0.02518850564956665 and
round_Exon_Count <= 284.0),
0.068969116586947529224 * (round_Degree > 35.5 and
round_Degree_mi <= 0.5 and
round_Blomen_KBM7 > -0.1796257197856903 and
round_LofTool <= 0.874500036239624),
0.011001837801926907245 * (round_s_het > 0.02518850564956665 and
round_Transcript_count > 15.5),
0.032911944288080231813 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi > 0.4623493552207947 and
round_LofTool <= 0.9787000417709351),
0.056478680464096327196 * (not Group == u'NDNE' and
round_Degree <= 12.5 and
round_s_het > 0.02518850564956665 and
round_Exon_Count > 160.5),
-0.033531490400225016923 * (not Group == u'NDNE' and
round_Degree > 4.5 and

round_Degree_mi <= 0.5 and
round_Phi <= 0.919446587562561),
0.032944099199805537692 * (-1.2874999046325684 <
round_Tajima__s_D_regulatory <= 1.4165000915527344 and
round_missense_Z <= 4.028769493103027 and
round_Transcript_count > 20.5),
0.028319937430381268706 * (round_Degree > 12.5 and
round_Closeness <= 0.33500000834465027 and
round_Tajima__s_D_regulatory <= 0.4235000014305115 and
round_LofTool <= 0.9922800064086914),
-0.0026354135038181500383 * (round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 <= -0.24819540977478027 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.0070229179471281545991 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.3149999976158142 and
round_Tajima__s_D_regulatory <= 0.4235000014305115 and
round_Exon_Count <= 156.5),
-0.42263729054912357874 * (Group == u'NDNE' and
round_Degree <= 3.5 and
round_StdDev_Transcript_length > 714.263427734375),
0.083714906610359116068 * (round_Closeness > 0.3149999976158142 and
round_Transcript_count > 19.5),
-0.024520343423184185611 * (round_missense_Z <= 0.4934942126274109 and
round_Exon_Count <= 96.5),
0.010109225596981909548 * (not Group == u'NDNE' and
round_Closeness > 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 218.5),
0.003267583662606738693 9 * (round_Degree_mi <= 0.5 and
round_missense_Z <= 3.335038185119629 and
round_StdDev_Transcript_length > 2377.02734375),
0.00060752664301474447461 * (Group == u'NDNE' and
round_Degree <= 4.5 and
round_StdDev_Transcript_length > 714.263427734375),
-0.010258215185811834017 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.2549999952316284 and
round_Phi <= 0.4316735863685608 and
round_Transcript_count <= 9.5),
0.03375802518351931486 * (not Group == u'MNC' and
round_Closeness > 0.33500000834465027 and
3.2976694107055664 < round_missense_Z <= 4.050085067749023),
-0.012795976026728770811 * (round_Degree <= 12.5 and
round_Closeness <= 0.7100000381469727 and
round_missense_Z <= 4.0613203048706055 and
round_Gene_Length_bp <= 352096.0),
0.0023318898906593155812 * (round_Closeness > 0.2549999952316284 and

round_Phi <= 0.4316735863685608 and
round_Transcript_count <= 9.5),
-0.0066755401432419267729 * (round_Degree <= 12.5 and
round_Closeness <= 0.3050000071525574 and
round_dN_dS_Chimp_mi <= 0.5 and
round_StdDev_Transcript_length > 168.6014404296875),
-0.0034145756247275114645 * (round_LofTool <= 0.6634999513626099 and
round_s_het <= 0.01709270477294922 and
round_Exon_Count <= 88.5),
0.020834681193880762867 * (round_Tajima__s_D_regulatory <=
1.4165000915527344 and
round_Blomen_KBM7 <= -0.5153244733810425 and
round_Transcript_count > 10.5 and
round_Gene_Length_bp > 10640.5),
-0.001953615307148606154 * (round_Phi <= 0.12447576969861984 and
round_LofTool > 0.9605500102043152 and
round_StdDev_Transcript_length > 1452.739013671875),
0.014062538630450481178 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_Average_Transcript_length > 2570.535888671875),
0.0052869251864379315439 * (round_Phi > 0.00015248148702085018 and
round_missense_Z <= 3.2976694107055664 and
round_Exon_Count <= 87.5),
-0.02333450869429674196 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory > 1.2934999465942383 and
round_Blomen_KBM7 > -0.4999815821647644),
-0.0013849296134431871106 * (Group == u'NDNE' and
round_s_het > 0.01700003445148468 and
round_StdDev_Transcript_length <= 704.5946044921875 and
round_Exon_Count <= 86.5),
0.011101717126482805661 * (round_Degree <= 12.5 and
round_s_het > 0.017613736912608147 and
round_Gene_Length_bp > 38001.5 and
round_Exon_Count > 221.5),
0.070135120685213117597 * (round_Closeness > 0.3149999976158142 and
round_Phi <= 0.8754478693008423),
-0.064495788823159316827 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_LofTool <= 0.8554999828338623),
0.00050612016669284083138 * (round_End <= 100419720.0 and
round_Tajima__s_D_regulatory > -1.2885000705718994 and
round_missense_Z_mi > 0.5),
-0.073034808226031561196 * (round_Gene_Length_bp > 2599.0 and
round_StdDev_Transcript_length <= 166.57864379882812),
-0.032090827359876047953 * (Group == u'NDNE' and
round_missense_Z > 2.7258143424987793),

```
        0.040659733289495375574 * (round_Tajima__s_D_regulatory <= -
1.2874999046325684 and
                    round_missense_Z <= 3.992063522338867 and
                    round_Exon_Count <= 224.5),
        0.026721043939407669587 * (not Group == u'NDNE' and
                    round_Closeness > 0.2549999952316284 and
                    round_LofTool <= 0.9787000417709351 and
                    round_StdDev_Transcript_length <= 1005.2410888671875),
       0.009139524908486320459 2 * (not Group == u'MNC' and
                    round_Degree > 5.5 and
                    round_Phi > 0.0015477617271244526 and
                    round_Blomen_KBM7 <= -0.18366125226020813),
       0.006062895954981630669 6 * (round_End <= 127637464.0 and
                    round_dN_dS_Chimp_mi <= 0.5 and
                    round_Gene_Length_bp > 9941.5 and
                    round_Average_Transcript_length <= 2037.5650634765625),
       -0.039510142557001665109 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
                    round_missense_Z > 3.8540451526641846),
        0.019811125846081169277 * (round_Degree <= 4.5 and
                    round_missense_Z <= 4.0604472160339355 and
                    round_Transcript_count <= 20.5 and
                    round_StdDev_Transcript_length > 1967.8238525390625),
      -0.001859945634251888857 4 * (Group == u'NDNE' and
                    round_End <= 100419720.0 and
                    round_Tajima__s_D_regulatory > -1.2885000705718994 and
                    round_Blomen_KBM7 <= -0.09869569540023804),
       0.002156215631248258046 7 * (not Group == u'MNC' and
                    round_dN_dS_Chimp_mi <= 0.5 and
                    round_Blomen_KBM7 > -0.5146373510360718 and
                    round_Gene_Length_bp > 37999.5),
       -0.049805293559377718238 * (Group == u'NDNE' and
                    round_StdDev_Transcript_length <= 621.5349731445312),
        0.014492269940647846405 * (round_End > 100419720.0 and
                    round_missense_Z > 4.076244354248047 and
                    round_LofTool <= 0.994350016117096),
         0.1236092272944794429 * (round_LofTool > 0.9922449588775635),
       -0.031030338902052499728 * (Group == u'NDNE' and
                    round_Degree <= 12.5 and
                    round_Average_Transcript_length <= 2169.2587890625),
        0.026392057958906565279 * (round_Blomen_KBM7 > -0.49977701902389526 and
                    round_LofTool > 0.9922449588775635),
        0.065464842259792752066 * (round_LofTool <= 0.9920099973678589 and
                    round_Exon_Count > 221.5),
       0.005569326062194727977 6 * (round_Degree > 3.5 and
                    round_Closeness <= 0.3149999976158142 and
                    round_Exon_Count > 151.5),
```

-0.012235121495508578457 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_s_het > 0.021862218156456947),
0.019996450446880205398 * (round_Tajima__s_D_regulatory >
0.4235000014305115 and
round_Blomen_KBM7 <= -0.21081802248954773 and
round_s_het > 0.015080630779266357 and
round_StdDev_Transcript_length > 580.992431640625),
-0.0010861882066284769995 * (Group == u'NDNE' and
round_Degree > 4.5 and
round_Degree_mi <= 0.5 and
round_Phi > 0.919446587562561),
-0.020932057542549405149 * (round_dN_dS_Chimp_mi > 0.5 and
round_Tajima__s_D_regulatory > -1.2909998893737793),
0.0043345273324120409467 * (not Group == u'MNC' and
round_dN_dS_Chimp_mi > 0.5 and
round_Tajima__s_D_regulatory > -1.2874999046325684 and
round_Blomen_KBM7 <= -0.4999815821647644),
-0.0056254294468995819437 * (not Group == u'CM' and
not Group == u'MNC' and
-1.2855000495910645 < round_Tajima__s_D_regulatory <=
1.4165000915527344),
-0.051135930907465425299 * (round_missense_Z <= 3.04426908493042 and
round_StdDev_Transcript_length <= 167.33212280273438),
0.0012261433532188758741 * (round_Degree > 4.5 and
round_Phi > 0.12357446551322937 and
round_Blomen_KBM7 <= -0.15807728469371796 and
round_StdDev_Transcript_length <= 1275.7037353515625),
0.023396898187144792025 * (not Group == u'NDNE' and
round_Closeness <= 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_Gene_Length_bp > 45328.0),
-0.00073349258122191681765 * (round_Degree > 4.5 and
round_Closeness <= 0.3149999976158142 and
round_StdDev_Transcript_length > 1275.7037353515625),
0.00187732465779288514 2 * (round_Degree <= 9.5 and
round_Phi > 0.1378774642944336 and
round_LofTool <= 0.659500002861023),
-7.2347323261013874146E-06 * (round_Exon_Count),
-0.0046239247092063531092 * (round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Average_Transcript_length > 2071.857421875),
-0.090305482829968683478 * (round_Phi <= 0.0002437000221107155 and
round_missense_Z <= 4.0604472160339355 and
round_StdDev_Transcript_length > 615.416259765625),
0.01648482491782999812 * (round_Degree <= 45.0 and

round_dN_dS_Chimp_mi > 0.5 and
round_Transcript_count <= 26.5 and
round_Gene_Length_bp > 2402.0),
-0.025610478117531781939 * (not Group == u'NDNE' and
round_Degree <= 57.5 and
round_missense_Z <= 2.642360210418701 and
round_Exon_Count <= 190.5),
0.013462495283131365245 * (not Group == u'CM' and
not Group == u'MNC' and
round_Degree > 70.0 and
round_LofTool > 0.8695000410079956),
0.010003967660577692267 * (round_dN_dS_Chimp_mi > 0.5 and
round_s_het <= 0.43051624298095703 and
round_Gene_Length_bp > 2821.0),
-0.25087679754859976144 * (round_Phi > 0.00015248148702085018 and
round_Blomen_KBM7 <= -0.15492522716522217 and
round_LofTool <= 0.6634999513626099 and
round_Exon_Count <= 157.5),
-0.20988131906886342559 * (round_Degree_mi > 0.5 and
round_Tajima__s_D_regulatory > 0.47749999165534973),
0.014989444923831543588 * (Group == u'MNC' and
round_Phi > 0.0015477617271244526 and
round_Exon_Count > 96.5),
-0.0083027455186904398216 * (not Group == u'MNC' and
round_Degree <= 5.5 and
round_Gene_Length_bp <= 43309.0),
-0.16073662255780979402 * (round_LofTool <= 0.6634999513626099 and
round_Exon_Count > 157.5),
0.035010641447032482543 * (2630.5 < round_Gene_Length_bp <= 9978.0 and
round_StdDev_Transcript_length <= 2377.90625),
-0.021649303102464188125 * (round_missense_Z <= 3.3354156017303467 and
round_StdDev_Transcript_length <= 636.7213134765625),
-0.02887978211493740649 * (round_missense_Z > 3.3354156017303467 and
round_StdDev_Transcript_length <= 636.7213134765625),
-0.015589426309587621142 * (not Group == u'CM' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Tajima__s_D_regulatory > 1.062999963760376 and
round_Blomen_KBM7_mi <= 0.5),
0.007869031006701419223 * (not Group == u'NDNE' and
round_Tajima__s_D_regulatory > 0.5564999580383301),
0.019198469558738117753 * (round_Closeness <= 0.2549999952316284 and
round_missense_Z <= 4.052361965179443 and
round_s_het > 0.014458265155553818 and
round_s_het_mi > 0.5),
-0.013523761746092448702 * (round_Closeness > 0.2549999952316284 and
round_missense_Z <= 1.43977952003479 and
round_StdDev_Transcript_length <= 902.3864135742188 and

round_Exon_Count <= 260.5),
0.056557835112290621993 * (not Group == u'NDNE' and
round_Degree > 4.5 and
round_LofTool > 0.9827499985694885),
1.0771704140300060 06E-06 * (round_StdDev_Transcript_length),
0.0067367214229233718728 * (not Group == u'NDNE' and
round_Closeness > 0.32499998807907104 and
round_missense_Z > 1.43977952003479),
-0.025280204319459625983 * (round_Blomen_KBM7 > -0.24819540977478027 and
round_missense_Z <= 3.2976694107055664),
-0.07144481186709027154 * (Group == u'NDNE' and
round_missense_Z <= 1.1351158618927002 and
round_Transcript_count <= 10.5),
-0.034050179534380775603 * (round_End <= 100419720.0 and
round_Closeness > 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z <= 2.852269172668457),
0.038061134138614192979 * (not Group == u'MNC' and
round_missense_Z <= 2.642360210418701 and
round_s_het > 0.014316117390990257 and
round_Exon_Count > 165.5),
-0.076712140314540558372 * (round_Closeness > 0.2549999952316284 and
round_Phi > 0.0002437000221107155 and
round_missense_Z <= 4.0604472160339355 and
round_StdDev_Transcript_length > 615.416259765625),
-0.039402444954045441616 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp > 0.5950000286102295),
-0.084793394898434820695 * (Group == u'NDNE' and
round_Degree > 4.5 and
round_Phi <= 0.13842733204364777),
-0.012276591717802870854 * (round_Closeness <= 0.33500000834465027 and
round_Blomen_KBM7 > -0.49977701902389526 and
round_missense_Z > 4.052361965179443 and
round_LofTool <= 0.9922449588775635),
0.0049006772599034468391 * (round_End <= 127093776.0 and
round_Tajima__s_D_regulatory > -1.2894999980926514 and
round_Blomen_KBM7 <= -0.49977701902389526 and
round_Exon_Count <= 218.5),
-0.017059843137643783406 * (0.659500002861023 < round_LofTool <=
0.9921150207519531 and
round_StdDev_Transcript_length <= 589.7366333007812 and
round_Exon_Count <= 174.5),
-0.007579869113845304453 17 * (round_End <= 100419720.0 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7_mi <= 0.5 and
round_Exon_Count <= 221.5),
0.010744566051742131946 * (12.5 < round_Degree <= 60.5 and

round_Tajima__s_D_regulatory <= 1.0529999732971191),
-0.38591243129541930035 * (Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi > 0.14362311363220215),
-0.040250397690003617002 * (not Group == u'NDNE' and
round_Phi <= 0.12447576969861984 and
round_Exon_Count <= 83.5),
0.0054042513292536526609 * (not Group == u'NDNE' and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 93.5),
0.012640846218055473704 * (round_Degree > 4.5 and
round_Blomen_KBM7 <= -0.1469864696264267 and
round_missense_Z > 4.099565505981445 and
round_Exon_Count <= 224.5),
-0.011599627423029612583 * (not Group == u'CM' and
3.5 < round_Degree <= 56.5 and
round_Exon_Count <= 253.5),
0.024923686490065564969 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Phi <= 0.9998619556427002 and
round_StdDev_Transcript_length > 2625.484375),
-0.024788357238958731721 * (not Group == u'CM' and
not Group == u'MNC' and
round_Tajima__s_D_regulatory <= 0.4104999899864197 and
round_Transcript_count <= 20.5),
-0.032616344153410088691 * (4.5 < round_Degree <= 29.5 and
round_StdDev_Transcript_length <= 897.559326171875),
-0.035027548355530679913 * (Group == u'NDNE' and
round_Degree <= 3.5 and
round_StdDev_Transcript_length > 1049.45654296875),
-0.0074999311629464779708 * (Group == u'NDNE' and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 1967.8238525390625),
0.033847444063908105338 * (round_Degree > 12.5 and
round_LofTool <= 0.992115020751953 1 and
round_s_het > 0.025150161236524582 and
round_Exon_Count <= 224.5),
0.018672957826625806443 * (round_Phi <= 0.00015248148702085018 and
round_missense_Z <= 4.052361965179443 and
round_LofTool > 0.11749999970197678),
-0.0066963862081202731036 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.3149999976158142 and
round_Tajima__s_D_regulatory > 0.4235000014305115 and
round_Exon_Count <= 156.5),
-0.19403437417903002249 * (round_Degree <= 4.5 and
round_Phi > 0.14253592491149902),
0.025474159388727657394 * (round_Degree > 4.5 and

round_Closeness > 0.3149999976158142 and
round_StdDev_Transcript_length > 1275.7037353515625),
-0.15200914772366930228 * (round_Gene_Length_bp <= 2814.5),
-0.012369173690002560964 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_missense_Z <= 2.7698380947113037 and
round_Gene_Length_bp <= 39341.0),
0.019096003975550827902 * (round_Degree <= 4.5 and
round_Degree_mi <= 0.5 and
round_StdDev_Transcript_length <= 1301.4395751953125),
-0.060655818768655632434 * (round_Blomen_KBM7 > -0.5146373510360718 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length > 2388.5126953125),
-0.022501726195653822676 * (not Group == u'NDNE' and
round_Degree <= 62.5 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 580.6193237304688),
0.06316842250916585022 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi <= 0.8768634796142578 and
round_LofTool <= 0.9829000234603882),
0.039685738909723246304 * (Group == u'NDNE' and
round_Degree <= 3.5 and
round_missense_Z > 1.472226858139038),
-0.002509813077558311397 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.32499998807907104 and
round_LofTool <= 0.9435499906539917 and
round_Gene_Length_bp <= 54517.0),
-0.039332187759655926063 * (round_dN_dS_Chimp > 0.5950000286102295),
-0.06643661699653181929 * (Group == u'NDNE' and
round_Closeness > 0.2549999952316284),
-0.021412501080032213946 * (round_Degree > 3.5 and
round_Phi <= 0.12447576969861984 and
round_Exon_Count <= 156.5),
-0.00029745582835470673125 * (round_Transcript_count),
0.057181720796839995147 * (round_Closeness > 0.3149999976158142 and
round_Phi > 0.12447576969861984),
-1.1772852955159365113E-05 * (round_Tajima__s_D_regulatory <=
0.4165000021457672 and
round_Phi > 0.00015248148702085018 and
round_missense_Z <= 3.308867931365967),
-0.078557558657171636107 * (round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 4.0604472160339355 and
round_StdDev_Transcript_length <= 615.416259765625),
0.017463172080480923037 * (round_Closeness > 0.2549999952316284 and
round_LofTool > 0.9937300086021423 and
round_Gene_Length_bp > 39268.5),

-0.0004124438619598738329 2 (round_Degree <= 60.5 and
round_dN_dS_Chimp_mi > 0.5 and
round_Blomen_KBM7 > -0.5153281092643738 and
round_Exon_Count <= 154.5),
-0.012984409356028072183 2 (round_Degree > 4.5 and
round_Phi > 0.0003650499857030809 and
round_Blomen_KBM7 <= -0.1469864696264267 and
round_missense_Z <= 4.099565505981445),
-0.013769490895467194 2 (not Group == u'MNC' and
round_Degree <= 34.5 and
round_Degree_mi <= 0.5 and
round_missense_Z <= 4.0604472160339355),
0.026191082501492171652 2 (round_missense_Z <= 2.695338487625122 and
round_Transcript_count > 26.5 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.0087539696043418078336 2 (not Group == u'CM' and
round_Closeness <= 0.7100000381469727 and
round_Blomen_KBM7 <= -0.49977701902389526 and
round_LofTool <= 0.6634999513626099),
0.015364239969804956848 2 (round_Degree > 37.5 and
round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.24797455966472626 and
round_missense_Z <= 2.721888303756714),
-0.013768597282704474888 2 (round_Degree > 5.5 and
round_Blomen_KBM7 <= -0.15774735808372498 and
round_Average_Transcript_length <= 2566.631103515625),
0.0016667767258665866937 2 (Group == u'NDNE' and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 93.5),
0.18128583819906532448 2 (round_Closeness > 0.3149999976158142 and
round_Blomen_KBM7 <= -0.5528146028518677 and
round_Exon_Count <= 224.5),
0.012943950146806795093 2 (round_Degree > 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_Gene_Length_bp > 2814.5),
-0.0049955332334434728037 2 (round_LofTool),
0.01416400139243850601 9 2 (round_Blomen_KBM7 <= -0.5153281092643738 and
round_missense_Z > 3.289116144180298),
0.050500102827979391484 2 (round_End > 124883040.0 and
-1.2874999046325684 < round_Tajima__s_D_regulatory <=
1.4165000915527344 and
round_Transcript_count <= 20.5),
0.060456640460729016429 2 (not Group == u'NDNE' and
round_Degree <= 12.5 and
round_StdDev_Transcript_length > 636.7213134765625 and
round_Exon_Count <= 253.5),

-0.033822224075698618939 * (round_Blomen_KBM7),
-0.0039964088486853953028 * (round_Closeness <= 0.33500000834465027 and
round_Blomen_KBM7 > -0.49977701902389526 and
round_missense_Z <= 4.052361965179443 and
round_LofTool <= 0.9922449588775635),
0.023975111109943367249 * (round_End <= 85205200.0 and
round_Gene_Length_bp > 9930.5 and
round_StdDev_Transcript_length > 2430.197265625),
-0.011463148479432971882 * (round_Tajima__s_D_regulatory >
1.0544999837875366 and
round_missense_Z <= 4.052361965179443 and
round_Average_Transcript_length <= 2288.774658203125),
-0.0051712503030842075363 * (round_Degree_mi <= 0.5 and
round_missense_Z <= 2.552614450454712 and
round_LofTool <= 0.6644999980926514 and
round_Average_Transcript_length <= 2570.535888671875),
0.037739634085657752793 * (round_missense_Z > 3.2976694107055664),
-0.021570209135655677574 * (round_Closeness > 0.3149999976158142 and
round_Phi <= 0.12447576969861984 and
round_Exon_Count <= 83.5),
-0.021510107011690166728 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi > 0.9640257358551025),
0.0014163650512331741119 * (round_Degree <= 12.5 and
round_missense_Z > 3.2976694107055664),
-0.010674563779895991297 * (4.050085067749023 < round_missense_Z <=
5.5646071434021),
0.041860350492729625493 * (round_Degree > 35.5 and
round_Degree_mi <= 0.5 and
round_Closeness > 0.2549999952316284 and
round_Phi <= 0.8162848949432373),
0.0056832531972258086908 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Transcript_count > 26.5 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.063871267111750978929 * (round_Degree_mi > 0.5 and
round_Phi <= 0.8162848949432373),
0.03489118407860661991 2 * (3.5 < round_Degree <= 63.0 and
round_Phi > 0.919446587562561),
-0.0018919554407697701439 * (round_Phi <= 0.0015477617271244526 and
round_Gene_Length_bp > 50399.0),
0.013992933384094302304 * (round_Phi > 0.0023738450836390257 and
round_Blomen_KBM7 <= -0.18453440070152283 and
round_Transcript_count > 26.5 and
round_Exon_Count > 87.5),
-0.051195693008280626635 * (round_Degree <= 3.5 and
round_s_het <= 0.02518850564956665),
0.055834058359271090954 * (not Group == u'NDNE' and

round_Degree <= 12.5 and
round_s_het > 0.02518850564956665 and
round_Exon_Count <= 160.5),
-0.0082870956775982305281 * (round_Closeness <= 0.2549999952316284 and
round_Tajima__s_D_regulatory > 0.4104999899864197),
0.038547688444224244286 * (Group == u'MNC' and
round_Closeness > 0.3149999976158142),
0.013900893462213393531 * (round_Closeness > 0.2549999952316284 and
round_LofTool > 0.7979999780654907 and
round_Gene_Length_bp <= 41216.0),
0.0050657262704566644387 * (round_Closeness > 0.2549999952316284 and
round_Phi <= 0.00013825652422383428 and
round_missense_Z <= 3.334120750427246),
-0.018169172998649386203 * (round_End <= 100419720.0 and
round_Degree_mi > 0.5),
0.0044089773122458876878 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
round_Blomen_KBM7 > -0.12487166374921799 and
round_s_het > 0.02633928880095482 and
round_Exon_Count > 53.0),
-0.071852306419135800186 * (round_dN_dS_Chimp_mi > 0.5 and
round_missense_Z <= 4.0604472160339355 and
round_StdDev_Transcript_length <= 615.416259765625),
0.073788341751617983477 * (round_Degree > 60.5 and
round_Tajima__s_D_regulatory <= 1.0529999732971191),
-0.0019138125607677928565 * (round_Blomen_KBM7 > -0.5153281092643738 and
round_missense_Z <= 4.103667736053467 and
round_Transcript_count <= 24.5 and
round_Gene_Length_bp > 53047.5),
0.0038228233805468426684 * (round_Closeness <= 0.3149999976158142 and
round_missense_Z <= 2.571293354034424 and
round_StdDev_Transcript_length > 1450.0491943359375),
0.009786301706510392578 5 * (not Group == u'NDNE' and
round_Closeness <= 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 218.5),
-0.064557916522001199122 * (round_Degree_mi > 0.5 and
round_Gene_Length_bp > 2510.5),
0.017207469487847969897 * (not Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_LofTool <= 0.9829000234603882 and
round_Average_Transcript_length > 1918.333251953125),
0.016078813237064535496 * (round_Tajima__s_D_regulatory <=
0.195499986410141 and
round_missense_Z > 3.2976694107055664),
0.0054438014970486035132 * (round_Closeness <= 0.2549999952316284 and
round_missense_Z <= 4.052361965179443 and

round_s_het > 0.014458265155553818 and
round_s_het_mi <= 0.5),
-0.056622008362259834691 * (round_Degree <= 3.5 and
0.6634999513626099 < round_LofTool <= 0.9905250072479248),
-0.009382110839505006239 * (round_Closeness <= 0.26499998569488525 and
round_LofTool <= 0.8105000257492065 and
round_Gene_Length_bp > 37966.5),
-0.022079473690573211964 * (round_missense_Z <= 3.2976694107055664 and
round_Transcript_count <= 25.5 and
round_Gene_Length_bp > 9941.5 and
round_Average_Transcript_length <= 2071.857421875),
0.026567551484747074092 * (round_Degree_mi > 0.5 and
round_LofTool > 0.6614999771118164),
0.012171923038907185924 * (not Group == u'MNC' and
round_Closeness <= 0.33500000834465027 and
3.2976694107055664 < round_missense_Z <= 4.050085067749023),
-0.007635595057066103170 91 * (round_Degree <= 5.5 and
round_LofTool > 0.6554999947547913 and
round_StdDev_Transcript_length <= 1337.2225341796875),
0.16325429384657416665 * (round_Tajima__s_D_regulatory <=
0.4235000014305115 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_s_het > 0.015603477135300636 and
round_Transcript_count > 21.5),
0.0072661225638375511598 * (round_End <= 100419720.0 and
round_StdDev_Transcript_length > 166.57864379882812 and
round_Exon_Count > 221.5),
0.036323776432227984634 * (round_Degree <= 37.5 and
round_Degree_mi <= 0.5 and
round_Blomen_KBM7 <= -0.24797455966472626 and
round_missense_Z <= 2.721888303756714),
-0.013138779464712927597 * (not Group == u'NDNE' and
round_Degree > 4.5 and
round_LofTool <= 0.656499981880188),
0.019063593314076931334 * (not Group == u'NDNE' and
round_Blomen_KBM7 > -0.5146430730819702 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 2418.21435546875),
-0.022312166919272659327 * (not Group == u'NDNE' and
round_missense_Z > 2.642360210418701 and
round_Exon_Count <= 285.0),
0.046627845220519749392 * (round_Degree_mi > 0.5 and
round_Tajima__s_D_regulatory > 0.47749999165534973 and
round_Blomen_KBM7 <= -0.12487166374921799),
-0.057144599564295774086 * (round_Phi_mi),
-0.042548091410918198463 * (Group == u'MNC' and

167.33212280273438 < round_StdDev_Transcript_length <=
615.8639526367188),
-0.078599340027315722779 * (Group == u'NDNE' and
round_StdDev_Transcript_length <= 714.5440063476562),
-0.035888903100163335735 * (Group == u'NDNE' and
round_Phi <= 0.12447576969861984 and
round_Exon_Count <= 83.5),
0.014099712983019761434 * (round_Tajima__s_D_regulatory >
0.4104999899864197 and
round_Exon_Count > 88.5),
-0.025909213199834423696 * (round_Degree > 4.5 and
round_Closeness <= 0.3149999976158142 and
round_Transcript_count <= 24.5 and
round_StdDev_Transcript_length > 897.559326171875),
-0.007484535984970042653 * (round_Tajima__s_D_regulatory >
0.43549999594688416 and
round_missense_Z <= 3.2976694107055664 and
round_LofTool <= 0.9603500366210938 and
round_Exon_Count <= 164.5),
-0.01077622969677029946 * (round_Degree_mi > 0.5 and
round_dN_dS_Chimp > 0.23499999940395355 and
round_Tajima__s_D_regulatory <= 0.47749999165534973 and
round_Blomen_KBM7 <= -0.12487166374921799),
-0.30401281812512331859 * (round_Degree <= 4.5 and
round_Phi <= 0.14253592491149902 and
round_StdDev_Transcript_length <= 616.5709228515625),
-0.028990481621997066936 * (round_Phi <= 0.12447576969861984 and
round_Transcript_count > 19.5 and
round_StdDev_Transcript_length > 1098.5848388671875),
-0.18916948110773898484 * (3.134337902069092 < round_missense_Z <=
3.308867931365967 and
round_s_het > 0.015855055302381516),
0.007932780135499323953 5 * (round_Blomen_KBM7 <= -0.5154723525047302 and
round_missense_Z <= 2.6123709678649902 and
round_StdDev_Transcript_length > 166.57864379882812),
0.009632205248113128587 3 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Phi > 0.8162848949432373 and
round_LofTool <= 0.9833999872207642),
-0.010776147984111764111 * (round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 3.308867931365967 and
round_Average_Transcript_length > 2037.5650634765625),
0.043118916775937328467 * (round_Phi <= 3.313508932478726e-05 and
round_missense_Z <= 4.052361965179443 and
round_LofTool > 0.11749999970197678),
-0.033825661237730453301 * (round_Closeness <= 0.2549999952316284 and
round_s_het <= 0.021862218156456947),

-0.047543359055394369961 * (not Group == u'MNC' and
round_s_het <= 0.025150161236524582 and
round_Average_Transcript_length <= 1873.0999755859375),
0.043005070335481156152 * (round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool > 0.9922449588775635),
-0.0193305402296465971 * (round_Phi > 0.9985461235046387 and
round_Transcript_count <= 10.5 and
round_Gene_Length_bp > 53048.5),
-0.0029334074033163580301 * (round_End <= 127204736.0 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Gene_Length_bp > 2814.5 and
round_Average_Transcript_length <= 2037.5650634765625),
-0.014088246133160155227 * (not Group == u'MNC' and
round_End <= 100419720.0 and
round_Degree <= 12.5 and
round_Blomen_KBM7 <= -0.11068554222583771),
-0.031492051646561769473 * (not Group == u'NDNE' and
round_End <= 100419720.0 and
round_Phi > 0.998741626739502 and
round_Transcript_count > 8.5),
0.064514690954963371805 * (round_End > 100419720.0 and
round_LofTool > 0.994350016117096),
0.0037794234205987689915 * (round_Gene_Length_bp <= 9930.5 and
round_StdDev_Transcript_length > 166.57864379882812),
-0.046813844593473395717 * (Group == u'NDNE' and
round_Transcript_count > 10.5),
0.0047308280825400263192 * (round_Closeness > 0.2549999952316284 and
0.6694999933242798 < round_LofTool <= 0.9937300086021423 and
round_Gene_Length_bp > 39268.5),
0.031799043486228008304 * (round_Phi > 0.00015248148702085018 and
round_Blomen_KBM7 <= -0.5508977174758911 and
round_missense_Z <= 4.0604472160339355 and
round_Exon_Count > 87.5),
0.024599510773753802129 * (Group == u'NDNE' and
round_Degree <= 4.5 and
round_StdDev_Transcript_length > 621.470947265625),
-0.024818309340896429344 * (round_Closeness <= 0.2549999952316284 and
round_LofTool <= 0.8535000085830688 and
round_StdDev_Transcript_length > 617.4835815429688),
-0.022986810189580389463 * (not Group == u'CM' and
round_Degree <= 35.5 and
round_Closeness > 0.3050000071525574 and
round_Blomen_KBM7 > -0.49977701902389526),
0.024818347921128428024 * (round_Closeness <= 0.33500000834465027 and
round_dN_dS_Chimp_mi > 0.5 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_Gene_Length_bp > 2814.5),

0.030785675449449489971 * (round_Degree > 9.5 and
round_Phi > 0.1378774642944336 and
round_LofTool <= 0.659500002861023),
0.071141168378547961493 * (round_Phi > 0.12447576969861984 and
round_Exon_Count > 218.5),
-0.0033207787269117858969 * (round_Blomen_KBM7 <= -0.14714285731315613 and
round_missense_Z <= 3.2974047660827637 and
round_LofTool <= 0.9922449588775635 and
round_StdDev_Transcript_length > 615.8639526367188),
-0.021924118897895659985 * (not Group == u'CM' and
not Group == u'MNC' and
round_Degree <= 35.5 and
round_Blomen_KBM7 > -0.49977701902389526),
-0.015856059453570228723 * (round_dN_dS_Chimp > 0.5950000286102295 and
round_LofTool > 0.04450000077486038 and
round_Exon_Count <= 217.0),
0.055378231519672874161 * (round_LofTool_mi),
-0.006130853536874806300598 * (0.3050000071525574 < round_Closeness <=
0.3149999976158142 and
round_Blomen_KBM7 > -0.5158457159996033 and
round_Gene_Length_bp <= 53140.5),
-0.021087248754085095859 * (not Group == u'CM' and
round_Tajima__s_D_regulatory > 0.4165000021457672 and
round_missense_Z <= 4.031401634216309 and
round_LofTool > 0.6634999513626099),
-0.028317726879795258876 * (round_Degree <= 12.5 and
round_Degree_mi <= 0.5 and
round_Phi > 0.12362469732761383 and
round_StdDev_Transcript_length <= 1011.6593627929688),
-0.00028572000111677241414 * (not Group == u'NDNE' and
round_Degree > 4.5 and
round_LofTool <= 0.9827499985694885 and
round_Gene_Length_bp > 39266.5),
-0.025054992712295911378 * (round_Degree <= 4.5 and
round_Degree_mi > 0.5 and
round_StdDev_Transcript_length > 1049.45654296875),
0.14453713378768545672 * (round_Closeness > 0.3149999976158142 and
2.642360210418701 < round_missense_Z <= 4.0604472160339355),
-0.019256270589407667448 * (round_StdDev_Transcript_length <=
169.24288940429688 and
round_Average_Transcript_length > 2318.75),
-0.072220477391367943198 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_s_het <= 0.02767527475953102),
0.036591368423358389128 * (round_dN_dS_Chimp_mi > 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_Average_Transcript_length > 2056.5712890625 and

round_Exon_Count <= 240.5),
0.009821830589055382 7403 * (round_missense_Z <= 4.050085067749023 and
round_Transcript_count > 10.5 and
round_Average_Transcript_length > 2037.5650634765625),
0.002504350471971457 8151 * (Group == u'CM' and
3.5 < round_Degree <= 56.5 and
round_Exon_Count <= 253.5),
-0.000272136153580052 34643 * (not Group == u'MNC' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 <= -0.12487166374921799 and
round_s_het > 0.014458265155553818),
-0.012962701962375372 186 * (round_Phi > 0.002197184134274721 and
round_LofTool <= 0.6634999513626099 and
round_Exon_Count > 88.5),
0.01726754044835551033 * (not Group == u'NDNE' and
round_missense_Z > 2.5613200664520264 and
round_s_het > 0.01700003445148468 and
round_Exon_Count <= 86.5),
0.047407821001623717816 * (not Group == u'CM' and
not Group == u'MNC' and
round_Blomen_KBM7 <= -0.5148604512214661 and
round_Transcript_count > 26.5),
-0.012239859477472964447 * (not Group == u'MNC' and
round_End <= 100419720.0 and
round_Degree <= 34.5 and
round_Tajima__s_D_regulatory > -1.2855000495910645),
-0.010213778329625312902 * (round_Degree <= 45.0 and
round_dN_dS_Chimp_mi > 0.5 and
round_Transcript_count > 1.5 and
round_StdDev_Transcript_length <= 1942.072509765625),
0.002697975622899676 3096 * (round_Degree_mi <= 0.5 and
round_Tajima__s_D_regulatory > -1.2885000705718994 and
round_Blomen_KBM7 <= -0.2516775131225586 and
round_s_het > 0.015896810218691826),
-0.034743371535693644281 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_missense_Z <= 1.43977952003479),
-0.000326704124917936 54431 * (round_Degree <= 60.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5153281092643738 and
round_Exon_Count <= 154.5),
-0.006933531040095820 4829 * (round_Degree <= 3.5 and
round_Closeness <= 0.3149999976158142 and
round_Blomen_KBM7 <= -0.25219112634658813 and
round_LofTool <= 0.8695000410079956),
-0.084412030034841656345 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and

round_LofTool > 0.8554999828338623),
-0.006800407811345367259 4 * (round_Degree > 18.5 and
round_Blomen_KBM7 <= -0.14956465363502502 and
round_LofTool > 0.659500002861023),
0.019252153925983447186 * (round_Degree > 62.5),
-0.0049752751404963432152 * (not Group == u'CM' and
round_dN_dS_Chimp_mi <= 0.5 and
round_Tajima__s_D_regulatory <= 1.062999963760376 and
round_Blomen_KBM7_mi <= 0.5),
0.003356968599162736891 * (round_Phi > 0.12447576969861984 and
round_missense_Z > 4.050085067749023),
-0.011793238037396919921 * (round_Closeness <= 0.2549999952316284 and
round_Blomen_KBM7_mi <= 0.5 and
round_s_het_mi <= 0.5 and
round_StdDev_Transcript_length <= 2625.484375),
-0.051565317012112775463 * (round_missense_Z > 4.052361965179443 and
round_LofTool <= 0.9499499797821045),
-0.019198707422328438466 * (round_End <= 100419720.0 and
round_Blomen_KBM7_mi > 0.5 and
round_Exon_Count <= 221.5),
0.028560411574742990137 * (round_Degree <= 12.5 and
round_Blomen_KBM7 <= -0.1114160418510437 and
round_StdDev_Transcript_length > 2377.90625 and
round_Average_Transcript_length > 2169.3193359375),
-0.008160214475651253017 * (round_Degree > 3.5 and
round_Closeness <= 0.3149999976158142 and
round_LofTool > 0.6514999866485596),
0.023083051924208931871 * (round_Degree <= 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_Gene_Length_bp > 2814.5),
0.046139548753990018704 * (round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool > 0.6634999513626099 and
round_Exon_Count > 162.5),
0.0082417292345994958014 * (Group == u'NDNE' and
round_Blomen_KBM7 > -0.5146430730819702 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 2418.21435546875),
0.041904863071896121529 * (round_End > 100419720.0 and
round_dN_dS_Chimp_mi > 0.5 and
round_LofTool <= 0.994350016117096),
-0.058523347259765781669 * (not Group == u'MNC' and
167.33212280273438 < round_StdDev_Transcript_length <=
615.8639526367188),
-0.13397370099236294294 * (not Group == u'MNC' and
round_Tajima__s_D_regulatory > -0.4115000069141388 and
round_Gene_Length_bp <= 9988.0),

-0.0025281542638851562527 * (not Group == u'CM' and
not Group == u'MNC' and
round_Blomen_KBM7 <= -0.5148604512214661 and
round_Transcript_count <= 26.5),
-0.014816630169283761392 * (Group == u'NDNE' and
round_Degree <= 12.5 and
round_missense_Z <= 3.2976694107055664 and
round_Exon_Count <= 223.5),
-0.0016748509383901300038 * (round_Closeness <= 0.3149999976158142 and
round_LofTool <= 0.6634999513626099 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 1968.24755859375),
0.021419335487036086224 * (round_Closeness > 0.2549999952316284 and
round_missense_Z <= 3.334120750427246 and
round_s_het > 0.014458265155553818),
0.053846292322283426102 * (round_Phi > 3.313508932478726e-05 and
3.2976694107055664 < round_missense_Z <= 4.052361965179443),
-0.094256185734778688556 * (not Group == u'CM' and
not Group == u'MNC' and
round_Degree <= 60.5 and
round_missense_Z <= 4.031624794006348),
0.044101606757228337119 * (round_Degree <= 35.5 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7 <= -0.12487166374921799 and
round_missense_Z <= 4.052361965179443),
0.034744638865665013194 * (not Group == u'NDNE' and
round_Degree <= 3.5 and
round_StdDev_Transcript_length > 1049.45654296875),
-0.0015117432304791230541 * (round_End > 127552112.0 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Tajima__s_D_regulatory <= 0.4165000021457672 and
round_missense_Z <= 4.052361965179443),
-0.0043696277889927453292 * (not Group == u'NDNE' and
0.2549999952316284 < round_Closeness <= 0.3149999976158142 and
round_Phi <= 0.8754478693008423),
0.035691787319774168075 * (not Group == u'NDNE' and
round_Closeness <= 0.3149999976158142 and
round_Phi > 0.8768634796142578),
-0.0038248717892900479035 * (not Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi > 0.8768634796142578 and
round_LofTool <= 0.9829000234603882),
-3.0717810457948354157E-07 * (round_Average_Transcript_length),
0.11374755291070159924 * (round_Tajima__s_D_regulatory <=
0.4235000014305115 and
round_Blomen_KBM7 <= -0.2521226406097412 and
round_s_het <= 0.015603477135300636),

-0.0047514279836939361107 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5181520581245422 and
round_Transcript_count <= 26.5 and
round_StdDev_Transcript_length > 169.24288940429688),
-0.012826511723195505726 * (not Group == u'NDNE' and
round_Degree <= 12.5 and
round_StdDev_Transcript_length <= 1418.164794921875),
0.016581885867794556033 * (round_Tajima__s_D_regulatory <=
1.4184999465942383 and
round_Phi > 7.604442998854211e-06 and
round_Gene_Length_bp <= 9942.0),
-0.021957726079143407433 * (not Group == u'NDNE' and
round_Closeness <= 0.2549999952316284 and
round_Phi > 0.4316735863685608),
-0.023063757621387480368 * (Group == u'CNM'),
0.006616251616266622334662 * (round_Degree <= 10.5 and
round_Degree_mi <= 0.5 and
round_Phi > 0.13315337896347046 and
round_LofTool <= 0.9311000108718872),
-0.011550473054156200695 * (round_dN_dS_Chimp > 0.5950000286102295 and
round_missense_Z_mi <= 0.5 and
round_LofTool > 0.04450000077486038),
0.020911928595311275736 * (round_Closeness <= 0.7100000381469727 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z <= 4.031624794006348 and
round_Gene_Length_bp <= 52859.0),
0.0081602582422299618781 * (round_Degree_mi > 0.5 and
round_Average_Transcript_length > 2550.24072265625),
0.00011519157793123054349 * (round_Blomen_KBM7 <= -0.2516775131225586 and
round_missense_Z <= 3.308867931365967 and
round_s_het > 0.015855055302381516),
-0.017088815875378346454 * (round_s_het <= 0.01700003445148468 and
round_Exon_Count <= 86.5),
-0.005247972595815775951 * (round_dN_dS_Chimp > 0.5849999785423279 and
round_Tajima__s_D_regulatory > 0.4165000021457672),
-0.006243045291876287913 9 * (round_Closeness <= 0.3249999807907104 and
round_missense_Z > 2.571293354034424 and
round_Transcript_count <= 10.5),
-0.011579335820359673917 * (round_End <= 100419720.0 and
round_Tajima__s_D_regulatory > -1.2885000705718994 and
round_missense_Z_mi <= 0.5 and
round_StdDev_Transcript_length <= 2093.128662109375),
-0.021952172218480371646 * (round_missense_Z > 4.052361965179443 and
round_Exon_Count <= 77.5),
2.1100283553908253599E-08 * (round_Gene_Length_bp),
0.030085702774348684757 * (round_Closeness > 0.3149999976158142 and
round_Phi > 0.0015477617271244526 and

round_StdDev_Transcript_length <= 626.7237548828125),
-0.004764866731496131 4426 * (round_Phi <= 0.00015248148702085018),
0.0082603409156458418305 * (not Group == u'CM' and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7_mi <= 0.5 and
round_missense_Z <= 4.031624794006348),
-0.073398950654784411718 * (round_End <= 100419720.0 and
0.3050000071525574 < round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295),
0.021089497152677241093 * (round_Closeness > 0.3149999976158142 and
round_Tajima__s_D_regulatory <= 0.5485000014305115 and
round_Blomen_KBM7 <= -0.25219112634658813 and
round_LofTool <= 0.6634999513626099),
0.076308296509059625468 * (round_Closeness <= 0.32499998807907104 and
round_Tajima__s_D_regulatory <= -1.2874999046325684 and
round_Blomen_KBM7 <= -0.41121259331703186),
0.030893043174464097922 * (not Group == u'CM' and
Group == u'MNC' and
round_s_het > 0.028055116534233093),
-0.035827548812507541143 * (round_Degree > 3.5 and
round_Phi <= 0.12447576969861984 and
round_Transcript_count <= 19.5 and
round_StdDev_Transcript_length > 1098.5848388671875),
-0.005052833399117633538 * (Group == u'NDNE' and
round_Degree > 3.5 and
round_Phi > 0.7055177092552185),
0.017529633814622507665 * (round_Degree > 5.5 and
round_Closeness > 0.3149999976158142 and
round_StdDev_Transcript_length > 649.8057861328125),
0.0043005984594983578603 * (round_Tajima__s_D_regulatory >
0.4104999899864197 and
round_Blomen_KBM7 <= -0.5105985403060913 and
round_s_het > 0.014667890034615993 and
round_Exon_Count <= 88.5),
-0.006794559515284487 3166 * (Group == u'NDNE' and
round_Degree <= 35.5 and
round_Gene_Length_bp <= 65653.0 and
round_StdDev_Transcript_length > 590.2589111328125),
-0.003922334003875199 1835 * (round_Closeness <= 0.3149999976158142 and
round_dN_dS_Chimp <= 0.5950000286102295 and
round_Blomen_KBM7 <= -0.25219112634658813 and
round_missense_Z <= 3.2974047660827637),
0.043641103738486321 24 * (-1.2874999046325684 <
round_Tajima__s_D_regulatory <= 1.4165000915527344 and
round_missense_Z > 4.028769493103027 and
round_Transcript_count > 20.5),
-0.004348216994198855 4202 * (Group == u'NDNE' and

round_Tajima__s_D_regulatory <= 0.5570000410079956 and
round_LofTool > 0.8105000257492065),
-0.067478416068065039113 * (not Group == u'MNC' and
round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length <= 987.5745849609375),
0.05856455668737312048 * (Group == u'MNC'),
0.0087343085663817605913 * (round_Tajima__s_D_regulatory >
0.4165000021457672 and
round_LofTool > 0.6634999513626099),
0.028953842017158313432 * (not Group == u'NDNE' and
round_Closeness > 0.2549999952316284 and
round_Phi <= 0.8768634796142578 and
round_LofTool > 0.9872499704360962),
-0.011814802013852038556 * (round_Closeness > 0.3149999976158142 and
round_dN_dS_Chimp_mi <= 0.5 and
round_Average_Transcript_length <= 2037.5650634765625),
0.024333688158167710719 * (round_Closeness > 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_missense_Z <= 3.302974224090576),
-0.034348155117580804474 * (round_Degree_mi <= 0.5 and
round_Closeness <= 0.2549999952316284 and
round_Phi <= 0.8162848949432373),
0.012477065486332037866 * (round_Degree <= 4.5 and
round_missense_Z > 1.43977952003479),
-0.0055992038299544763177 * (round_dN_dS_Chimp_mi <= 0.5 and
0.6634999513626099 < round_LofTool <= 0.9922449588775635 and
round_Exon_Count <= 162.5),
0.012688866869836799844 * (Group == u'END'),
0.051122519055613512007 * (round_Degree > 70.0 and
round_Phi > 0.12447576969861984 and
round_Exon_Count <= 156.5),
0.033172549145610998045 * (not Group == u'MNC' and
round_End <= 100331008.0 and
round_missense_Z <= 3.308867931365967),
0.042438817871726271236 * (round_Blomen_KBM7 <= -0.5158457159996033 and
round_Gene_Length_bp <= 53140.5),
0.018517202101923656982 * (round_End > 100419720.0 and
round_missense_Z > 3.99676229370117),
0.0097170358540753076076 * (not Group == u'NDNE' and
round_StdDev_Transcript_length > 590.2589111328125 and
round_Exon_Count > 253.5),
0.041889875414339285131 * (Group == u'NDNE' and
round_Degree <= 3.5 and
round_Phi > 0.14934945106506348),
0.1159141151153197935 * (round_dN_dS_Chimp_mi),
-0.036397079913846719368 * (round_Degree <= 3.5 and

round_Phi <= 0.12447576969861984 and
round_Transcript_count <= 19.5 and
round_StdDev_Transcript_length > 1098.5848388671875),
-0.040646373044763282889 * (round_Closeness <= 0.2549999952316284 and
round_missense_Z <= 3.2976694107055664 and
round_StdDev_Transcript_length > 987.5745849609375),
0.0034491772056603540834 * (not Group == u'NDNE' and
round_Degree <= 62.5 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length > 580.6193237304688),
0.010753897831233117169 * (round_Closeness > 0.3149999976158142 and
round_Blomen_KBM7 <= -0.5146373510360718),
-0.0029093503719793034797 * (round_dN_dS_Chimp <= 0.5950000286102295 and
round_missense_Z <= 3.2976694107055664 and
round_LofTool <= 0.9922449588775635),
0.0085107244663394625989 * (round_s_het_mi),
-0.016375026470318413546 * (not Group == u'CM' and
not Group == u'MNC' and
round_End <= 121391536.0 and
round_Blomen_KBM7 <= -0.49977701902389526),
0.006494039635649984099 1 * (not Group == u'NDNE' and
round_Closeness <= 0.32499998807907104 and
round_missense_Z > 1.43977952003479 and
round_Gene_Length_bp > 47610.5),
0.1894264397753082918 * (round_Degree <= 4.5 and
round_LofTool > 0.9868500232696533 and
round_StdDev_Transcript_length > 1049.45654296875),
-0.023331859677601131386 * (not Group == u'MNC' and
round_Degree <= 35.5 and
round_Tajima__s_D_regulatory > -1.2855000495910645 and
round_Gene_Length_bp <= 315852.5),
0.029766225515022122494 * (round_dN_dS_Chimp <= 0.5950000286102295 and
round_Gene_Length_bp > 343581.5),
0.16369767760948211732 * (Group == u'NDNE' and
round_Degree <= 4.5 and
round_StdDev_Transcript_length <= 714.263427734375),
-0.024442885761271002099 * (Group == u'NDNE' and
round_Degree_mi <= 0.5 and
round_StdDev_Transcript_length <= 661.5775146484375),
-0.0042331491676081359904 * (round_Closeness <= 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_LofTool > 0.8825000524520874 and
round_Gene_Length_bp <= 54522.0),
0.0376150466960103666 * (round_Tajima__s_D_regulatory <=
0.4235000014305115 and
round_LofTool > 0.9922800064086914),
0.030185637919566413873 * (not Group == u'NDNE' and

round_Degree_mi <= 0.5 and
round_Closeness <= 0.3149999976158142 and
round_Average_Transcript_length > 2570.535888671875),
0.007569280695528393838 9 * (round_Closeness <= 0.3149999976158142 and
round_Phi > 0.12447576969861984 and
round_LofTool <= 0.9910449981689453 and
round_Gene_Length_bp > 54522.0),
0.093970991187771663045 * (round_Closeness),
0.005168232707957641329 5 * (round_Blomen_KBM7 > -0.5153281092643738 and
0.49326902627944946 < round_missense_Z <= 4.031624794006348
and
round_Gene_Length_bp <= 53047.5),
-0.000636820449720149647 61 * (not Group == u'CM' and
not Group == u'MNC' and
round_End > 121391536.0 and
round_Blomen_KBM7 <= -0.49977701902389526),
0.13059023194145646 94 * (not Group == u'NDNE' and
round_StdDev_Transcript_length > 636.7213134765625 and
round_Exon_Count > 253.5),
-0.028105123991677912615 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5146373510360718 and
round_LofTool <= 0.9922449588775635 and
round_Transcript_count > 20.5),
-0.071733384292753485378 * (round_dN_dS_Chimp_mi <= 0.5 and
round_Blomen_KBM7 > -0.5146373510360718 and
round_Transcript_count <= 20.5 and
round_StdDev_Transcript_length <= 2388.5126953125),
-0.001494487018499959449 2 * (round_Closeness > 0.3149999976158142 and
round_missense_Z <= 2.571293354034424 and
round_StdDev_Transcript_length > 1450.0491943359375),
-0.037305572747207686735 * (round_Degree <= 12.5 and
round_missense_Z > 4.052361965179443 and
round_Gene_Length_bp > 2814.5 and
round_Exon_Count <= 413.0),
-0.013677846268730356125 * (round_Closeness <= 0.3449999988079071 and
3.2976694107055664 < round_missense_Z <= 4.052361965179443 and
round_Average_Transcript_length <= 2037.5650634765625),
-0.011008296654006112167 * (round_Degree <= 3.5 and
round_Phi > 0.0009046811610460281 and
round_LofTool <= 0.6634999513626099),
-0.066509812215818378545 * (round_Degree <= 12.5 and
round_dN_dS_Chimp_mi <= 0.5 and
0.5915000438690186 < round_LofTool <= 0.6634999513626099),
0.043755311507365633739 * (round_dN_dS_Chimp_mi <= 0.5 and
round_missense_Z <= 3.2976694107055664 and
round_Average_Transcript_length > 2056.5712890625 and
round_Exon_Count <= 240.5),

0.01113235474873714749 * (round_Degree <= 60.5 and
round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool <= 0.6634999513626099 and
round_Transcript_count > 20.5),
0.029791550463155375139 * (round_End > 100419720.0 and
round_dN_dS_Chimp_mi <= 0.5 and
round_LofTool <= 0.994350016117096),
-0.0065287331094041282237 * (round_Degree <= 60.0 and
round_Phi <= 0.12447576969861984 and
round_Exon_Count > 83.5),
-0.0045345154906442954046 * (round_Tajima__s_D_regulatory > -
1.2874999046325684 and
round_missense_Z <= 3.2976694107055664 and
round_missense_Z_mi <= 0.5 and
round_LofTool > 0.6634999513626099),
-0.0042090971907015710396 * (round_Closeness <= 0.2549999952316284 and
round_Phi > 0.0015477617271244526 and
round_StdDev_Transcript_length <= 2427.245361328125 and
round_Exon_Count <= 96.5),
0.006482760392506849676 8 * (round_Degree <= 10.5 and
round_LofTool <= 0.659500002861023 and
round_Gene_Length_bp > 37942.5),
0.090064277946272239261 * (round_Degree > 3.5 and
round_Closeness <= 0.3149999976158142 and
round_Gene_Length_bp <= 41501.0),
-0.016228231105145558139 * (round_Degree_mi > 0.5 and
round_StdDev_Transcript_length <= 613.3181762695312),
-0.024541945681890484088 * (round_s_het <= 0.025150161236524582 and
round_Transcript_count <= 13.5 and
round_Average_Transcript_length > 1873.0999755859375),
0.088655920423249531814 * (round_End <= 110010712.0 and
round_Degree <= 61.5 and
round_Degree_mi <= 0.5 and
round_missense_Z <= 4.052361965179443),
0.033084801199255609028 * (Group == u'NDNE' and
round_Degree <= 12.5 and
round_s_het > 0.02518850564956665),
0.018352753415038773 61 * (round_Degree > 12.5 and
round_Degree_mi <= 0.5 and
round_Closeness > 0.3149999976158142 and
round_StdDev_Transcript_length > 1285.35791015625),
-0.041044206485432332965 * (Group == u'NDNE' and
round_Degree <= 4.5 and
round_StdDev_Transcript_length <= 1049.45654296875),
-0.039696369932485681131 * (round_Phi > 0.998741626739502 and
round_Gene_Length_bp > 9906.5 and
round_StdDev_Transcript_length > 166.57864379882812 and

```
                        round_Average_Transcript_length > 1956.6083984375),
        -0.012496735141249921616 * (round_dN_dS_Chimp <= 0.5950000286102295 and
                        round_missense_Z > 4.052361965179443 and
                        round_Transcript_count <= 10.5 and
                        round_Average_Transcript_length <= 6597.4873046875),
         0.039674947245171322818 * (not Group == u'NDNE' and
                        round_Closeness <= 0.3149999976158142 and
                        round_Phi > 0.12447576969861984 and
                        round_Gene_Length_bp <= 45328.0),
         -0.30567366547201163529 * (round_Degree <= 4.5 and
                        round_Degree_mi <= 0.5 and
                        round_StdDev_Transcript_length > 1049.45654296875)    ])

def get_type_conversion():
   return {
      u'Tajima\'s D regulatory': {'convert_func': parse_nonstandard_na, 'convert_args':
None},
      u'Degree': {'convert_func': parse_nonstandard_na, 'convert_args': None},
      u'dN/dS Chimp': {'convert_func': parse_nonstandard_na, 'convert_args': None},
      u'Closeness': {'convert_func': parse_nonstandard_na, 'convert_args': None},}
INDICATOR_COLS = [u'Blomen KBM7', u'Degree', u'LofTool', u'Phi', u'Tajima\'s D regulatory',
u'dN/dS Chimp', u'missense_Z', u's_het']

IMPUTE_VALUES = {
   u'Average Transcript length': 1932.128571,
   u'Blomen KBM7': -0.499215,
   u'Closeness': 0.250000,
   u'Degree': 3.000000,
   u'End': 58139967.000000,
   u'Exon Count': 36.000000,
   u'Gene Length bp': 28439.000000,
   u'LofTool': 0.502000,
   u'Phi': 0.002401,
   u'StdDev Transcript length': 899.036969,
   u'Tajima\'s D regulatory': 0.143000,
   u'Transcript count': 6.000000,
   u'dN/dS Chimp': 0.230000,
   u'missense_Z': 0.493769,
   u's_het': 0.017813,}


def bag_of_words(text):
   """ set of whole words  in a block of text """
   if type(text) == float:
      return set()

   return set(word.lower() for word in
```

```
            re.findall(r'\w+', text, re.UNICODE | re.IGNORECASE))


def parse_date(x, date_format):
    """ convert date strings to numeric values. """
    try:
        # float values no longer pass isinstance(x, np.float64)
        if isinstance(x, (np.float64, float)):
            x = long_type(x)
        if '%f' in date_format and date_format.startswith('v2'):
            temp = str(x)
            if re.search('[\+-][0-9]+$', temp):
                temp = re.sub('[\+-][0-9]+$', '', temp)

            date_format = date_format[2:]
            dt = datetime.strptime(temp, date_format)
            sec = calendar.timegm(dt.timetuple())
            return sec * 1000 + dt.microsecond // 1000
        elif '%M' in date_format:
            temp = str(x)
            if re.search('[\+-][0-9]+$', temp):
                temp = re.sub('[\+-][0-9]+$', '', temp)

            return calendar.timegm(datetime.strptime(temp, date_format).timetuple())
        else:
            return datetime.strptime(str(x), date_format).toordinal()
    except:
        return float('nan')


def parse_percentage(s):
    """ remove percent sign so percentage variables can be converted to numeric """
    if isinstance(s, float):
        return s
    if isinstance(s, int):
        return float(s)
    try:
        return float(s.replace('%', ''))
    except:
        return float('nan')

def parse_nonstandard_na(s):
    """ if a column contains numbers and a unique non-numeric,
        then the non-numeric is considered to be N/A
    """
    try:
        ret = float(s)
```

```
        if np.isinf(ret):
            return float('nan')
        return ret
    except:
        return float('nan')

def parse_length(s):
    """ convert feet and inches as string to inches as numeric """
    try:
        if "'" in s and '"' in s:
            sp = s.split('"')
            return float(sp[0]) * 12 + float(sp[1].replace("'", ''))
        else:
            if '"' in s:
                return float(s.replace('"', '')) * 12
            else:
                return float(s.replace("'", ''))
    except:
        return float('nan')

def parse_currency(s):
    """ strip currency characters and commas from currency columns """
    if not isinstance(s, text_type):
        return float('nan')
    s = re.sub(u'[\$\u20AC\u00A3\uFFE1\u00A5\uFFE5]|(EUR)', '', s)
    s = s.replace(',', '')
    try:
        return float(s)
    except:
        return float('nan')


def parse_currency_replace_cents_period(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(",", "")
        val = float(val)
    except ValueError:
        val = float('nan')
```

```
    return val


def parse_currency_replace_cents_comma(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(".", "")
        val = val.replace(",", ".")
        val = float(val)
    except ValueError:
        val = float('nan')
    return val


def parse_currency_replace_no_cents(val, currency_symbol):
    try:
        if np.isnan(val):
            return val
    except TypeError:
        pass
    if not isinstance(val, string_types):
        raise ValueError('Found wrong value for currency: {}'.format(val))
    try:
        val = val.replace(currency_symbol, "", 1)
        val = val.replace(" ", "")
        val = val.replace(",", "")
        val = val.replace(".", "")
        val = float(val)
    except ValueError:
        val = float('nan')
    return val

def parse_numeric_types(ds):
    """ convert strings with numeric types (date, currency, etc.)
        to actual numeric values """
    TYPE_CONVERSION = get_type_conversion()
    for col in ds.columns:
        if col in TYPE_CONVERSION:
            convert_func = TYPE_CONVERSION[col]['convert_func']
```

```
            convert_args = TYPE_CONVERSION[col]['convert_args']
            ds[col] = ds[col].apply(convert_func, args=convert_args)
    return ds

def sanitize_name(name):
    safe = name.strip().replace("-", "_").replace("$", "_").replace(".", "_")
    safe = safe.replace("{", "_").replace("}", "_")
    safe = safe.replace('"', '_')
    safe = safe.replace("\n", "_")
    safe = safe.replace("\r", "_")
    return safe

def rename_columns(ds):
    new_names = {}
    existing_names = set()
    blank_index = 0
    for old_col in ds.columns:
        col = sanitize_name(old_col)
        if col == '':
            col = 'Unnamed: %d' % blank_index
            blank_index += 1
        if col in existing_names:
            raise ValueError('Duplication detected. Column with name=['
                    + old_col + '] was preprocessed to['
                    + col + '] that already exists')
        existing_names.add(col)
        new_names[old_col] = col
    ds.rename(columns=new_names, inplace=True)
    return ds

def add_missing_indicators(ds):
    for col in INDICATOR_COLS:
        ds[col + '-mi'] = ds[col].isnull().astype(int)
    return ds

def impute_values(ds):
    for col in ds:
        if col in IMPUTE_VALUES:
            ds.loc[ds[col].isnull(), col] = IMPUTE_VALUES[col]
    return ds

BIG_LEVELS = {
    u'Group': [
        u'CM',
        u'CNM',
        u'END',
        u'MNC',
```

```
      u'NDNE',
   ],
}


SMALL_NULLS = {
   u'Group': 1,
}


VAR_TYPES = {
   u'Average Transcript length': 'N',
   u'Blomen KBM7': 'N',
   u'Closeness': 'N',
   u'Degree': 'N',
   u'End': 'N',
   u'Exon Count': 'N',
   u'Gene Length bp': 'N',
   u'Group': 'C',
   u'LofTool': 'N',
   u'Phi': 'N',
   u'StdDev Transcript length': 'N',
   u'Tajima\'s D regulatory': 'N',
   u'Transcript count': 'N',
   u'dN/dS Chimp': 'N',
   u'missense_Z': 'N',
   u's_het': 'N',
}


def combine_small_levels(ds):
   for col in ds:
      if BIG_LEVELS.get(col, None) is not None:
         mask = np.logical_and(~ds[col].isin(BIG_LEVELS[col]), ds[col].notnull())
         if np.any(mask):
            ds.loc[mask, col] = 'small_count'
      if SMALL_NULLS.get(col):
         mask = ds[col].isnull()
         if np.any(mask):
            ds.loc[mask, col] = 'small_count'
      if VAR_TYPES.get(col) == 'C' or VAR_TYPES.get(col) == 'T':
         mask = ds[col].isnull()
         if np.any(mask):
            if ds[col].dtype == float:
               ds[col] = ds[col].astype(object)
            ds.loc[mask, col] = 'nan'
   return ds
```

```
# N/A strings in addition to the ones used by Pandas read_csv()
NA_VALUES = ['null', 'na', 'n/a', '#N/A', 'N/A', '?', '.', '', 'Inf', 'INF', 'inf', '-inf', '-Inf', '-INF', ' ',
'None', 'NaN', '-nan', 'NULL', 'NA', '-1.#IND', '1.#IND', '-1.#QNAN', '1.#QNAN', '#NA', '#N/A
N/A', '-NaN', 'nan']

# True/False strings in addition to the ones used by Pandas read_csv()
TRUE_VALUES = ['TRUE', 'True', 'true']
FALSE_VALUES = ['FALSE', 'False', 'false']

DEFAULT_ENCODING = 'utf8'

REQUIRED_COLUMNS = [u"Average Transcript length",u"Blomen
KBM7",u"Closeness",u"Degree",u"End",u"Exon Count",u"Gene Length
bp",u"Group",u"LofTool",u"Phi",u"StdDev Transcript length",u"Tajima's D
regulatory",u"Transcript count",u"dN/dS Chimp",u"missense_Z",u"s_het"]


def validate_columns(column_list):
   if set(REQUIRED_COLUMNS) <= set(column_list):
      return True
   else :
      raise ValueError("Required columns missing: %s" %
                  (set(REQUIRED_COLUMNS) - set(column_list)))

def convert_bool(ds):
   TYPE_CONVERSION = get_type_conversion()
   for col in ds.columns:
      if VAR_TYPES.get(col) == 'C' and ds[col].dtype in (int, float):
         mask = ds[col].notnull()
         ds[col] = ds[col].astype(object)
         ds.loc[mask, col] = ds.loc[mask, col].astype(text_type)
      elif VAR_TYPES.get(col) == 'N' and ds[col].dtype == bool:
         ds[col] = ds[col].astype(float)
      elif ds[col].dtype == bool:
         ds[col] = ds[col].astype(text_type)
      elif ds[col].dtype == object:
         if VAR_TYPES.get(col) == 'N' and col not in TYPE_CONVERSION:
            mask = ds[col].apply(lambda x: x in TRUE_VALUES)
            if np.any(mask):
               ds.loc[mask, col] = 1
            mask = ds[col].apply(lambda x: x in FALSE_VALUES)
            if np.any(mask):
               ds.loc[mask, col] = 0
            ds[col] = ds[col].astype(float)
         elif TYPE_CONVERSION.get(col) is None:
            mask = ds[col].notnull()
```

```
            ds.loc[mask, col] = ds.loc[mask, col].astype(text_type)
    return ds

def get_dtypes():
    return {a: object for a, b in VAR_TYPES.items() if b == 'C'}

def predict_dataframe(ds):
    return ds.apply(predict, axis=1)

def run_dataframe(ds):
    ds = rename_columns(ds)
    ds = convert_bool(ds)
    validate_columns(ds.columns)
    ds = parse_numeric_types(ds)
    ds = add_missing_indicators(ds)
    ds = impute_values(ds)
    ds = combine_small_levels(ds)
    prediction = 1/(1 + np.exp(-predict_dataframe(ds)))
    return prediction


def run(dataset_path, output_path, encoding=None):
    if encoding is None:
        encoding = DEFAULT_ENCODING

    ds = pd.read_csv(dataset_path, na_values=NA_VALUES, low_memory=False,
                dtype=get_dtypes(), encoding=encoding)

    prediction = run_dataframe(ds)
    prediction_file = output_path
    prediction.name = 'Prediction'
    prediction.to_csv(prediction_file, header=True, index_label='Index')


def _construct_parser():
    import argparse

    parser = argparse.ArgumentParser(description='Make offline predictions with DataRobot
Prime')

    parser.add_argument(
        '--encoding',
        type=str,
        help=('the encoding of the dataset you are going to make predictions with. '
            'DataRobot Prime defaults to UTF-8 if not otherwise specified. See the '
            '"Codecs" column of the Python-supported standards chart '
            '(https://docs.python.org/2/library/codecs.html#standard-encodings) '
```

```
            'for possible alternative entries.'),
        metavar='<encoding>'
    )
    parser.add_argument(
        'input_path',
        type=str,
        help=('a .csv file (your dataset); columns must correspond to the '
            'feature set used to generate the DataRobot Prime model.'),
        metavar='<data_file>'
    )
    parser.add_argument(
        'output_path',
        type=str,
        help='the filename where DataRobot writes the results.',
        metavar='<output_file>'
    )

    return parser


def _parse_command(args):
    parser = _construct_parser()
    parsed_args = parser.parse_args(args[1:])

    if parsed_args.encoding is None:
        sys.stderr.write('Warning: For input data encodings other than UTF-8, '
                         'search "Prime examples" in the DataRobot Users Guide at
https://app.datarobot.com/docs/users-guide/index.html')
        parsed_args.encoding = DEFAULT_ENCODING

    return parsed_args


if __name__ == '__main__':
    args = _parse_command(sys.argv)
    run(args.input_path, args.output_path, encoding=args.encoding)
```