Identifying Functionally Important Missense Mutations in Cancer by Dynamics-Based Analysis and Predicting Pathogenicity/Disease Category of Missense Mutations

by

Jan Fehmi Sayılgan

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Computational Science and Engineering

KOÇ ÜNİVERSİTESİ

Februrary 10, 2021

Identifying Functionally Important Missense Mutations in Cancer by Dynamics-Based Analysis and Predicting Pathogenicity/Disease Category of Missense Mutations

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Jan Fehmi Sayılgan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Mehmet Gönen (Advisor)

_____

Prof. Attila Gürsoy

_____

Prof. Füsun Can

_____

Prof. Türkan Haliloğlu

_____

Assoc. Prof. Nurcan Tunçbağ

Date: _____

To my dearest mother

ABSTRACT

Identifying Functionally Important Missense Mutations in Cancer by Dynamics-Based Analysis and Predicting Pathogenicity/Disease Category of Missense Mutations

Jan Fehmi Sayılgan

Doctor of Philosophy in Computational Science and Engineering

Februrary 10, 2021

Missense mutations have various effects on protein structures, also leading to distorted protein dynamics that plausibly affects the function. We hypothesized that missense mutations in cancer-related genes selectively target hinge-neighboring residues that orchestrate collective structural dynamics. To test our hypothesis, we selected 69 cancer-related genes from the Cancer Gene Census (CGC) database and their representative protein structures from the Protein Data Bank. We first identified the hinge residues in two global modes of motion by applying the Gaussian Network Model. We then showed that missense mutations are significantly enriched on hinge-neighboring residues in oncogenes and tumor suppressor genes. We observed that several oncogenes (e.g., MAP2K1, PTPN11, and KRAS) and tumor suppressor genes (e.g., EZH2, CDKN2C, and RHOA) strongly exhibit this phenomenon. Next, we developed a computational pipeline to detect significantly enriched three-dimensional (3D) clustering of missense mutations around hinge residues by using the the Cancer Genome Atlas (TCGA) dataset. The hinge residues were also detected by applying a Gaussian network model for the modes 1 to 5. By systematically analyzing the Pan-Cancer compendium of somatic missense mutations in nearly 10,000 tumors from TCGA, we identified candidate genes and mutations in addition to well known ones. For instance, we found significantly enriched 3D clustering of missense mutations in known cancer genes including CDK4, CDKN2A, TCL1A, and MAPK1. Besides these known genes, we also identified significantly enriched 3D clustering of missense mutations around hinge residues in PLA2G4A, which may lead to excessive phosphorylation of the extracellular signal-regulated kinases. Our results show that the consideration of clustering around hinge residues can help us explain the functional

role of the mutations in known cancer genes and identify candidate genes. Furthermore, we proposed new features, named hinge-based, for pathogenicity prediction for missense mutations and show that hinge-based features improve pathogenicity prediction. Pathogenicity prediction of human missense variants remains a challenging problem. Existing computational models are basically binary classifiers predicting whether given missense variants are deleterious or neutral. We demonstrated a multilabel classification method that predicts not only the pathogenicity but also the disease category type of given missense variants. Moreover, existing computational models are based on sequence-, structural-, or protein dynamics-based analysis. We also showed that network topological properties of proteins significantly improve determining the pathogenicity of missense variants. We trained and tested our model PathDis with 20,361 missense variants. Then, we benchmarked by the area under the ROC curve (AUROC) evaluation metric score with a well-established prediction model which uses the same dataset. We observed that our model PathDis improves AUROC by 3%. Then, we tested PathDis with a different dataset. Also benchmarking based on this different dataset against other well-established prediction models demonstrated that PathDis' AUROC score is approximately 3% higher than the second highest AUROC score. In addition to high pathogenicity prediction results, PathDis has approximately 79% accuracy for predicting the disease category types (i.e., No Disease, Cancer, and Non-Cancer). Besides introducing a prediction model, we also characterized the missense variants by our sequence-, structure-, dynamics-, and network-based features for the disease category types. We observed that sequence-based, network-based, and structure-/dynamics-based features characterize No Disease, Cancer, and Non-Cancer missense variants, respectively.

ÖZETÇE

Kanserde Fonksiyonel Öneme Sahip Yanlış Anlam Mutasyonların Dinamik Tabanlı Analizle Belirlenmesi ve Yanlış Anlam Mutasyonların Patojenite/Hastalık Kategorisinin Tahmin Edilmesi

Jan Fehmi Sayılgan

Hesaplamalı Bilimler ve Mühendislik, Doktora

10 Şubat 2021

Yanlış anlam mutasyonlarının protein yapıları üzerinde çeşitli etkilerinin yanı sıra protein dinamiğini etkileyerek de protein işlev bozukluklarına yol açabilir. Kanserle ilgili genlerdeki yanlış anlam mutasyonların kolektif yapısal dinamiklerini düzenleyen menteşe-komşu rezidülarını seçici olarak hedef aldığını varsaydık. Hipotezimizi test etmek için, Cancer Gene Census (CGC) veritabanından kanserle ilgili 69 geni ve Protein Data Bank'tan temsilci protein yapılarını seçtik. İlk olarak, Gauss Ağ Modelini uygulayarak iki küresel hareket modunda menteşe rezidülarını tanımladık. Daha sonra, yanlış anlam mutasyonlarının, onkojenlerdeki ve tümör baskılayıcı genlerdeki menteşe-komşu rezidülar üzerinde önemli ölçüde zenginleştiğini gösterdik. Birkaç onkojenin (ör., MAP2K1, PTPN11 ve KRAS) ve tümör baskılayıcı genlerin (ör., EZH2, CDKN2C ve RHOA) bu fenomeni güçlü bir şekilde sergilediğini gözlemledik. Daha sonra, the Cancer Genome Atlas (TCGA) veri setini kullanarak menteşe rezidülarının etrafındaki yanlış anlam mutasyonların önemli ölçüde zenginleştirilmiş üç boyutlu (3B) kümelenmesini tespit etmek için hesaplamalı bir işlem hattı geliştirdik. Menteşe rezidüları, 1 ve 5 arası modlar için bir Gauss Ağı Modeli uygulanarak tespit edildi. TCGA'dan yaklaşık 10.000 tümördeki somatik yanlış anlam mutasyonların sistematik olarak analiz ederek, iyi bilinenlere ek olarak aday sürücü genleri ve mutasyonları belirledik. Örneğin, CDK4, CDKN2A, TCL1A ve MAPK1 dahil olmak üzere bilinen kanser genlerinde yanlış anlam mutasyonların önemli ölçüde zenginleştirilmiş 3B kümelenmesini bulduk. Bilinen bu sürücü genlerin yanı sıra, PLA2G4A'daki menteşe rezidülarının etrafındaki yanlış anlam mutasyonlarının önemli ölçüde zenginleştirilmiş 3B kümelenmesini de belirledik. Bu kümelenme hücre dışı sinyalle düzenlenen kinazların aşırı fosforilasy-

onuna yol açabilir. Sonuçlarımız, menteşe rezidüları etrafında kümelenmenin, bilinen kanser genlerindeki mutasyonların işlevsel rolünü açıklamamıza ve aday sürücü genleri/mutasyonları belirlememize yardımcı olabileceğini göstermektedir. Ayrıca, yanlış anlam mutasyonları için patojenite tahmini için menteşe tabanlı olarak adlandırdığımız yeni öznitelikler önerdik ve menteşe tabanlı özniteliklerin patojenite tahminini ilerlettiğini gösterdik. İnsan yanlış anlam varyantlarının patojenite tahmini zorlu bir problem olmaya devam etmektedir. Mevcut hesaplama modelleri, temelde, verilen yanlış anlam varyantların zararlı veya nötr olup olmadığını tahmin eden ikili sınıflandırıcılardır. Sadece patojeniteyi değil, aynı zamanda verilen yanlış anlam varyantlarının hastalık kategori tipini de öngören çok etiketli bir sınıflandırma yöntemi gösterdik. Dahası, mevcut hesaplama modelleri sekans, yapısal veya protein dinamiğine dayalı analize dayanmaktadır. Ayrıca proteinlerin ağ topolojik özelliklerinin, yanlış anlam varyantların patojenikliğini belirlemeyi önemli ölçüde iyileştirdiğini de gösterdik. PathDis modelimizi 20.361 yanlış anlam varyantla eğitip test ettik. Daha sonra, aynı veri setini kullanan başka bir tahminleme modeliyle ROC eğrisi altındaki ortalama alan (AUROC) metriğine göre kıyasladık. PathDis modelimizin AUROC'u %3 oranında iyileştirdiğini gözlemledik. Ardından PathDis'i farklı bir veri kümesiyle test ettik. Ayrıca, bu farklı veri kümesine dayalı olarak literatürdeki başka tahminleme modelleriyle kıyaslayarak PathDis'in AUROC değerinin ikinci en yüksek AUROC değerinden yaklaşık %3 daha yüksek olduğunu gözlemledik. PathDis, yüksek patojenite tahmin sonuçlarına ek olarak, hastalık kategorisi türlerini (yani, Hastalık Yok, Kanser ve Kanser Olmayan) tahmin etmede yaklaşık %79 doğruluğa sahiptir. Bir tahmin modeli sunmanın yanı sıra, yanlış anlam varyantları, hastalık kategorisi türleri için sekans, yapı, dinamik ve ağ tabanlı özniteliklerimizle de karakterize ettik. Sekans tabanlı, ağ tabanlı ve yapı/dinamik tabanlı özelliklerin sırasıyla Hastalık Yok, Kanser ve Kanser Olmayan yanlış anlam varyantlarını karakterize ettiğini gözlemledik.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Assoc. Prof. Mehmet Gönen for his splendid scientific guidance. I am also deeply grateful to Prof. Türkan Haliloğlu for her excellent scientific collaboration. I am enormously thankful to Assist. Prof. Michael Lawrence for sharing his awesome scientific passion. I would like to express my speical gratitude to Ayda Şentürk for her wonderful friendship and talks on writing skills. Also, I would like to thank to my other thesis committee members, Prof. Attila Gürsoy, Prof. Füsun Can, and Assoc. Prof. Nurcan Tunçbağ for their interest and questions. I also thank to my dear friends at Koç and at the Machine Intelligence and Data Analysis in Science (Midas) Lab. Finally, I thank my mother Karol Ann for her never ending worlds worth support.

TABLE OF CONTENTS

List of Tables **xi**

List of Figures **xii**

Abbreviations **xiv**

Chapter 1: Introduction **1**

Chapter 2: Materials and Methods **6**

2.1 Data Preprocessing 6

2.2 Annotating the Missense Mutations on PDB Structures 8

2.3 Locating the Hinge Residues 9

2.4 Analyzing the Missense Mutations Near Hinge Residues 11

2.5 Analyzing the Missense Mutations Cluster Around Hinge Residues . . 12

2.6 Features and Disease Category Types 12

2.7 Computational Tools Used in the Prediction Model 14

Chapter 3: Protein Dynamics Analysis Reveals That Missense Mutations in Cancer-Related Genes Appear Frequently on Hinge-Neighboring Residues **15**

3.1 Analysis of Oncogenes and Tumor Suppressor Genes 15

3.2 Evaluation of Cancer-Related Genes 20

Chapter 4: Protein Dynamics Analysis Identifies Candidate Cancer Driver Genes and Mutations in TCGA Data **24**

4.1 Identification of Hinge-Mutation Clusters 24

4.2 Cases of Proteins with Hinge-Mutation Clusters 26

4.2.1 CDK4 26
4.2.2 p16 INK4a 27
4.2.3 TCL1A 27
4.2.4 ERK2 29
4.2.5 PLA2G4A 30
4.3 Hinge-Based Features in Pathogenicity Prediction 32

Chapter 5: Predicting the Pathogenicity and the Disease Category of Missense Variants 37

5.1 Performance Evaluation of PathDis 37
5.2 Characterizing Missense Variants by NET and SEQ Features for Disease Category Types 42
5.3 Characterizing Missense Variants by STR and DYN Features for Disease Category Types 46
5.4 A Test Case with De Novo PTEN Missense Variants 47

Chapter 6: Conclusion 49

Bibliography 54

Appendix A: 71

LIST OF TABLES

3.1 Distribution of hinge-neighboring mutations. 17
3.2 List of genes with statistically significantly enriched hinge-neighboring mutations. 18
3.3 Distribution of hinge-neighboring mutations for Mode 1 and 1&2. 20

4.1 List of genes with statistically significantly enriched hinge-mutation clusters. 25

LIST OF FIGURES

2.1 Schematic workflow for the characterization of the relationship between missense mutations and hinge residues on protein structures. 9

3.1 The distribution of hinge-neighboring mutations for all proteins, oncogenes, and TSGs. 16
3.2 The percentages of the number of hinges relative to the gene length. . 19
3.3 The ratios between the numbers of residues covered by all hinge sites and the total numbers of residues at four different distance thresholds. 21
3.4 Computational analysis for the hinge and mutated residues of TCL1A. 22
3.5 The hinge and mutated residues of EZH2. 23

4.1 Volcano plots reporting p-values against fold changes. 26
4.2 Significantly mutated genes and cancer types by hinge-mutation clusters in TCGA. 28
4.3 Examples of hinge-mutation clusters in 1D and 3D for (A) TCL1A and (B) PLA2G4A. 29
4.4 The evaluation metrics of Rhapsody, Rhapsody plus HingeM1-M2, and Rhapsody plus HingeM1-M5 models. 32
4.5 The Spearman's correlation scores of the features 33
4.6 The hinge-mutation clusters identified using the Rhapsody dataset . . 34
4.7 The evaluation metrics of HingeM1-M2, HingeM1-M5 models. 35

5.1 Performance evaluation of PathDis in pathogenicity prediction. 38
5.2 Features of PathDis. 40

5.3 Performance evaluation of PathDis in disease category prediction. . . 41
5.4 Distributions of all NET and two SEQ features. 44
5.5 Distributions of one STR and DYN features. 45
5.6 Evaluation of pathogenicity predictions for de novo PTEN missense variants. 48

ABBREVIATIONS

CGC	The Cancer Gene Census
TCGA	The Cancer Genome Atlas
COSMIC	The Catalogue of Somatic Mutations in Cancer
GNM	Gaussian Network Model
ANM	Anisotropic Network Model
PDB	Protein Data Bank
TSG	Tumor Suppressor Gene
AUROC	Area Under the Receiver Operating Characteristics Curve
ROC	The Receiver Operating Characteristics Curve
FDR	False Discovery Rate
ATP	Adenosine triphosphate
GTP	Guanosine-5'-triphosphate
DNA	Deoxyribonucleic acid
BLOSUM62	Blocks Substitution Matrix For Midrange Related Proteins
SASA	Solvent Accessible Surface Area
wtPSIC	Wild Type Position-Specific Independent Counts
DeltaPSIC	Delta Position-Specific Independent Counts
MSF	Mean Square Fluctuations
SEQ	Sequence-Based Features
STR	Structure-Based Features
DYN	Dynamics-Based Features
NET	Network-Based Features
HNSC	Head and Neck Squamous Cell Carcinoma
COAD	Colon Adenocarcinoma
GBM	Glioblastoma Multiforme

Chapter 1

INTRODUCTION

Functional properties of a protein are not only determined by its structure, but also by its dynamics. Any molecular change caused by missense mutations might affect protein dynamics. Our knowledge on the effects of missense mutations on the development and progression of certain diseases such as cancer remains incomplete. An examination of the relationship between somatic missense mutations and cancer-related proteins from the view of protein dynamics is necessary.

Protein dynamics ranges from local to global fluctuations on many different time scales. Collective global modes of motions are evolutionary conserved structural motions essential for a protein's functional activity. In that respect, our work is based on the premise that somatic missense mutations likely appear as global perturbations to the protein's intrinsic dynamics [Haliloglu and Bahar, 2015].

Protein dynamics have two main global motions: hinge and shear [Gerstein et al., 1994]. Shear motions are a series of small sliding movements amongst closely packed domains. In contrast, hinge motions are rotating/bending movements of rigid parts, including domains of the protein which coordinate their large scale global dynamic behavior. An illustrative example for hinge motions is the opening and closing movements of a door. The hinges of the door maintain opening and closing movements according to the hinge axis they form. As a broken hinge of a door might hinder opening or closing, a missense mutation on a hinge-neighboring residue might drastically distort the protein dynamics, therefore the protein function. Hence, detecting hinge residues becomes crucial in understanding hinge movements. Many methods have been developed to detect hinge residues [Emekli et al., 2008, Flores et al., 2008, Kundu et al., 2004, Shatsky et al., 2004]. Here, we applied a method

based on the Gaussian Network Model (GNM) [Emekli et al., 2008, Bahar et al., 1997b, Haliloglu et al., 1998]. After detecting hinge residues, we statistically studied the distribution of somatic missense mutations on hinge-neighboring residues in cancer-related proteins.

Previous studies have identified many driver mutations using sequence-based or structure-based approaches. Sequence-based approaches fail in identifying some driver mutations since they ignore close spatial proximity in three-dimensional (3D) structures that may be very distant in linear sequence [Ryslik et al., 2016], but contributed greatly by generating evolutionary based features. One recent study revealed that missense mutations form clusters in various regions of the structure including protein interaction interfaces [Kamburov et al., 2015]. The interaction interfaces of oncoproteins and tumor suppressors were also significantly enriched by cancer somatic missense mutations [Engin et al., 2016]. Likewise, mutations can also be enriched on functional sites. For example, ATP- and GTP-binding sites in oncoproteins significantly harbor missense mutations [Stehr et al., 2011]. DNA-binding proteins were demonstrated to be enriched by missense mutations. Interestingly, these variants were rarely located in the proteins' DNA-interacting interfaces [Gress et al., 2017].Moreover, It was shown that deleterious missense variants tend to be located at singlet hot spots at protein–protein interaction interfaces [Ozdemir et al., 2018]. These studies contributed enormously in understanding and detecting many driver missense mutations, however, they were not able to capture all of the drivers.

Furthermore, several studies have illustrated the importance of protein dynamics descriptors in determining driver missense mutations [Ponzoni and Bahar, 2018, Kumar et al., 2019, Ponzoni et al., 2020]. Our first aim is to show whether known cancer-related proteins are enriched by missense mutations selectively targeting hinge-neighboring residues. In Chapter3, we demonstrate that missense mutations in cancer-related genes appear frequently around hinge residues by utilizing the Catalogue of Somatic Mutations in Cancer (COSMIC) collection [Forbes et al., 2017].

Another important database, The Cancer Genome Atlas (TCGA) project, has generated comprehensive catalogs of somatic alterations by sequencing the exomes

or genomes of many human tumors [Cancer Genome Atlas Research Network et al., 2013]. The vast majority of these somatic alterations have no functional impact on tumor initiation or progression and are termed passenger mutations. In contrast, mutations providing a selective advantage to cancer cells are termed driver mutations [Vogelstein et al., 2013]. One of the main aims of TCGA has been to provide data for discriminating driver mutations from a large number of passenger mutations [Bailey et al., 2018, The ICGC/TCGA Pan-Cancer Analysis and of Whole Genomes Consortium, 2020]. In the direction of this aim, we developed similarly again a protein dynamics-based computational pipeline to detect driver mutations that may disrupt wild type protein dynamics in the TCGA dataset.

Our work in Chapter 4 extends Chapter 3 by devising a similar protein dynamics approach, but also considering the hinge residues from the next three slowest modes besides the two most global motions, using the TCGA PanCancer data set and incorporating protein interaction residue information to exclude the missense mutations near to these residues, hence, analyzing only the missense mutations clustering around hinge residues. In Chapter 4, we report significant 3D clusters of missense mutations around hinge residues in protein structures, which can assist in explaining the functional role of the mutations in known cancer genes and in identifying candidate cancer genes. Out of 1,969 genes, our pan-cancer analysis identified 54 unique genes with at least one statistically significant cluster around hinge residues. Our analysis does not only identify known driver mutations but also candidate driver mutations in known cancer genes such as MAPK1 Y316C, which are not listed as oncogenic in the OncoKB database [Chakravarty et al., 2017], as well as unidentified candidate driver genes such as PLA2G4A, which are not listed in the Cancer Gene Census (CGC) [Futreal et al., 2004]. After observing highly statistically significant clustering, we proposed hinge-based features for pathogenicity prediction. Furthermore, we demonstrate that hinge-based features improves pathogenicity prediction for missense variants.

Several machine learning methods have been developed for predicting pathogenicity of missense variants [Niroula and Vihinen, 2016]. Our work in Chapter 5 presents

a supervised machine learning method not only for predicting the pathogenicity of given missense variants but also their associated diseases. Our approach is to perceive the pathogenicity prediction challenge as a multilabel problem in which each missense variant is labeled with its disease category and its pathogenicity. By doing so, we were also able to compare and contrast disease category types (i.e., No Disease, Cancer, and Non-Cancer) from the perspective of missense variants.

According to the Human Gene Mutation Database, germline missense variants are the most common disease-associated mutation type in human inherited disorders [Stenson et al., 2017]. These inherited disorders can be monogenic disorders such as achondroplasia or more frequently observed polygenic disorders such as orofacial cleft. Pathogenic germline variants were also identified in cancers such as breast cancer [Golubeva et al., 2019]. Besides inherited variants, especially somatic missense variants play major roles in cancer (Vogelstein et al., 2013). As a result, germline and somatic missense variants are associated with various complex diseases.

Other mutation types like nonsense or duplication have clearer impacts on proteins, however missense variants may have effects ranging from no impact to drastic impacts on protein function. A missense variant distant to active sites of a protein can have severe effects on the protein's thermostability or enzymatic activity [Freeman et al., 2011]. A missense variant near hinge residues which disrupts shaker K channels activity can be neutralized by another missense variant nearby [Magidovich and Yifrach, 2004a]. Hotspot missense mutations in cancer can be neutral [Hess et al., 2019]. Therefore, it is difficult identifying which missense variant is neutral or deleterious.

In pathogenicity prediction, some methods are only sequence-based methods [Choi et al., 2012, Hopf et al., 2017, Ng and Henikoff, 2003, Rentzsch et al., 2019], some are sequence- and structure-based [Adzhubei et al., 2010, Capriotti and Altman, 2011, Carter et al., 2009, Gray et al., 2018, Pejaver et al., 2020, Tokheim and Karchin, 2019a], a recent method is also dynamics-based [Ponzoni et al., 2020]. In Chapter 5, we present our model PathDis, which is a sequence-, structure-, dynamics- and network-based multilabel random forest classification method. We

included network-based features such as Hub, Bottleneck, Degree, Closeness, and Betweenness, which describe the protein's network topological properties. Network topological properties are crucial in determining essential proteins [Yu et al., 2007].

Briefly in our thesis, we show that missense mutations near hinge residues are enriched in cancer-related proteins from the Cancer Gene Census, then we further identify missense mutation clusters around hinge residues using the TCGA dataset. So, we were also able to propose candidate driver cancer genes and candidate driver missense mutations. Next, we propose hinge-based features for predicting whether a given missense mutation is deleterious or neutral. We demonstrate that these features indeed improve pathogenicity prediction. Following this, we developed our pathogenicity prediction model. Besides using hinge-based features, we input network-based features, and besides predicting pathogenicity, our model also predicts disease category types.

Chapter 2

MATERIALS AND METHODS

2.1 Data Preprocessing

The Catalogue of Somatic Mutations in Cancer (COSMIC) [Forbes et al., 2017] database contains thousands of somatic mutations related to human cancer. One of the sub-projects of COSMIC is the Cancer Gene Census (CGC) [Futreal et al., 2004] that has 671 genes catalogued as cancer-related according to the latest release date, 13th February 2018. We downloaded the Cancer Gene Census file with 671 cancer genes. Next, we cross-mapped the cancer genes to their representative PDB structures using SIFTS [Velankar et al., 2013]. After downloading the PDB structures available in the Protein Data Bank (PDB) [Berman et al., 2000] at http://www.rcsb.org, we filtered out the ones accommodating modified residues and drugs. The drug list was provided by the PDBSum [Laskowski et al., 1997] database. Then, we checked the protein sequence coverage of these PDB structures and excluded the structures with less than 70% coverage. We calculated the coverage by downloading the gene sequences from the UNIPROT database [The UniProt Consortium, 2017] and performing sequence alignment with MUSCLE [Edgar, 2004] using the R package bio3d [Grant et al., 2006]. Consequently, we had 70 PDB structures in our final list. Each of them corresponds to one distinct protein. Furthermore, we chose a single chain of the structures to represent the corresponding protein.

We downloaded the All Mutations in Census Genes file (https://cancer.sanger.ac.uk/cosmic/download, accessed January 15, 2018) with the mutation information of all of the 671 cancer genes in COSMIC. We extracted the missense mutations of the 70 genes included in our list. Since we did not detect any hinge residues for

MTCP1, we excluded MTCP1. Finally, for 69 genes, we obtained 87,243 missense mutations which originated from 77,472 tumor samples of 39 cancer types.

We downloaded the entire set of missense mutations from the release version v17.0 of the TCGA collection (https://portal.gdc.cancer.gov). This consisted of 540,440 missense mutations in 6,164 genes of 9,710 samples from 33 cancer types. TCGA contains all detected genomic alterations, therefore, the missense mutations are a mixture of driver and passenger mutations. Next, we cross-mapped the genes to their representative PDB structures using SIFTS [Velankar et al., 2012] descriptors, which is a project of the Protein Data Bank (PDB) in Europe. We proceeded only with those genes whose protein structures are deposited in the PDB [Berman et al., 2000]. We additionally filtered out structures with less than 70% coverage of the gene sequence and structures with modified residues and drugs using the drug list from PDBSum [Laskowski et al., 1997]. The coverage was calculated by performing sequence alignment between the gene sequence and the PDB sequence using MUSCLE [Edgar, 2004] with the R package "bio3d" [Grant et al., 2006]. The single chain of the PDB structure with the highest coverage was selected as the representative chain for the gene. Finally, we obtained 106,951 missense mutations in 1,969 genes of 8,648 samples from 33 cancer types. Then, we annotated the missense mutations on the PDB structures. In some cases, the residue numbers of the missense mutations and the residue numbers of the PDB structure does not exactly match, so the sequence alignment between gene sequence and PDB sequence was used to match these residue numbers.

The dataset we used for pathogenicity prediction is the Rhapsody dataset which consists of 13,626 deleterious and 6,735 neutral missense variants [Ponzoni et al., 2020]. This dataset was obtained by first gathering these publicly available datasets: HumVar [Adzhubei et al., 2010], ExoVar [Li et al., 2013a], PredictSNP [Bendl et al., 2014], VariBench [Thusberg et al., 2011], SwissVar [Mottaz et al., 2010], ClinVar [Landrum et al., 2016] and Humsavar [The UniProt Consortium, 2017]. Second, these datasets were filtered according to whether a missense variant with at least one ClinVar review star can be annotated to a Protein Data Bank (PDB) structure.

Finally, the Rhapsody dataset has 20,361 missense variants on 2,422 proteins.

To test our model PathDis with a different dataset, we used 3,230 missense variants for the tumor suppressor protein PTEN which were able to be PDB annotated and had an abundance class label (Table A.7.). The pathogenicity of PTEN variants was measured by Variant Abundance by Massively Parallel Sequencing (VAMP-seq) [Matreyek et al., 2018]. The abundance classes 'low' and 'possibly low' were relabeled as deleterious, and the abundance classes 'wt-like' and 'possibly wt-like' were relabeled as neutral. We retrained PathDis without any PTEN variants before testing for the de novo PTEN missense variants. In addition, we used the Rhapsody webserver (http://rhapsody.csb.pitt.edu/index.php), the Polyphen2 HumVar webserver (http://genetics.bwh.harvard.edu/pph2/), the Provean webserver (http://provean.jcvi.org/index.php) and the MutPred2 webserver (http://mutpred2.mutdb.org) for testing the de novo PTEN missense variants for each model, respectively.

2.2 Annotating the Missense Mutations on PDB Structures

At this point, we mapped the missense mutations on the PDB structures, since the residue numbers of the missense mutations and the residue numbers of the PDB structure might not exactly match (see Figure 2.1). Similar to calculating the protein sequence coverage, we performed sequence alignment with MUSCLE using the R package bio3d. We also excluded 2,386 missense mutations that were not mapped to any PDB structure after the calculations.

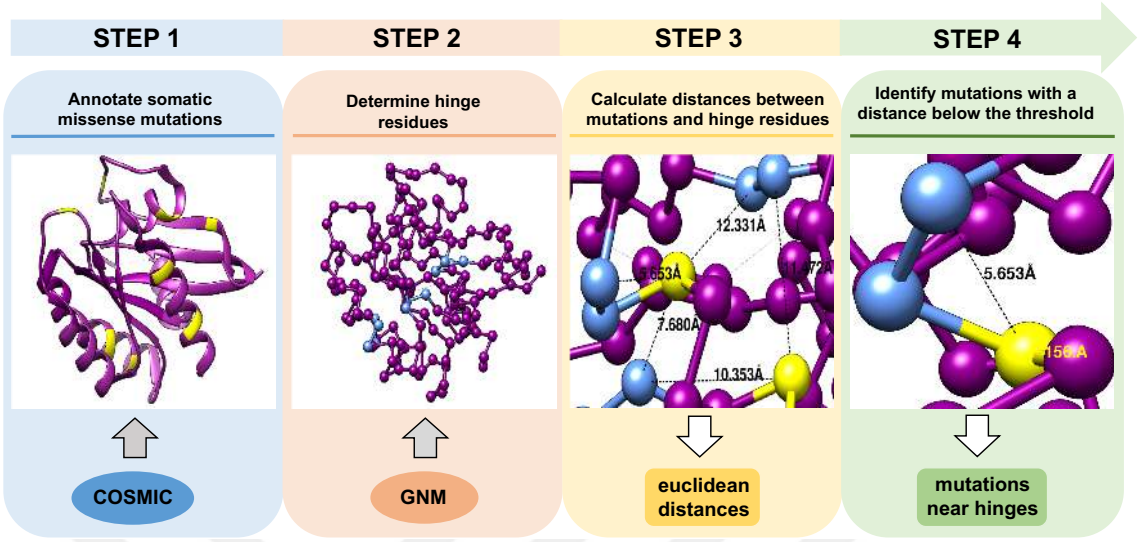

Figure 2.1: Schematic workflow for the characterization of the relationship between missense mutations and hinge residues on protein structures.

2.3 Locating the Hinge Residues

To determine the hinge residues in a PDB structure, we applied the Gaussian Network Model (GNM) [Bahar et al., 1997b, Haliloglu et al., 1998]. The GNM is a coarse-grained, elastic mass-and-spring network model which characterizes the dynamics of the biological macromolecule. In the GNM, the nodes represent the α-carbon atoms of the residues. The model assumes that α-carbon atoms are in contact when they are separated less than a cut-off distance, R_c. We decided to set R_c as 7.3 Å, which was demonstrated as the optimal cut-off distance according to the agreement between the experimental β-factors and of those calculated by the GNM method [Li et al., 2017].

The GNM method finally delivers cross-correlations between each residue of a given mode for a PDB structure. The cross-correlations indicate which structural segments move in the same direction, i.e., the positions where the structures flex to define dynamic domains. To calculate the cross-correlations, we first began with the

definition of the connectivity matrix:

$$\Gamma = \begin{cases} -1, & \text{if } i \neq j \text{ and } R_{ij} \leq R_c, \\ 0, & \text{if } i \neq j \text{ and } R_{ij} > R_c, \\ -\Sigma_{i,i\neq j}\Gamma_{ij}, & \text{if } i = j, \end{cases}$$

where R_{ij} denotes the distance between the i^{th} and j^{th} α-carbon atoms. The cross-correlations between residue fluctuations are then found by

$$\langle \Delta\mathbf{R_i} \cdot \Delta\mathbf{R_j} \rangle = \frac{3k_BT}{\gamma}[\mathbf{\Gamma}^{-1}]_{ij}$$

where k_B is the Boltzmann constant, T is the absolute temperature, and γ is the force constant of the Hookean pairwise potential, and $\Delta\mathbf{R_i}$ denotes the fluctuation in the position vector of residue i. Since the determinant of $\mathbf{\Gamma}$ is zero, the pseudo-inverse $\mathbf{\Gamma}^{-1}$ is calculated by the eigenvalue decomposition $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^T$, where $\mathbf{U}$ is an orthogonal matrix whose columns, $\mathbf{u_k}$, are the eigenvectors of $\mathbf{\Gamma}$, and $\mathbf{\Delta}$ is the diagonal matrix of the eigenvalues, λ_k, of $\mathbf{\Gamma}$.

By substituting, we can rewrite the cross-correlation as

$$\begin{aligned} \langle \Delta\mathbf{R_i} \cdot \Delta\mathbf{R_j} \rangle &= \frac{3k_BT}{\gamma}[\mathbf{U}\mathbf{\Delta}^{-1}\mathbf{U}^T]_{ij} \\ &= \frac{3k_BT}{\gamma}\sum_{k=1}^{N-1}\lambda_k^{-1}[\mathbf{u_k}\mathbf{u_k^T}]_{ij} \end{aligned}$$

The elements of the k^{th} eigenvector $\mathbf{u_k}$ represent the distribution of residue fluctuations along the k^{th} mode axis, and the corresponding eigenvalue λ_k scales with the frequency of the mode. We focused on the first two modes which capture the most cooperative global movements of the protein structure [Emekli et al., 2008, Bahar et al., 1997a]. It is debatable which slow modes or how many slow modes are functionally the most informative for detecting hinges. We detected the hinge residues for the slow modes 1, 2, 3, 4, and 5, which characterize a dynamic spectrum from the global to more intermediate modes of motions.

In a given mode, a positive cross-correlation value between two nearby residues means that these two residues move in the same direction, while a negative cross-correlation value implies that they do not. We computed a symmetric cross-correlation

matrix in which its components are the cross-correlations between each pair of residues in a mode for a given structure. A typical cross-correlation matrix represents rigid regions of residues with positive or negative signs. A region of residues with a positive sign indicates that all the residues in the region move together but not with the residues of the regions with a negative sign, and vice versa. Moreover, flexible residues are present at the borders of positive and negative rigid regions. We defined the two flexible residues at each border as hinge residues. In addition, we merged the regions that comprise of less than 15 residues with the neighboring rigid regions [Emekli et al., 2008].

2.4 Analyzing the Missense Mutations Near Hinge Residues

We measured the Euclidean distances between the missense mutations and hinge residues of each PDB structure in our final list. The distances are calculated between the coordinates of the α-carbon atom of the mutated residue and the coordinates of the α-carbon atom of the hinge residue. We classified the missense mutations with a distance less than the cut-off distance to one of the hinge residues as hinge-neighboring mutations. Further, we calculated the percentages of hinge-neighboring mutations. To control whether the distribution of hinge-neighboring mutations was statistically significant or not, we carried out a bootstrap analysis by randomly resampling the mutation positions with replacement 10,000 times. We computed the p-values by dividing the number of bootstraps that scored higher frequencies of hinge-neighboring mutations than the value generated using the original COSMIC data by 10,000. Then, p-values were corrected for multiple testing using the Benjamini-Hochberg method [Benjamini and Hochberg, 1995] as implemented in the *p.adjust* function of R (htttps://www.r-project.org/). Also, our p-value threshold for statistical significance is 0.05.

2.5 Analyzing the Missense Mutations Cluster Around Hinge Residues

For the identifying hinge-mutation cluster, first we again measured the Euclidean distances between the coordinates of the α-carbon atom of the missense mutations and the coordinates of the α-carbon atom of the hinge residues for each PDB structure. Hence, we discovered the missense mutations that spatially clustered with a distance less than the cut-off distance 7.3 Å around the hinge residues. We termed such clusters as hinge-mutation clusters if there are three or more missense mutations. Then, we filtered out the missense mutations that are near known interaction residues with less than 7.3 Å to exclusively work with missense mutations that only cluster around hinge residues, since a missense mutation may be near both hinge residues and interaction residues. Both cut-off distances were set to 7.3 Å to be consistent with the cut-off used in GNM calculations. The Protein–Protein, Protein–DNA/RNA, Protein–Ligand, and Protein–Metal interaction residue information of the PDB structures were downloaded from the PDBSum database. We assessed the significance of the hinge-neighboring mutations by using a null model assuming a uniform distribution of mutations on the given PDB structure. The null distribution was obtained through 10,000 resampling with replacement of missense mutations by preserving the number of missense mutations on each PDB. The resulting p-values were corrected for multiple testing using the Benjamini-Hochberg method [Benjamini and Hochberg, 1995] by the "p.adjust" function of R. The false discovery rate cut-off for statistical significance was set to 0.05. UCSF Chimera software was used for structure visualization [Pettersen et al., 2004].

2.6 Features and Disease Category Types

In our prediction model, we have four groups of features: sequence-based (SEQ), structure-based (STR), dynamics-based (DYN) and network-based (NET) (see Table A.8. for the Rhapsody dataset and all corresponding features). Some of the SEQ, STR, DYN features were available in the Rhapsody dataset. These available features are wtPSIC and DeltaPSIC which are the Position Specific Independent

Counts (PSIC) scores [Sunyaev et al., 1999] of the wild type amino acid PSIC score, and the difference of wild type amino acid PSIC score and the mutant amino acid PSIC score respectively, BLOSUM62 amino acid substitution scores [Henikoff and Henikoff, 2000], Solvent Accessible Surface Area (SASA) [Touw et al., 2015], and the DYN features Mean Square Fluctuations (MSF), Effectiveness, Sensitivity, and Stiffness [Bakan et al., 2011]. We included the NET features Hub, Bottleneck, Degree, Closeness, and Betweenness obtained from TargetMine [Chen et al., 2019]. We also added the secondary structures (i.e. isHelix, isBeta, isLoop) [Frishman and Argos, 1995], the PASE [Li et al., 2013b] that is the Euclidean distance between the amino acid substitution by using seven physicochemical properties (i.e. transfer of free energy from octanol to water, normalized van der Waals volume, isoelectric point, polarity, normalized frequency of turn, normalized frequency of alpha-helix, free energy of solution in water) of amino acids in the AAindex [Kawashima et al., 2008]. As final features, we included the DYN features called hinge-based features (i.e. HingeM1, HingeM2, HingeM3, HingeM4, HingeM5) which are the minimum Euclidean distances between the missense variants and hinge residues of each mode on the corresponding PDB structures. We detected the hinge residues by applying the Gaussian Network Model and using modes 1 to 5 which characterize a spectrum from the global to intermediate modes of protein dynamics [Haliloglu et al., 1997, Emekli et al., 2008]. For annotating each missense variant with its associated disease type, we utilized databases Humsavar [The UniProt Consortium, 2017], ClinVar [Landrum et al., 2016], CanProVar [Li et al., 2010], and DoCM [Ainscough et al., 2016]. 951 disease types were annotated to 10,282 deleterious missense variants. We categorized these disease types into two categories: 'Cancer', and 'Non-Cancer'. Less than 1% of the dataset entries had an overlap between Cancer and Non-Cancer category types. We assigned these entries as Cancer disease category type because it has much lesser variants. 6,735 neutral variants were labeled as 'No Disease'. The 3,344 missense variants of the deleterious variants were not able to be annotated using the databases.

2.7 Computational Tools Used in the Prediction Model

The 3,344 missense variants with no annotated disease category were labeled by implementing the positive-unlabeled machine learning framework [Denis et al., 2005]. We trained our positive data (i.e. Cancer and Non-Cancer data) with randomly chosen 80% of the unlabeled data, and tested with the 20% of the unlabeled data in 1,000 replications with randomForest R package version 4.6.14 [Liaw and Wiener, 2002]. Next, we took the average of the predicted probabilities of each unlabeled missense variant to assign either to Cancer or Non-Cancer category type. As a result, we have 1,856 Cancer and 11,700 Non-Cancer variants. For obtaining our pathogenicity and disease category predictions, we applied utiml R package version 0.1.6.9001 [Rivolli and de Carvalho, 2018] utilizing the multilabel random forest method. For calculating the evaluation metric and feature importance scores, we used caret R package version 6.0.86 [Kuhn, 2008] and ROCR R package version 1.0.11 [Sing et al., 2005], and pROC R package version 1.16.2 [Robin et al., 2011]. We set the number of trees in the forest to 1,500 and the maximum number of features hyperparameters to 2 after 5-fold cross-validation based on the AUROC metric. In addition, we used mice R package version 3.12.0 [van Buuren and Groothuis-Oudshoorn, 2011] utilizing predictive mean matching method for imputation of the missing NET features for 7,349 missense variants. We applied randomForest R package also for the Rhapsody model since they applied random forest method with their optimized hyperparameters 1,500 and 2 for the number of trees and the maximum number of features, respectively [Ponzoni et al., 2020].

Chapter 3

PROTEIN DYNAMICS ANALYSIS REVEALS THAT MISSENSE MUTATIONS IN CANCER-RELATED GENES APPEAR FREQUENTLY ON HINGE-NEIGHBORING RESIDUES

3.1 Analysis of Oncogenes and Tumor Suppressor Genes

To analyze the distribution of missense mutations in cancer-related genes on protein structures, we selected 69 out of 671 cancer proteins listed in the CGC. First, we annotated 87,243 missense mutations from COSMIC on the representative structures from PDB of the selected proteins. We then identified hinge residues of two global modes and calculated the distances between the α-carbon atoms of the hinge residues and of the missense mutations.

Our aim was to investigate the distribution of missense mutations over the 69 protein structures. Therefore, we set four distance thresholds (i.e., 0.0 Å, 5.0 Å, 7.3 Å, and 10.0 Å) to compute the percentage of mutations on hinge-neighboring residues (see Figure 3.1 and Table 3.1). If the distance between the α-carbon atoms of a mutated residue and the α-carbon atoms of a hinge residue was smaller than or equal to the distance threshold, we classified this mutation as related to hinge movements of the protein and termed it as a hinge-neighboring mutation.

Among the 69 genes, 26, 24, and 19 of them were identified as oncogenes, tumor suppressor genes (TSGs), and unknown, respectively [Futreal et al., 2004]. We examined the missense mutation distributions on hinge-neighboring residues for TSGs, oncogenes, and all of the 69 genes separately. For the TSGs, the enrichment on hinge-neighboring residues were statistically significant (i.e., p-value < 0.05) for all of the cut-off thresholds (see Figure 3.1 and Table 3.1). However, for the oncogenes

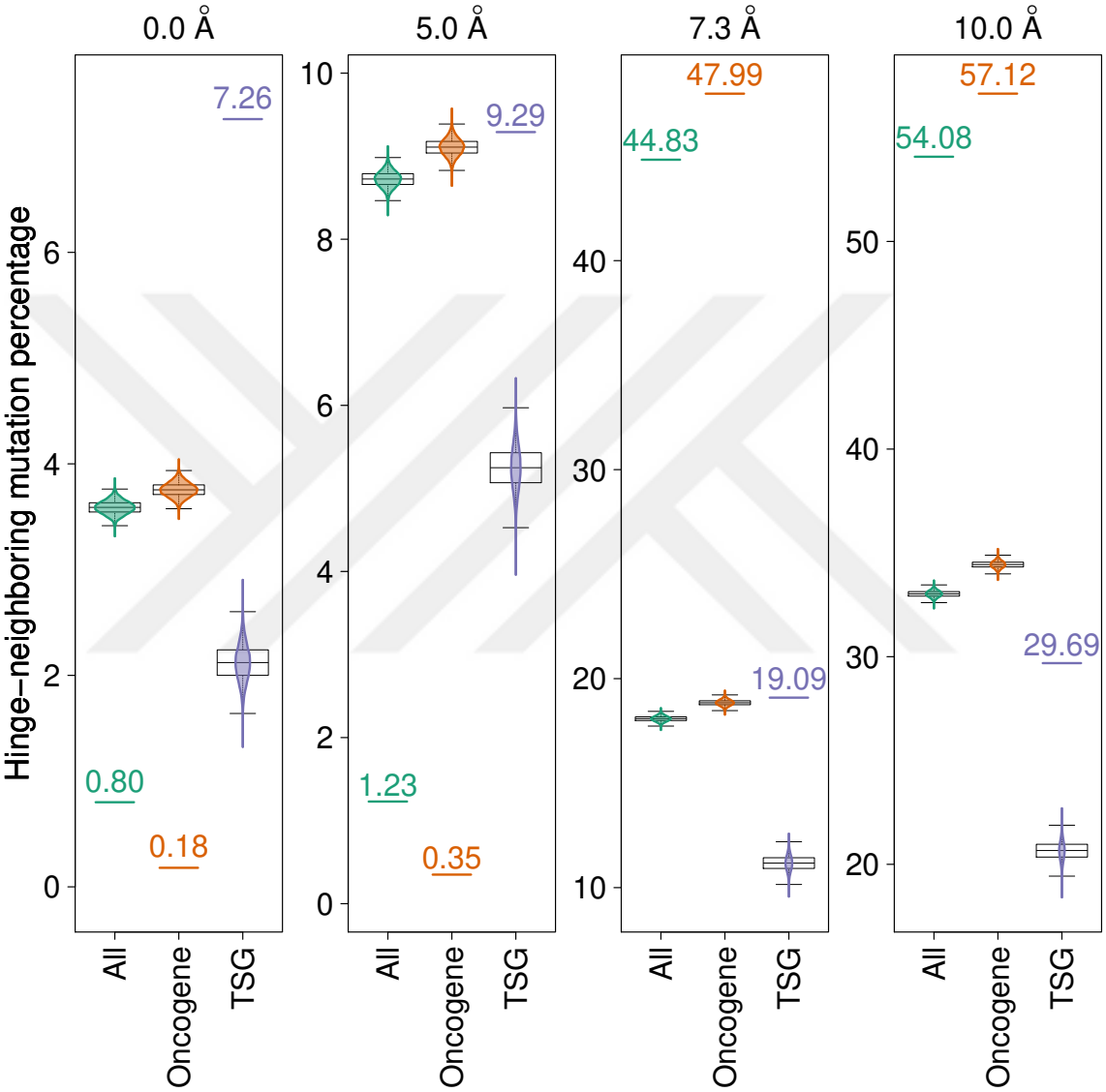

Figure 3.1: The distribution of hinge-neighboring mutations for all proteins, oncogenes, and TSGs. The dashed lines and violin plots illustrate the results from the COSMIC data and bootstrap analysis, respectively, for the cut-off thresholds 0.0 Å, 5.0 Å, 7.3 Å, and 10.0 Å.

the enrichment was statistically significant only for the cut-off thresholds 7.3 Å and 10.0 Å (see Figure 3.1 and Table 3.1).

The difference between the results for TSGs and oncogenes might be due to that the oncogenic missense mutations tend to be at locations where they can effect both hinge-neighboring residues and functional/binding/interaction sites. Moreover, for all of the 69 genes the distribution expresses a pattern similar to oncogenes in Figure 3.1. This pattern might be related to the number of missense mutations on PDB structures of oncogenes and TSGs, which are 75,902 and 6,649, respectively. This considerable difference also confirms that missense mutations largely target oncogenes, while truncating mutations target TSGs [Vogelstein et al., 2013].

Table 3.1: Distribution of hinge-neighboring mutations.

Gene group	**COSMIC results**	**Bootstrap results (mean ± standard deviation)**	**Cut-off threshold**
All	0.80%	3.59% ± 0.063%	0.0 Å
All	1.23%	8.73% ± 0.096%	5.0 Å
All	44.83%	18.08% ± 0.130%	7.3 Å
All	54.08%	33.03% ± 0.156%	10.0 Å
Oncogene	0.18%	3.76% ± 0.068%	0.0 Å
Oncogene	0.35%	9.11% ± 0.104%	5.0 Å
Oncogene	47.99%	18.84% ± 0.140%	7.3 Å
Oncogene	57.12%	34.44% ± 0.167%	10.0 Å
TSG	7.26%	2.12% ± 0.176%	0.0 Å
TSG	9.29%	5.25% ± 0.268%	5.0 Å
TSG	19.09%	11.17% ± 0.370%	7.3 Å
TSG	29.69%	20.66% ± 0.467%	10.0 Å

In addition, distribution of hinge-neighboring mutations according to modes have large difference. The proportion of hinge-neighboring mutations on all proteins was 28-fold higher for the first and second modes together for the cut-off threshold 7.3 Å, even though the number of hinge residues increased only by two-fold from the first mode to the second mode (Table 3.3).

Table 3.2: List of genes with statistically significantly enriched hinge-neighboring mutations.

Gene	**Role**	**Cut-off threshold**	**Hinge-neighboring mutation percentage**	***p-value***
EZH2	TSG	0.0 Å	49.87%	< 1e-4
TCL1A	Oncogene	0.0 Å	20.00%	0.0124
EED	Unknown	0.0 Å	9.09%	0.0211
CDKN2C	TSG	0.0 Å	7.14%	0.0427
MAP2K1	Oncogene	0.0 Å	6.74%	0.0028
EZH2	TSG	5.0 Å	51.53%	< 1e-4
IL2	Unknown	5.0 Å	32.26%	0.0209
TCL1A	Oncogene	5.0 Å	32.00%	0.0314
HLA-A	Unknown	5.0 Å	9.29%	0.0081
KRAS	Oncogene	7.3 Å	84.39%	< 1e-4
RHOA	TSG	7.3 Å	66.38%	< 1e-4
EZH2	TSG	7.3 Å	55.22%	< 1e-4
ALDH2	Unknown	7.3 Å	21.78%	0.0264
MAP2K4	TSG	7.3 Å	20.61%	0.0016
PTPN11	Oncogene	7.3 Å	19.81%	< 1e-4
HLA-A	Unknown	7.3 Å	18.19%	< 1e-4
POLG	Unknown	7.3 Å	13.89%	0.0281
KRAS	Oncogene	10.0 Å	98.78%	< 1e-4
B2M	TSG	10.0 Å	89.87%	< 1e-4
RHOA	TSG	10.0 Å	80.94%	< 1e-4
PTPN11	Oncogene	10.0 Å	63.46%	< 1e-4
EZH2	TSG	10.0 Å	59.67%	< 1e-4
NT5C2	Oncogene	10.0 Å	53.73%	< 1e-4
MAP2K1	Oncogene	10.0 Å	41.84%	0.0001
CDKN2C	TSG	10.0 Å	39.29%	0.0222
CDKN2A	TSG	10.0 Å	34.54%	0.0021
ALDH2	Unknown	10.0 Å	32.67%	0.0402
STAG2	TSG	10.0 Å	10.53%	0.0469

Hinge-neighboring mutation percentage gives the percentage of mutations near hinges for the COSMIC data.

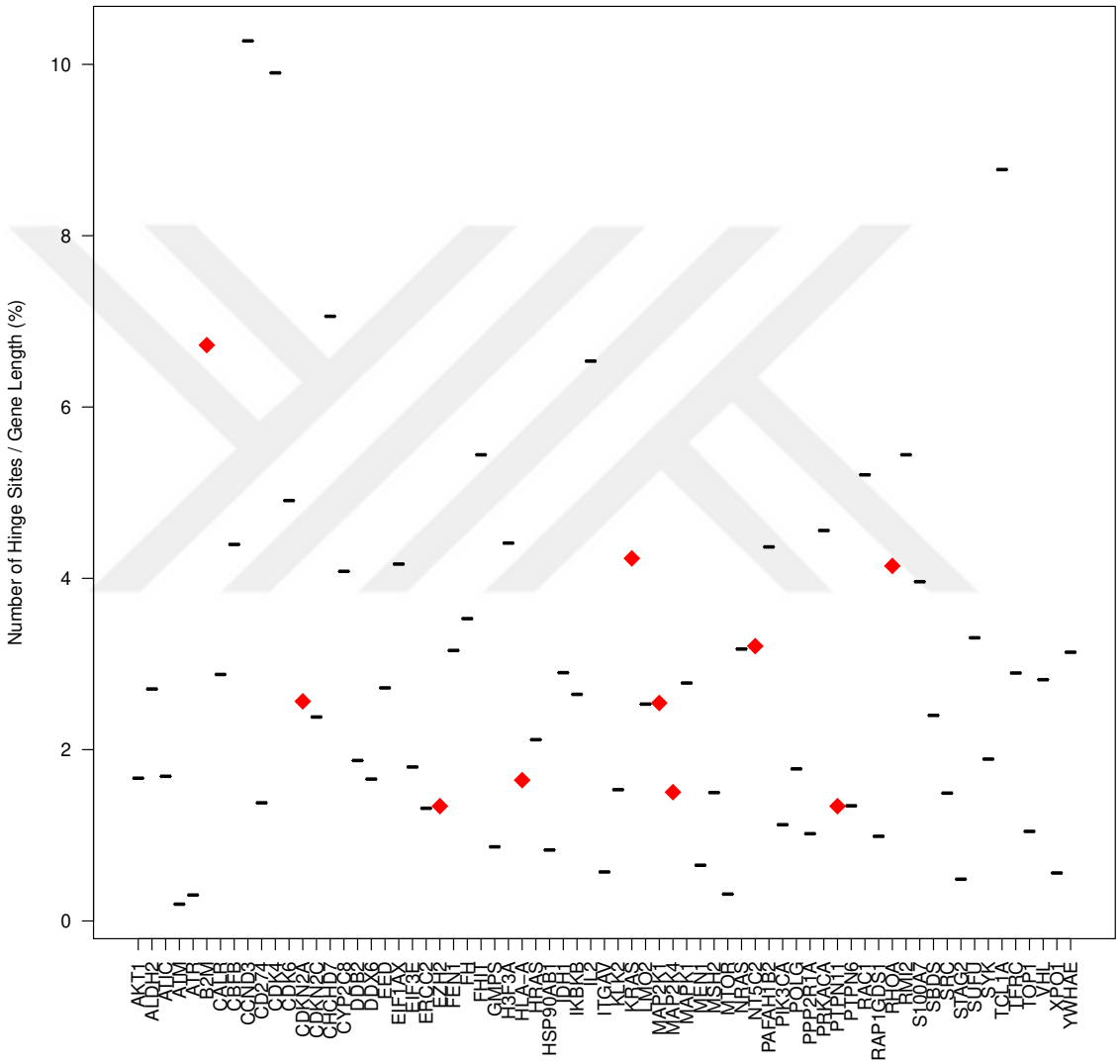

Figure 3.2: The percentages of the number of hinges relative to the gene length. The gene names are listed on the x-axis. Red diamond shapes indicate the statistical significant proteins.

Table 3.3: Distribution of hinge-neighboring mutations for Mode 1 and 1&2.

Mode	Gene group	COSMIC results	Cut-off threshold
1	All	0.61%	0.0 Å
1	All	0.83%	5.0 Å
1	All	1.59%	7.3 Å
1	All	10.18%	10.0 Å
1&2	All	0.80%	0.0 Å
1&2	All	1.23%	5.0 Å
1&2	All	44.83%	7.3 Å
1&2	All	54.02%	10.0 Å

3.2 Evaluation of Cancer-Related Genes

We further analyzed the distributions of missense mutations on structures for each gene individually (see Table 3.2 and Table A.1.). We noted that 10 out of 69 genes contained statistically significant hinge-neighboring mutations for at least one cut-off threshold. The percentages of hinge-neighboring mutations across the genes varied from 6.74% to 98.78%. The oncogene KRAS had the highest hinge percentage in all cut-off thresholds, whereas the tumor suppressor B2M had the second most at the cut-off threshold 10.0 Å. Both proteins have only eight hinge residues (Table A.2.). Considering their sequence lengths (189 for KRAS and 119 for B2M), the proportions of the number of hinge residues to the sequence lengths are 4.2% and 6.7%, respectively (Table A.3. and Figure 3.2). Therefore, these high values were not due to the high numbers of hinge residues.

Furthermore, the ratios between the numbers of hinge residues and the sequence lengths for proteins with statistically significant results were below 7%. We also examined the ratios between the numbers of residues covered by all hinge sites and the total numbers of residues at four different distance thresholds (Figure 3.3). Here, the ratio for B2M at the distance threshold 10.0 Å was 74%. This value is clearly an outlier since other values range from 1.38% to 46%.

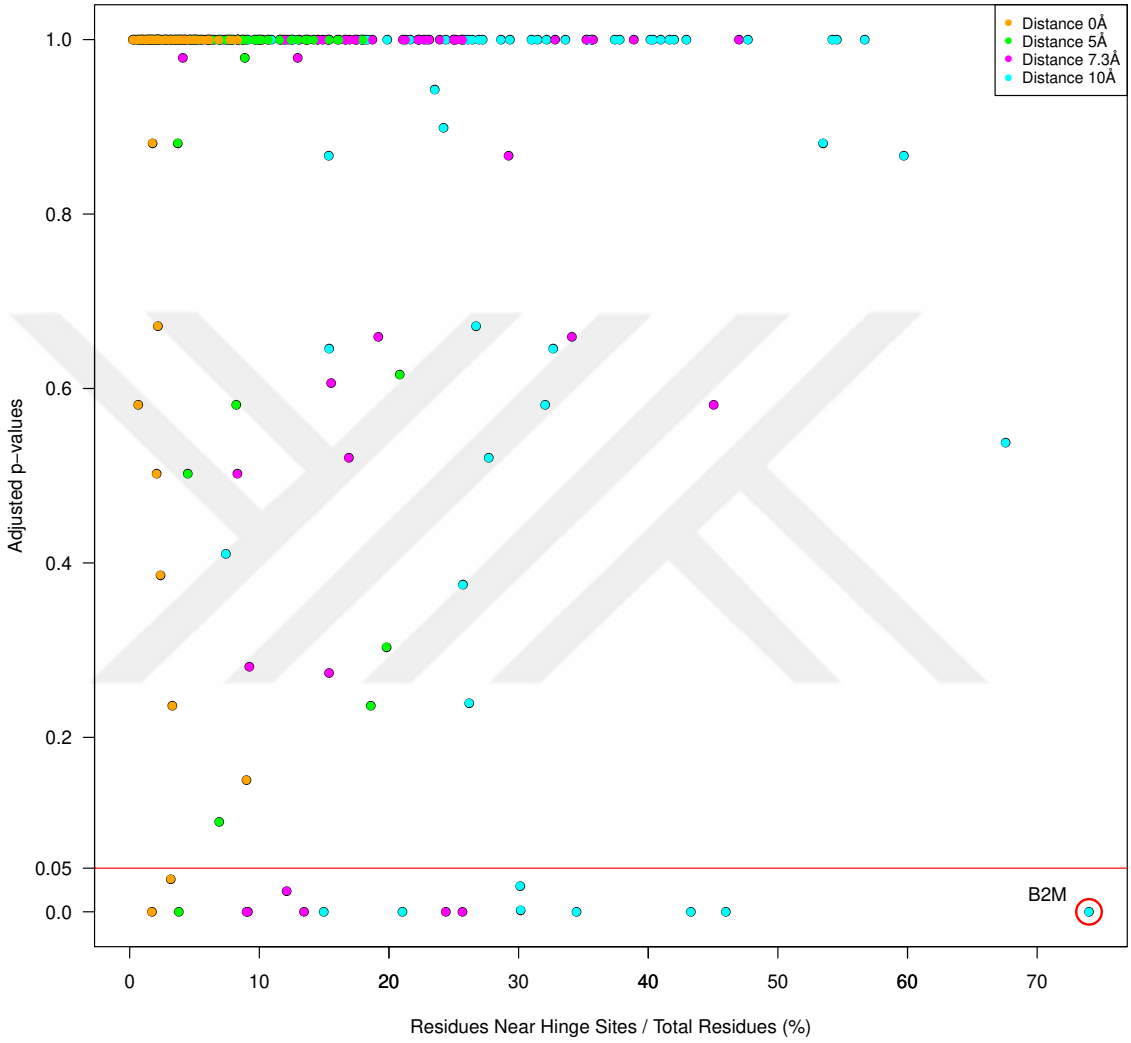

Figure 3.3: The ratios between the numbers of residues covered by all hinge sites and the total numbers of residues at four different distance thresholds. On the y-axis, we have the adjusted p-values, and on the x-axis, we have the percentages of the residues that are covered by all the hinge sites for each protein at four different distance thresholds. B2M at distance threshold 10 Å is in red circle to show that it is an outlier among the statistical significant proteins.

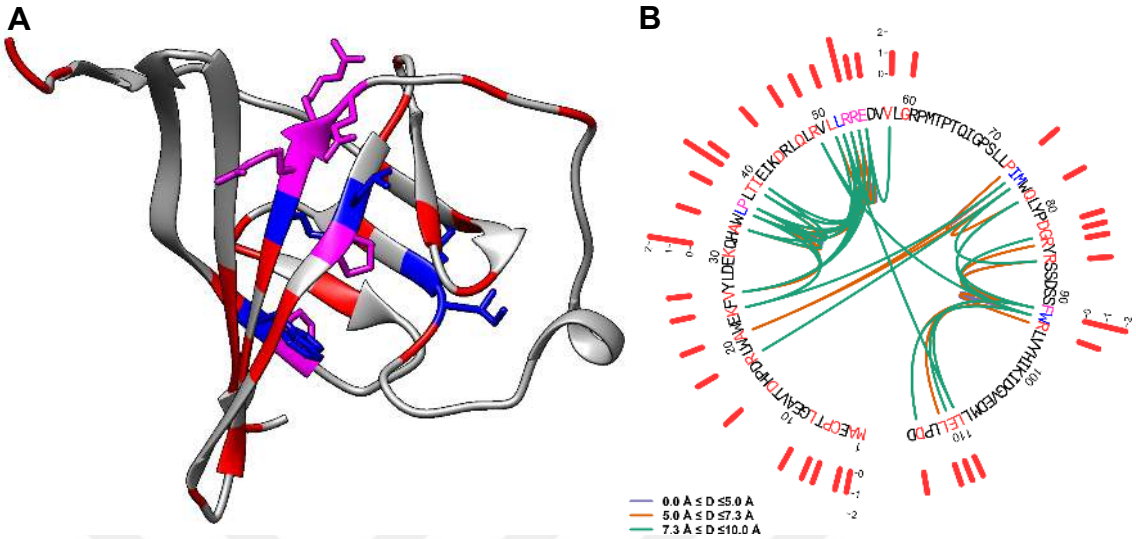

Figure 3.4: Computational analysis for the hinge and mutated residues of TCL1A. (A) 3D structure of TCL1A (PDB ID code 1JSG). Red color shows the mutated residues, blue color shows the hinge residues, and magenta color shows the mutated hinge residues. (B) Protein sequence of TCL1A with mutated residues from the COSMIC. Red color shows the mutated residues, blue color shows the hinge residues, and magenta color shows the mutated hinge residues. The arcs indicate the distances between a mutation and a hinge residue. D in the legend denotes the distance between two residues.

Even though we did not observe statistically significant enrichment for the overall oncogenes at the cut-off thresholds 0.0 Å and 5.0 Å, the oncogenes TCL1A (see Figure 3.4) and MAP2K1 have significant enrichment of hinge-neighboring mutations at these cut-off thresholds. The percentage for MAP2K1 is the lowest value and is also low in absolute terms, however, the value is still significantly higher than it could occur by random chance.

EZH2 is the only gene with an overrepresentation of hinge-neighboring mutations for all of the cut-off thresholds. The CGC has stated EZH2 as a TSG, but EZH2 also functions as an oncogene in some cancer types [Yamaguchi and Hung, 2014]. Y641F/N/S/H/C, A677G, and A687V mutations were identified as activating mu-

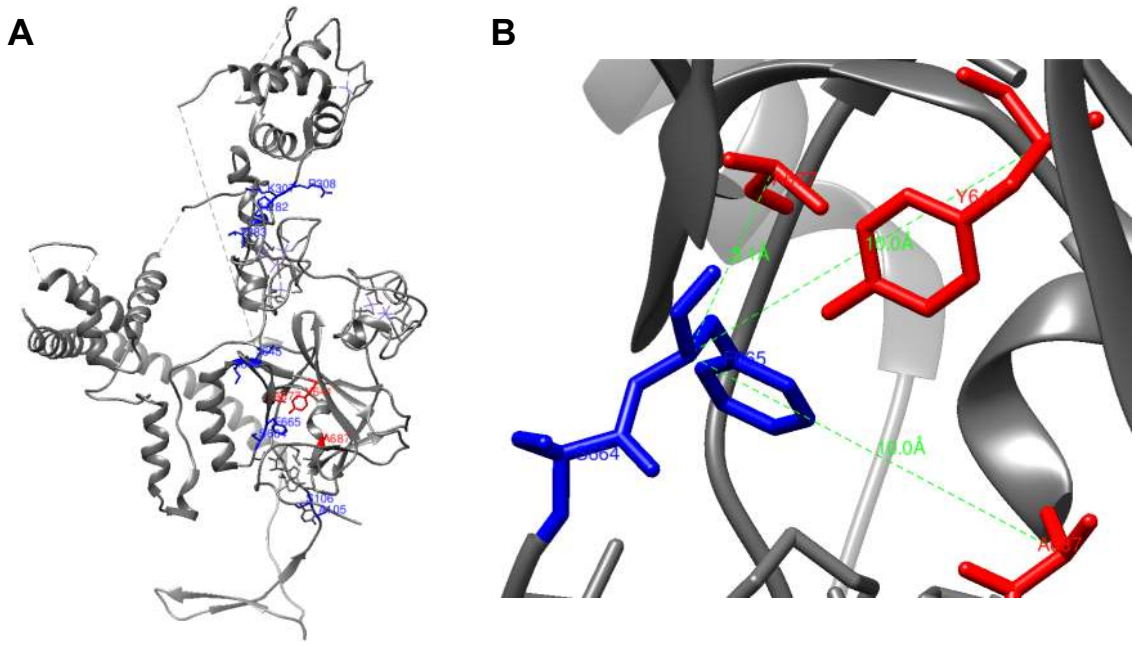

Figure 3.5: The hinge and mutated residues of EZH2. (A) 3D structure of EZH2 (PDB ID code 5HYN_A). Blue color shows the hinge residues, which are aligned on the hinge axis, and red color shows the mutated residues (i.e., Y641F/N/S/H/C, A677G, and A687V). (B) The green lines show the distances between the α-carbon atom of the hinge residue F665 and the α-carbon atoms of the three mutated residues.

tations of EZH2 in lymphoma leading to hyper-trimethylation of H3K27 [Sneeringer et al., 2010, McCabe et al., 2012, Majer et al., 2012]. I645, I646, S664, and F665 are among the ten hinge residues we detected for EZH2 using the PDB structure 5HYN_A. A677, A687, and Y641 are within 5.1 Å, 10.0 Å, and 10.0 Å α-carbon distances from F665, respectively (see Figure 3.5). The mutation clustering around the hinge residue F665 may disrupt the dynamics along the global hinge axis. This disruption may explain the structural mechanism changes of EZH2 causing hyper-trimethylation of H3K27.

Chapter 4

PROTEIN DYNAMICS ANALYSIS IDENTIFIES CANDIDATE CANCER DRIVER GENES AND MUTATIONS IN TCGA DATA

4.1 Identification of Hinge-Mutation Clusters

Our approach revealed 71 hinge-mutation clusters on 54 proteins out of 1,969 proteins with a false discovery rate cutoff of 0.05 ($q \leq 0.05$) (Table 4.1 and Figure 4.1). Figure 4.1 depicts the 54 proteins according to their modes and shows their log transformed p-values with respect to the fold change of the number of hinge-neighboring mutations from the TCGA data set and the randomly resampled number of mutations near hinge residues. These proteins also include known cancer driver genes such as TCL1A, PTPN11, CDK4, BAX, MAPK1, and CDKN2A [Tate et al., 2019] (Table 1 and Figure 4.2A).

Besides these known cancer driver genes, we identified genes that play a role in many biological processes: (i) DDB1, CUL4B, and SKP1 in histone H2A monoubiquitination, (ii) PYGL and ITKP1 in programmed necrotic cell death, and (iii) IL5RA and PLA2G4A in response to toxic substance [The Gene Ontology Consortium, 2017]. In addition, the patients who harbored hinge-mutation clusters in head and neck squamous cell carcinoma (HNSC), colon adenocarcinoma (COAD), and glioblastoma multiforme (GBM) cohorts have a significantly shorter survival time compared to patients without hinge-mutation clusters (p-value = 0.01, 0.05, and 0.03, respectively, log-rank test) (Figure 4.2B, 4.2C, and 4.2D). In this study, we present selected genes from our list of highly hinge-mutated genes such as CDK4, CDKN2A, TCL1A, MAPK1, and PLA2G4A, where we have prior evidence that some of them are already known to play a role in HNSC, COAD, and GBM.

Table 4.1: List of genes with statistically significantly enriched hinge-mutation clusters.

Gene	Mode	HingeResidues	ClusterSize	ClusterRank
CDKN2A	5	112-113	27	5
MAPK1	4	317-318	23	3
PTP4A3	3	107-108	12	6
PTP4A3	4	141-142	12	6
PTPN11	2	282-283	12	7
PIK3CG	3	191-192	11	6
PLA2G4A	1	144-145	10	5
AMY1A	1	383-384	10	4
ACTB	1	335-336	10	3
TCL1A	1	35-36	10	7
CLC	1	81-82	9	5
CUL4B	3	367-368	9	5
TLR5	4	175-176	9	3
DHPS	4	60-61	9	4
PRDX6	2	104-105	8	6
TCL1A	1	53-54	8	5
PTP4A3	1	143-144	8	4
AMY2A	3	175-176	8	6
CRYGB	3	61-62	8	4
ITPK1	4	257-258	8	5
CAMK1D	5	106-107	8	5
RHO	1	301-302	7	4
HIBCH	2	377-378	7	5
AMY2A	3	367-368	7	6
BAX	3	130-131	7	4
IMPA1	4	23-24	7	4
NAALAD	5	664-665	7	5
IL5RA	5	169-170	7	5
NME1	5	126-127	7	5
CDK4	2	218-219	5	3

Genes with statistically significantly enriched hinge-neighboring mutations with at least five mutations. Mode and HingeResidues columns indicate the slow mode and the residue numbers of the hinge residues, respectively. ClusterSize and ClusterRank indicate the number of mutations and the number of unique mutations near the hinge residues, respectively. Table A.5. shows the complete list of genes with statistically significantly enriched hinge-neighboring mutations.

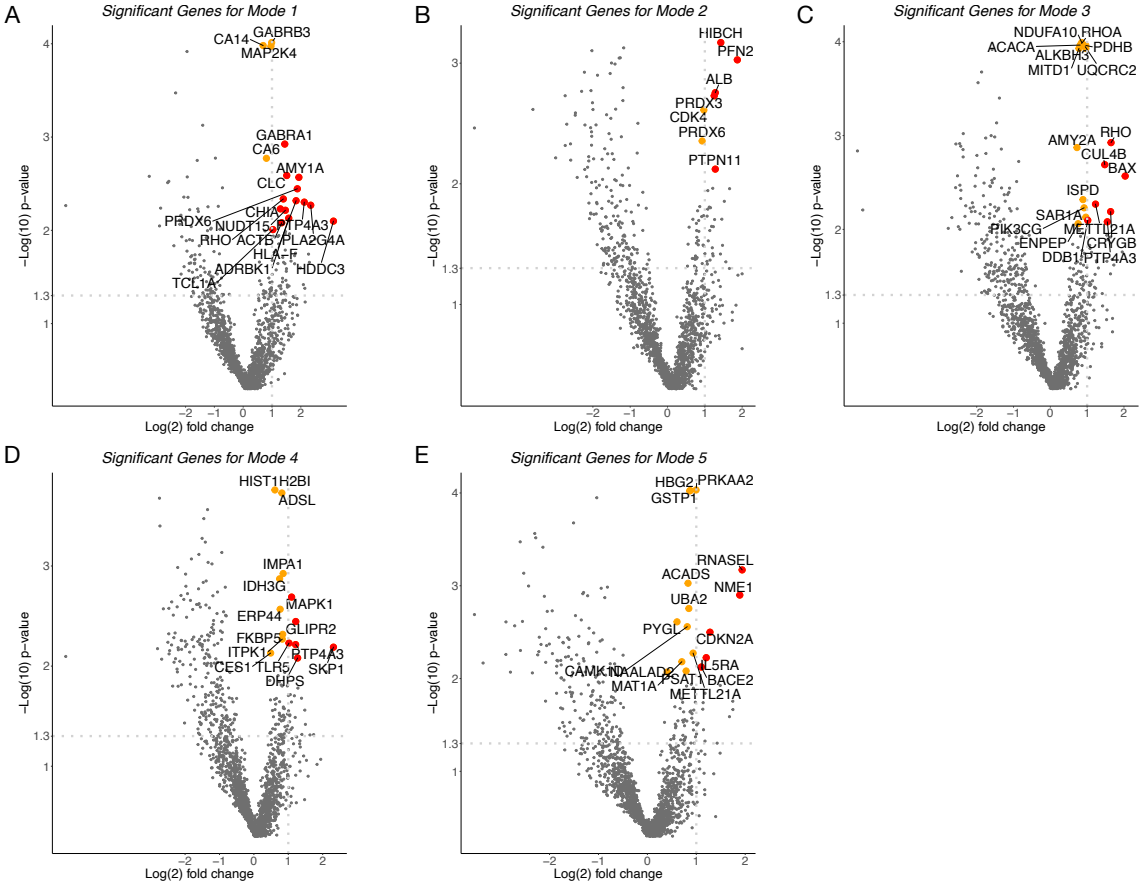

Figure 4.1: Volcano plots reporting p-values against fold changes. Each panel marks the statistically significant genes according to FDR adjusted p-values for each mode from 1 to 5 from (A) to (E), respectively. Red dots show log(2) fold change of higher than 1 and orange dots show log(2) fold change of lower than 1 and positive. The value 1.3 equals to $-\log(10)(p\text{-value} = 0.05)$.

4.2 Cases of Proteins with Hinge-Mutation Clusters

4.2.1 CDK4

CDK4 is a serine/threonine kinase with a catalytic domain that is inactive when it is under-phosphorylated and monomeric [Morgan, 1997]. Activation of CDK4 requires phosphorylation of the catalytic domain at the conserved residue Thr172 by cyclin-dependent kinase CAK1 [Kato et al., 1994]. CDK4 associates with D-

type cyclins forming a crucial complex in cell cycle G1 phase progression [Baker and Reddy, 2012]. One way to turn off an active CDK4-cyclin D complex is by the INK4 family of proteins such as p16 INK4a, which inhibits cyclin D activity by specifically associating with CDK4 [Blain, 2008]. We detected in our work the hinge-mutation cluster E206K, E206D, L213F, G216E, G216V near the hinge residues L218-F219 for CDK4 (PDB: 3g33), which is remote from both the cyclin D and the p16 INK4a interaction interfaces, respectively. In addition, since this cluster is near to hinge residues of slow mode 2, which characterizes global motions, the mutations in this cluster might interrupt either the interaction with cyclin D or with p16 INKa allosterically.

4.2.2 p16 INK4a

p16 INK4a tumor suppressor (CDKN2A) controls cell growth at G1 phase by binding to and inhibiting CDK4/6 [Serrano et al., 1993]. Tumor-derived p16 INK4a mutants are known to be damaging for CDK4/6 binding and inhibition. This defect in function is caused by mutations such as D108N [Huot et al., 2002] and P114L [Tevelev et al., 1996]. Along with these two mutations, our data indicates that the mutations R107H (n=1), D108Y (n=10), D108N (n=4), D108V (n=1), D108G (n=2), G111D (n=1), R112H (n=2), P114T (n=2), and P114L (n=4) cluster around the hinge residues R112-L113 of mode 5 for p16 INK4a (PDB: 1a5e). Interestingly, it has been shown that P114L perturbs the conformation of p16 INK4a significantly [Tevelev et al., 1996]. We believe that the other mutations in the cluster beside P114L also might show similar effects on p16 INK4a.

4.2.3 TCL1A

TCL1A is also one of the cancer driver genes we have identified as containing a hinge-mutation cluster. TCL1A plays an important role in the phosphorylation and activation of AKT1, therefore, regulates cell proliferation and survival. We detected its hinge residues L35-P36 of mode 1 on the PDB structure 1jsg. The missense mutations V25M (n=1), A33T (n=1), T38I (n=1), L50F (n=1), R52H (n=1), R52C

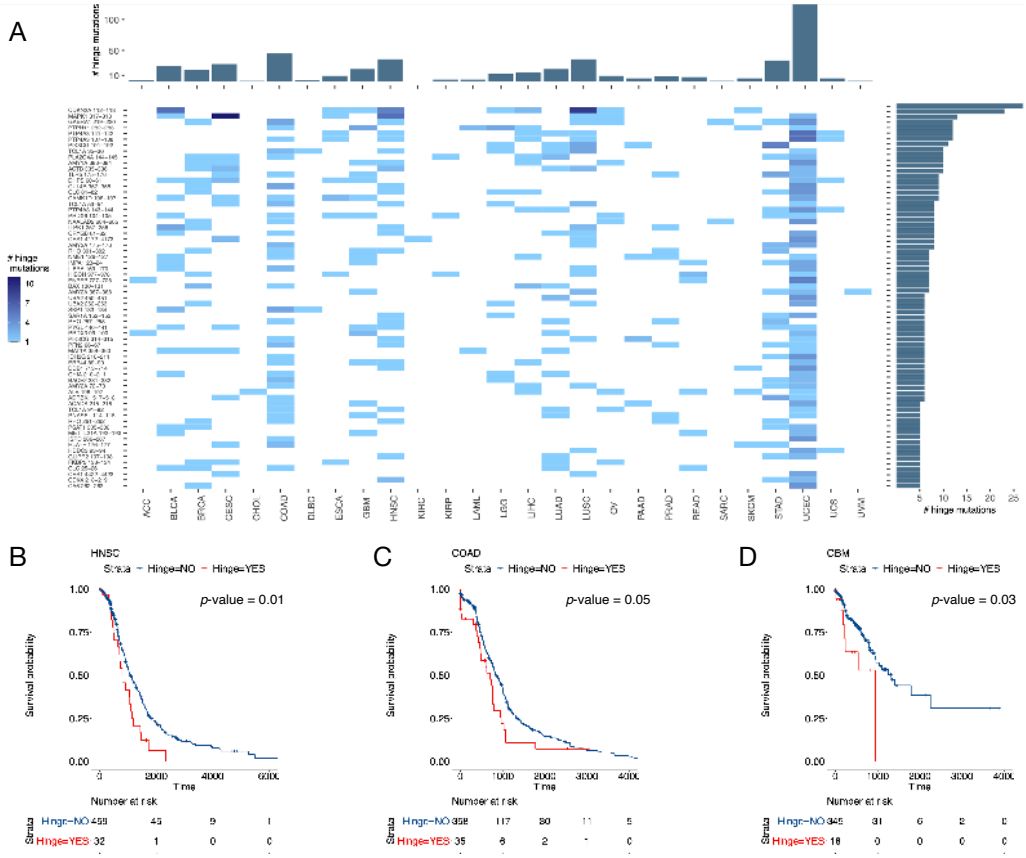

Figure 4.2: Significantly mutated genes and cancer types by hinge-mutation clusters in TCGA. (A) shows gene and hinge residues on the y axis and the cancer types on the x axis. The color bar indicates the count of mutations as in the legend. The bar charts are the summations of each row and column. (B), (C), (D) show the Kaplan-Meier plots for HNSC, COAD, and GBM, respectively.

(n=2), R52S (n=1), R53W (n=1), and E54Q (n=1) cluster around the hinge residues (Figure 4.3A). Being not indicated as oncogenic in the OncoKB database, makes these missense mutations promising candidates for further experimental validation of tumor progression regulated by TCL1A.

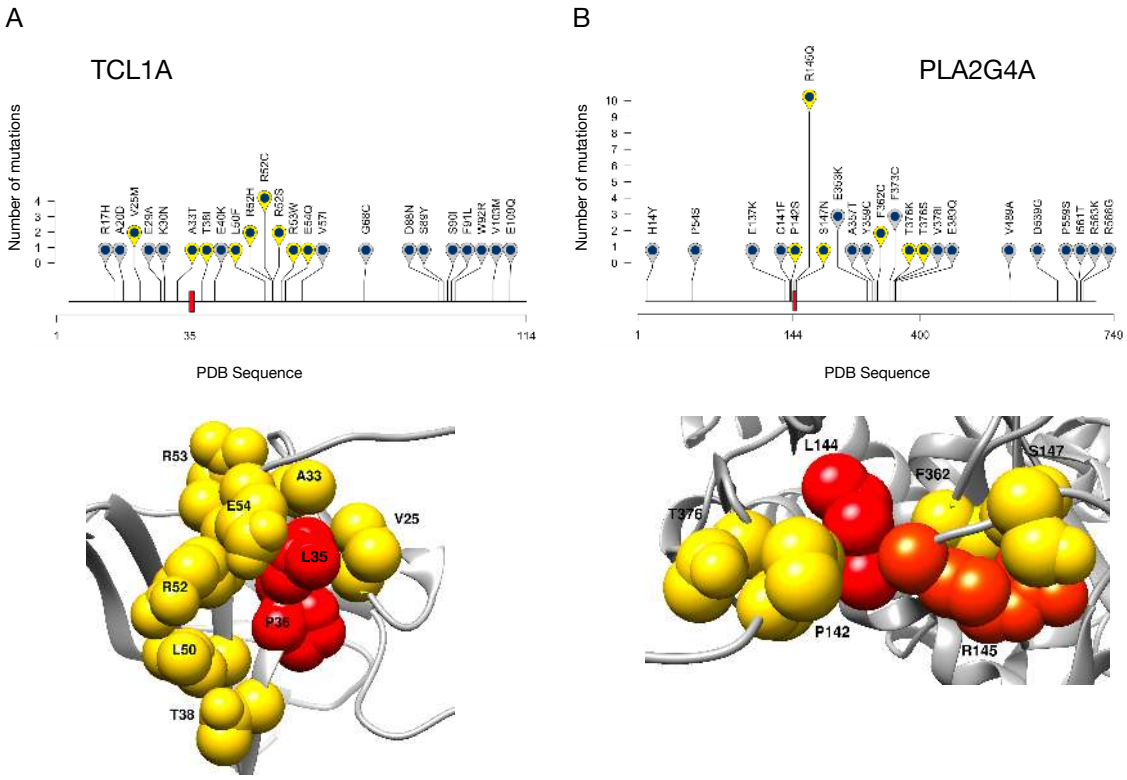

Figure 4.3: Examples of hinge-mutation clusters in 1D and 3D for (A) TCL1A and (B) PLA2G4A. In 1D representations, we show hinge-mutations in yellow, other missense mutations in gray, and hinge residues in red. In 3D representations, we show residues with hinge-mutations as yellow spheres and hinge residues (L35-P36 in TCL1A and L144-R145 in PLA2G4A) as red spheres. Note that R145 in PLA2G4A is a hinge residue and has missense mutations at the same time.

4.2.4 ERK2

Another important cancer driver protein is ERK2 (MAPK1). Activation of ERK2 needs phosphorylation of the threonine and tyrosine residues in the T-E-Y motif resides in the kinase activation loop. Upon phosphorylation, ERK2 undergoes conformational changes that enable it to interact with its substrates [Hoofnagle et al., 2001]. Phosphorylation of the ERK2 T-E-Y motif occurs in a systematic manner;

first, the Y185 residue and then the T183 residue are phosphorylated by MEK1/2 [Aoki et al., 2011]. When ERK2 is phosphorylated on a threonine and a tyrosine residue, the phosphorylation lip refolds in a state that the phosphothreonine and phosphotyrosine align with the arginine-rich binding sites. This results in conformational changes in L16 (residues between A327 and D337) which leads to interactions such as ERK2 dimerization [Canagarajah et al., 1997]. ERK2 dimerization is physiologically important such that preventing dimerization reduces ERK2's ability to accumulate in the nucleus and attenuates cellular proliferation, transformation, and tumor development [Casar et al., 2008, Khokhlatchev et al., 1998]. In our study, we identified a hinge-mutation cluster around Y317-D318 of mode 4 for ERK2 (PDB: 1wzy) consisting of the mutations Y316C/F, D321E/N/G, and E322K (n=18). The ERK2 E322K mutation is a well-known hotspot mutation that is seen in approximately 1.3% of HNSC and 8.0% of CESC and enhances constitutive activation of ERK2 [Wen et al., 2016]. The E322K mutation is on the ERK2 common docking (CD) site (residues between L313 and E326), which is important for binding to its regulators and substrates [Tanoue et al., 2000]. Furthermore, the E322K mutation destabilizes the S320-A327 loop in the L16 and CD sites and causes decreased rates of ERK2 dephosphorylation by DUSP6. The ERK2 E322K mutant destabilizes the unphosphorylated configuration so that it adopts the docked configuration [Taylor et al., 2019]. Interestingly, the D321N ERK2 structure similarly interferes with the interaction of ERK2 with DUSP6 but does not alter the wild type structure [Taylor et al., 2019]. Detecting such important clusters in these cancer driver proteins encourages us to investigate the hinge-mutation clusters in candidate cancer driver proteins such as Cytosolic phospholipase A2.

4.2.5 PLA2G4A

Cytosolic phospholipase A2 (PLA2G4A) selectively hydrolyzes arachidonyl phospholipids to produce arachidonic acid, which is further catalyzed by Cyclooxygenase-2 (COX2) to prostaglandin E2 (PGE2) [Zhang et al., 2018, Yang and Chen, 2008]. PGE2 activates membrane prostanoid receptors such as EP4 [Chell et al., 2006]. It

has been shown that PGE2 activation of EP4 receptors stimulates a PI3K-dependent pathway leading to the phosphorylation of the extracellular signal-regulated kinases (ERKs) [Fujino et al., 2002]. Stimulated ERKs may explain how the knockout of the PLA2G4A gene dramatically decreases lung and prostate tumorigenesis in mice [Meyer et al., 2004] and prostate tumorigenesis in mice [Patel et al., 2008]. On the other hand, the released arachidonic acid triggers sphingomyelinases by activation of PLA2G4A to generate ceramide, which in turn induces G0/G1 cell cycle arrest and apoptosis. Overexpressed COX2 might deplete arachidonic acid. Then, the downstream ceramide pathway reduces apoptotic signaling [Ilsley et al., 2005]. In parallel with this, a reduced PLA2G4A expression and an increased expression of COX2 are associated with unfavorable survival of patients with gastric cancer [Zhang et al., 2013]. Moreover, the knockout of the PLA2G4A gene enhanced mouse colonic tumor development (Ilsley et al., 2005). Interestingly, we identified a hinge-mutation cluster in PLA2G4A (PDB: 1cjy) consisting of the missense mutations P142S (n=1), R145Q (n=5), S147N (n=1), F362C (n=1), and T376K (n=1) around the hinge residues L144-R145 of mode 1 (Figure 4.3B). In our data set, the patients who harbor R145Q mutations suffer from endometrial, cervical, gastric, and colon cancers. Moreover, the hinge residues L144-R145 are on the loop that connects the two main domains of Cytosolic phospholipase A2, which are PLA2c and C2 domains. Following activation by Ca2+ concentration and/or phosphorylation on the C2 domain, it is posited that the active conformation of the protein is the configuration when the catalytic domain moves closer to the C2 domain to form a groove and the phospholipid moves into this groove to be hydrolyzed [Das and Cho, 2002]. Therefore, R145Q may prevent the active conformation to occur and decrease PLA2G4A's activity in gastric and colon tumors. Thus, besides the reduced PLA2G4A expression, the R145Q may cause PLA2G4A to be fixed in the inactive conformation state in gastric and colon tumors.

4.3 Hinge-Based Features in Pathogenicity Prediction

In addition, we tested as additional features whether distances between hinge residues and missense mutations can improve pathogenicity prediction for missense mutations. There are tools that distinguish cancer driver and passenger missense mutations such as CHASM [Carter et al., 2009] and CHASMplus [Tokheim and Karchin, 2019b] as well as tools for general-purpose pathogenicity prediction such as PolyPhen-2 [Adzhubei et al., 2010] and Rhapsody [Ponzoni et al., 2020].

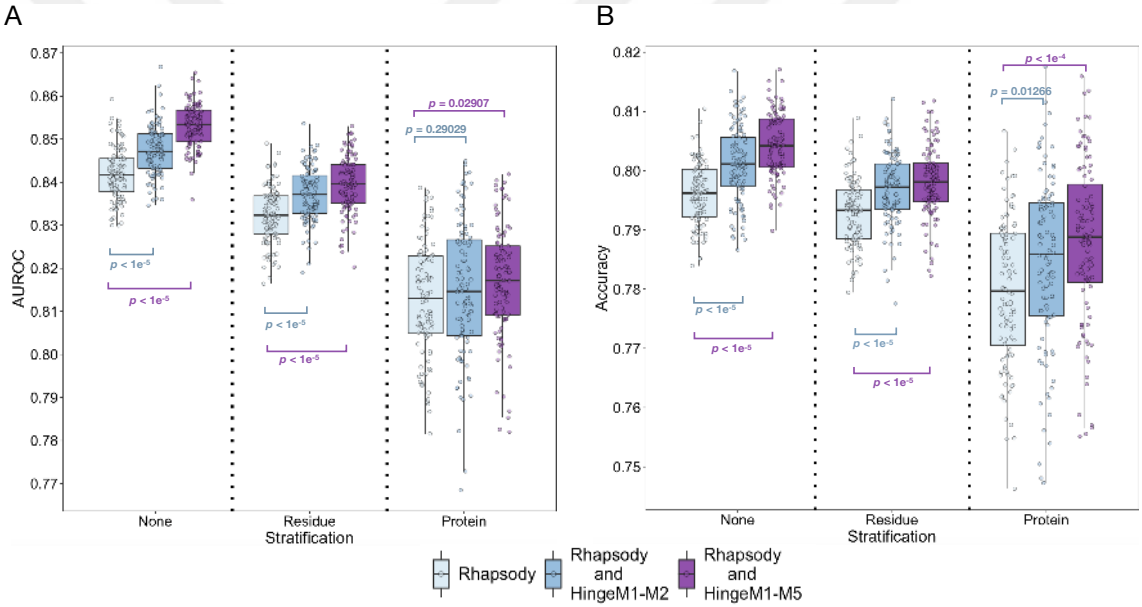

Figure 4.4: The evaluation metrics of Rhapsody, Rhapsody plus HingeM1-M2, and Rhapsody plus HingeM1-M5 models. (A) shows AUROC scores, (B) Accuracy scores for 100 replication runs. The p-values are results from t-tests to compare how different the models' scores are.).

We benchmarked against Rhapsody, since it uses sequence-, structure- and dynamics-based features, we were able to train and test using its full data set without filtering it for missense mutations that can be annotated to a PDB structure, therefore avoiding any sampling bias while benchmarking. We ran a random forest classifier with the hyperparameter values of 1,500 and 2 (for the number of decision trees and the

number of features used to split a node, respectively) as done in Rhapsody.

We also did two stratifications: residue-stratification and protein-stratification. In these stratifications, we separated the variants of the same residue and of the same protein either into the training set or test set. In all the stratification cases, we trained our classifier with 80% of the data and tested with 20% of it, and ran 100 replications. Our hinge-based features are defined as the minimum distances between the missense mutations and the hinge residues for each given mode. Previously, we showed that missense mutations near hinge residues are enriched in cancer-related proteins from the Cancer Gene Census [Sayılgan et al., 2019].

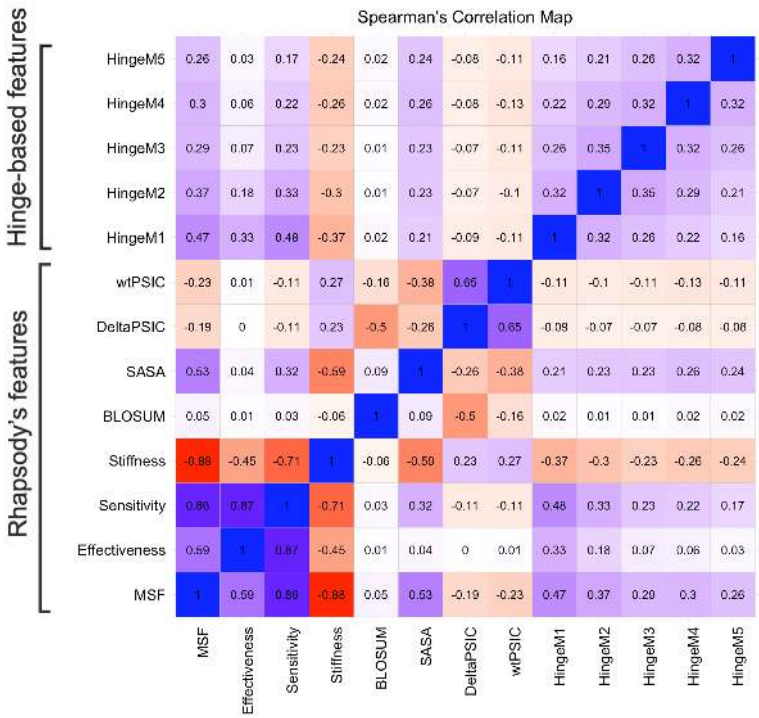

Figure 4.5: The Spearman's correlation scores of the features. HingeM1-M5 indicates hinge-based features for mode 1 to 5, respectively. Other features are from Rhapsody.

Figure 4.4A shows the AUROC scores of Rhapsody, of Rhapsody plus hinge-based features from mode 1 and mode 2, and of Rhapsody plus hinge-based features

from mode 1 to mode 5. We observed that the hinge-based features from mode 1 to mode 5 statistically significantly improves (i.e., more than 1%) the AUROC scores, where the p-values were obtained from two-tailed paired t-tests. Accuracy scores were also statistically significantly improved (Figure 4.4). Furthermore, we calculated the Spearman's correlation scores among features from Rhapsody and hinge-based features (Figure 4.5). Hinge-based features are not highly correlated with ANM-based features (i.e., mean square fluctuations (MSFs), effectiveness, sensitivity, stiffness), which also indicates that hinge-based features adds new information.

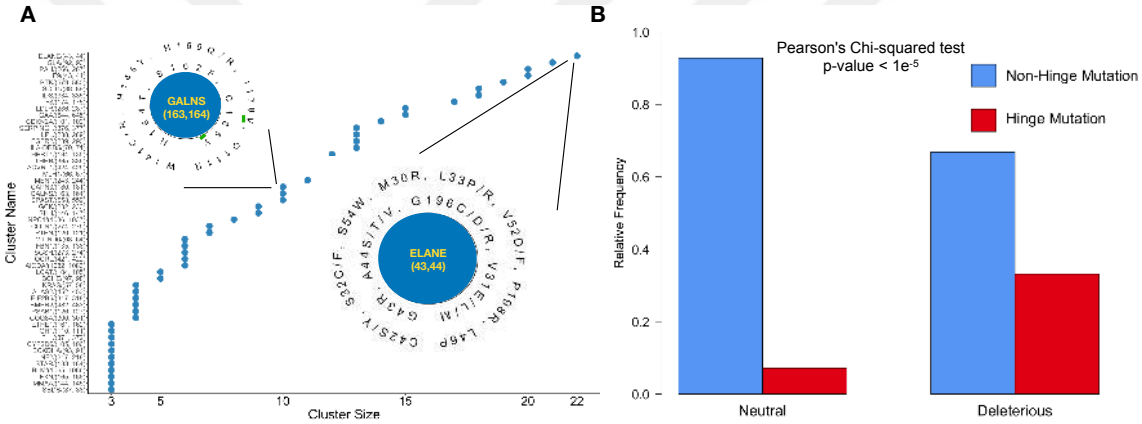

Figure 4.6: (A) shows the hinge-mutation clusters identified using the Rhapsody dataset. Here, we present a subset of clusters due to visualization concerns (For the full list of clusters please see Table A.6.). The gene name and hinge residues are written in the center of the blue dot, and the mapped mutations are circled around according to their distance to the hinge residues. (B) shows the distribution of neutral and deleterious mutations that are hinge-mutations (that clustered around hinge residues less than the cut-off distance 7.3 Å).

MSFs are displacements of α-carbon atoms calculated analytically from the modes of an elastic network model. MSFs contain hinge information by averaging the modes. The correlation scores among MSFs and hinge-based features are between 0.26 and 0.48, which indicates some correlation. However, our approach is to take hinge contribution of each individual mode rather than considering the av-

erage behavior. Furthermore, the correlation scores among the hinge-based features is ∼0.3 which indicates that each individual mode is generally distinctive from the others.

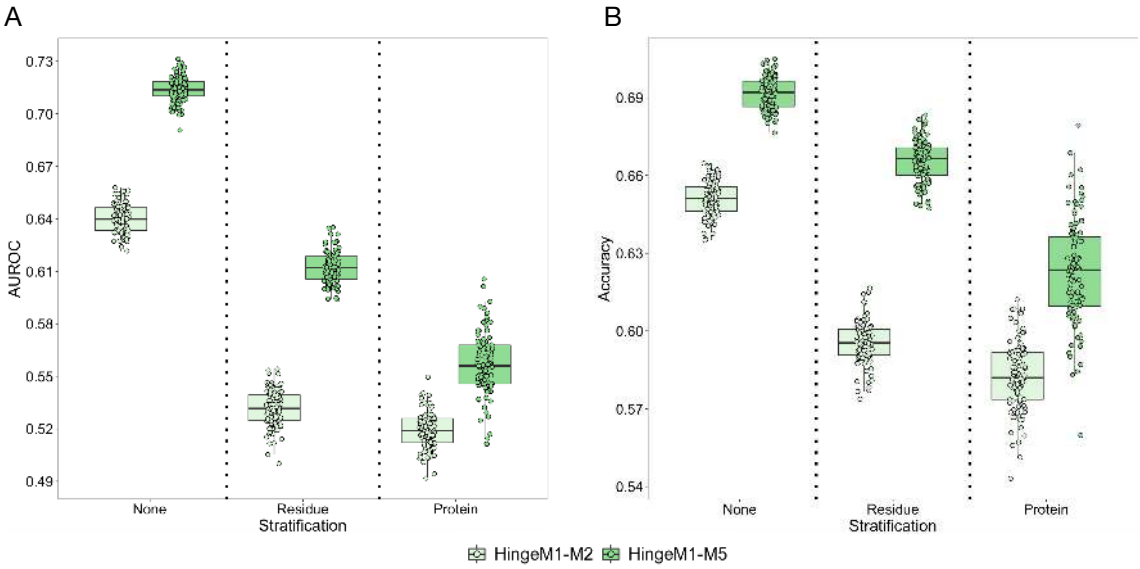

Figure 4.7: The evaluation metrics of HingeM1-M2, HingeM1-M5 models. (A) shows AUROC scores, (B) Accuracy scores for 100 replication runs.

Furthermore, we identified the hinge-mutation clusters in the Rhapsody dataset. To identify hinge-mutation clusters, we checked whether there are three or more missense mutations around hinge residues with a distance less than the cut-off distance 7.3 Å (Figure 4.6A) (For the full list of clusters please see Table A.6.). We found that 4,518 out of 13,636 deleterious missense mutations are clustered around hinge-residues (i.e. hinge-mutations), while 485 out of 6,735 neutral missense mutations are hinge-mutations. Moreover, these deviations are not by chance (Pearson's Chi-squared test, p-value $<$ 1e-5) (Figure Figure 4.6B).

In addition, we ran two models where the first model consists of only the hinge-based features from modes 1 and 2, and the second model consists of only the hinge-based features from modes 1 to 5. By solely having features based on hinge residues and no other features, we can observe whether hinge-based features have

any discriminative power in deleterious and neutral missense mutations. The values of AUROC score (i.e. 0.72) in the none-stratification indicates that the classifier was able to learn that if there are deleterious missense mutations around hinge residues then other missense mutations near hinge residues are likely to be deleterious too (Figure 4.7A).

The metric values of residue- and protein-stratification are expected to be lower since it becomes more difficult for the classifier to make predictions in these cases. However, we believe that the mean values of the accuracy metric for the model with hinge-based features from modes 1 to 5 is 0.69, 0.66, and 0.62 for none-, residue- and protein-stratification, respectively, are impressive (Figure 4.7B). From the metric results, we inferred that hinge-based features can indeed discriminate between some deleterious and neutral missense mutations.

Chapter 5

PREDICTING THE PATHOGENICITY AND THE DISEASE CATEGORY OF MISSENSE VARIANTS

5.1 Performance Evaluation of PathDis

PathDis is a supervised multilabel classification method for predicting pathogenicity and associated disease category type for the missense variants using four groups of features (i.e. sequence-based (SEQ), structure-based (STR), dynamics-based (DYN) and network-based (NET)). Before training our model, we stratified our dataset according to three stratification types. 'none-stratification' is randomly splitting the dataset into 80% training data and 20% testing data. 'residue-stratification' is randomly splitting the dataset with the criterion that the variants on the same residue are either in the training data or in the testing data. 'protein-stratification' is randomly splitting dataset with the criterion that the variants on the same protein are either in the training data or in the testing data. In the last two stratifications, the dataset is split into 80% training data and 20% testing data. These stratifications are done for avoiding spurious increases in prediction accuracy [Grimm et al., 2015]. Also, for each stratification type, we trained and tested our model in 100 replications.

In Figure 5.1A, we illustrate the average area under the ROC curve (AUROC) for our model (i.e. PathDis) with mean values 0.8694, 0.8641 and 0.8214 for the none-, residue- and protein-stratification cases, respectively. We compare PathDis with the Rhapsody model, the SEQ model where we only input SEQ features, the SEQ+STR model where we input SEQ and STR features and the SEQ+STR+DYN model where we input SEQ, STR and DYN features. In addition, the accuracy scores for PathDis are 0.8120, 0.8102, and 0.7872 for the none-, residue- and protein-

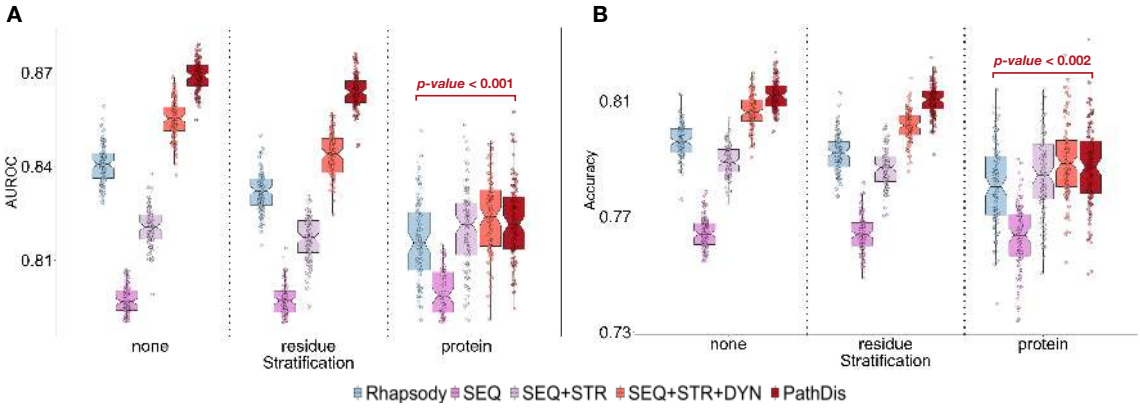

Figure 5.1: Performance evaluation and features of PathDis in pathogenicity prediction. (A) Box plots show the AUROC scores and (B) Accuracy scores on the testing dataset obtained through 100 replications of each model for none-, residue- and protein-stratification, respectively. Missense variants can be found on same residues or proteins, so we stratified the training and test datasets according to residues and proteins so that missense variants on the same residue or protein are either in the training dataset or testing dataset. The dark red color box plots and light blue box plots indicate AUROC scores of PathDis and an existing computational model Rhapsody, respectively. The other box plots show AUROC scores of the depending on the feature group that were input. The p-value was obtained by two-tailed paired t-test between PathDis and Rhapsody in the protein stratification case.

stratification cases, respectively (Figure 5.1B).

Benchmarking with the Rhapsody model's AUROC results (i.e. 0.8407, 0.8317 and 0.8141), PathDis improves AUROC metric results approximately 3% in the none- and residue-stratification cases, and 1% for the protein-stratification case. Moreover, the AUROC results for the SEQ, the SEQ+STR and the SEQ+STR+DYN models show that adding STR, DYN and NET features contribute in the improving in none- and residue-stratification cases. However, we do not observe a similar improving trend in protein-stratification case. There is no statistically significant difference among the SEQ+STR, the SEQ+STR+DYN and PathDis in this case.

The standard deviation for PathDis in the protein-stratification case is almost three times higher than in the other stratification cases (0.0052, 0.0055, and 0.0139, respectively). This may imply that some proteins in the training set are more informative for predicting than other proteins. So, there are not enough informative proteins in the dataset. Increasing the dataset with more available PDB structures in the future may let us observe a similar improvement in the protein-stratification case as in other stratification cases.

Despite there is not any statistically significant difference among the models mentioned above in the protein-stratification case, the difference between PathDis and Rhapsody in this case is statistically significant (p-value < 0.001, paired t-test). This result emphasizes the power of our hinge-based features, network-based features and feature PASE. In the previous chapter, we showed that adding hinge-based features to the Rhapsody model statistically significantly improves AUROC results in every case [Sayılgan et al., 2021].

The Spearman's correlation demonstrates four clusters for each feature group (Figure 5.2A). The SEQ features show correlation scores with magnitudes ranging from 0.09 to 0.65 at the bottom left corner, NET features show correlation scores with magnitudes ranging from 0.29 to 0.63 at the top right corner. In the bottom middle, the secondary structure features of STR have some correlation, and the Solvent Accessible Surface Area (SASA) of STR has higher correlation with the NET features rather than it has with the secondary structure features.

The DYN features Mean Square Fluctuations (MSF), Effectiveness, Sensitivity, and Stiffness from the Rhapsody model have high correlation scores with magnitudes ranging from 0.49 to 0.88. The hinge-based features (i.e. HingeM1, HingeM2, HingeM3, HingeM4 and HingeM5) we created are not highly correlated with these DYN features. This indicates that hinge-based features add valuable and non-redundant information.

In Figure 5.2B, we illustrate the feature importance for the predicting whether a missense variant is deleterious or neutral. We observe that sequence-based features such as DeltaPSIC and wtPSIC are very important for learning in our model. We

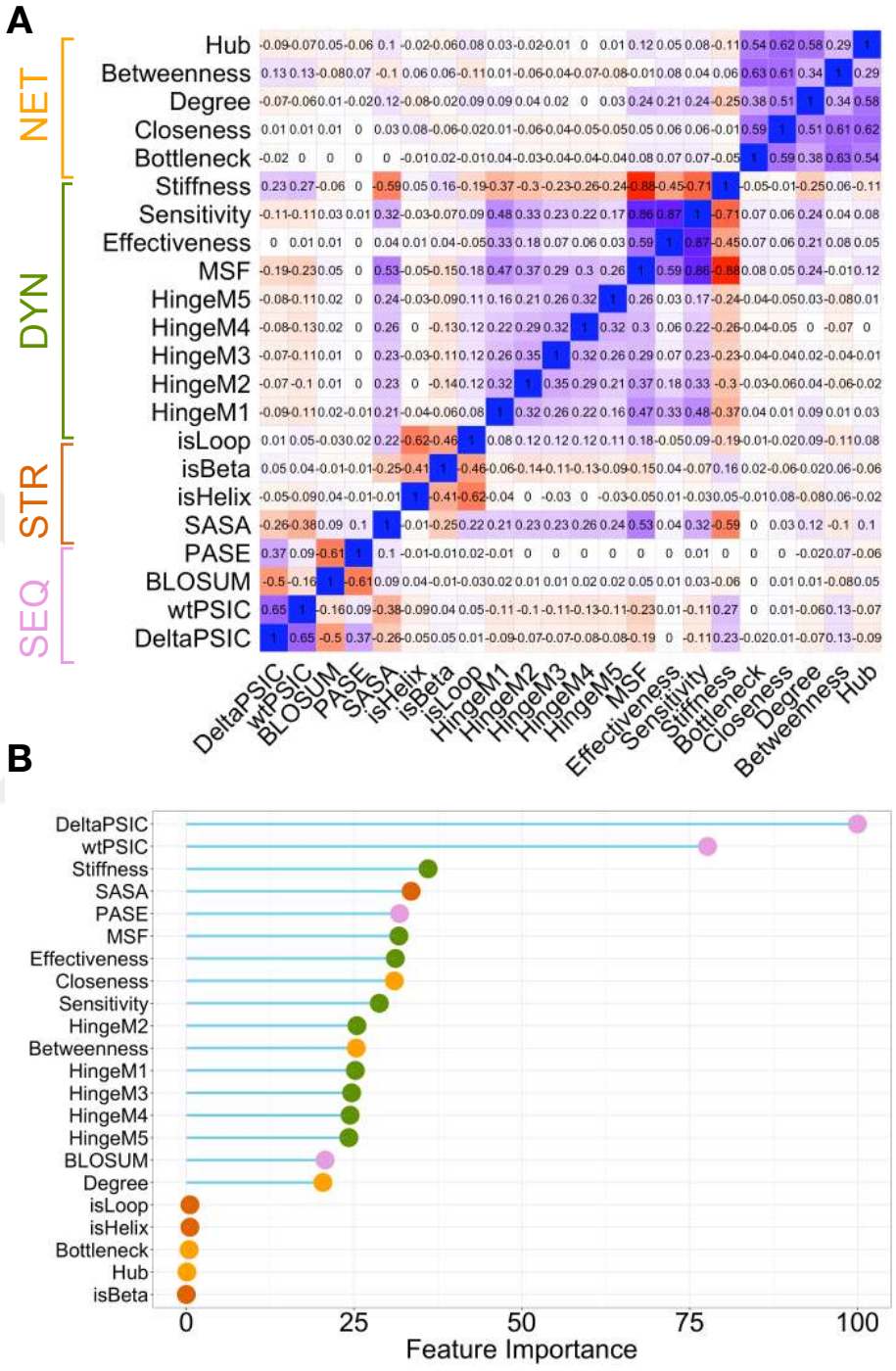

Figure 5.2: Features of PathDis. (A) illustrates Spearman's correlation scores between all pairs of features of PathDis. SEQ, STR, DYN and NET are abbreviations of sequence-, structure-, dynamics- and network-based features, respectively. (B) demonstrates feature importance for predicting whether a missense variant is deleterious or neutral. The colors of the circles are related to the colors of SEQ, STR, DYN and NET abbreviations in (B). The color of the circle indicates the feature group of the corresponding feature.

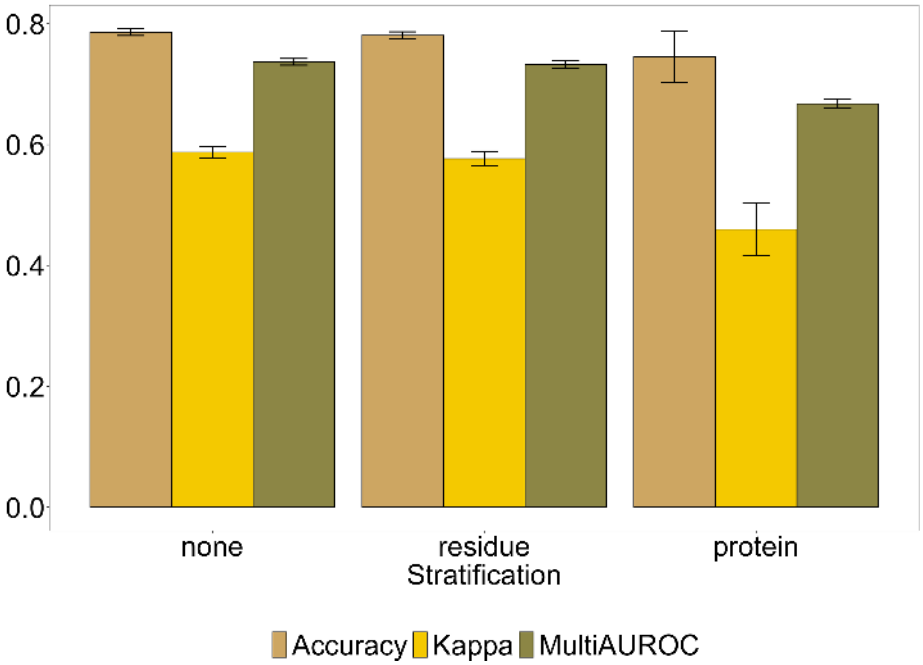

Figure 5.3: Performance evaluation and features of PathDis in disease category prediction. Accuracy, kappa and multiAUROC (i.e. multi-class AUROC) evaluation scores of predicting disease category types (i.e. No Disease, Cancer and Non-Cancer) for none-, residue- and protein-stratification are displayed.

also observe that many features are contributing to our model's learning while the secondary structures features, and Hub and Bottleneck features do not so much. This partially can be because these features are categorical. On the other hand, Hub and Bottleneck features becomes important for learning the disease category type.

Predictions for the disease category types (i.e. No Disease, Cancer and Non-Cancer), our model PathDis has mean accuracy scores 0.7848, 0.7793 and 0.7404, mean Kappa scores 0.5870, 0.5757 and 0.4557, and mean multiAUROC scores 0.7361, 0.7314 and 0.6675 for none-, residue-, and protein-stratification, respectively, on the testing dataset for 100 replications (Figure 5.3).

5.2 Characterizing Missense Variants by NET and SEQ Features for Disease Category Types

We monitored that the most divergent features between Cancer, and No Disease/Non-Cancer disease category types are at SEQ and NET features (Figure 5.4A). SEQ features demonstrate high level of average feature importance for No Disease/Non-Cancer, while not for Cancer. On the other hand, NET features show high level importance for Cancer, while not for No Disease and Non-Cancer. 48.6%, 84.3% and 25.1% of the No Disease, Cancer and Non-Cancer missense variants are on Hub proteins, respectively (Figure 5.4B). Here, we see that Cancer missense variants are selectively on Hub proteins, while Non-Cancer variants are selectively on non-Hub proteins. No Disease variants show no clear preference between Hub or non-Hub proteins.

Degree distributions show a similar trend as the Hub and non-Hub frequency (Figure 5.4C). No Disease missense variants are on proteins with higher Degree than Non-Cancer variants. 48.7%, 80% and 45.7% of the No Disease, Cancer and Non-Cancer missense variants are on Bottleneck proteins (Figure 5.4D). Here, we see a clear selectivity for Cancer missense variants on Bottleneck proteins. However, No Disease/Non-Cancer variants show no clear preference between Bottleneck and non-Bottleneck proteins.

Closeness distributions show that Cancer missense variants have higher Closeness values, while No Disease/Non-Cancer variants have similar values (Figure 5.4E). Thus, Cancer missense variants are on proteins that are closer to other proteins. Betweenness distributions demonstrate that Cancer/Non-Cancer variants prefer proteins with similarly high Betweenness values, while No Disease variants prefer very low Betweenness values (Figure 5.4F).

Overall, we observe that Cancer variants have higher value network topological properties than Non-Cancer variants. Interestingly, only in Betweenness they have similarly high values. Thus, Non-Cancer variants are on especially non-Hub but highly between proteins. In addition, the largest difference between No Disease and

Non-Cancer variants was on Hubness and Betweenness. No Disease variants show no preference on Hubness but on very low Betweenness.

Large positive values of Delta PSIC implies that the amino acid substitution is rarely or never observed in the protein family, values near zero implies that the substitution is almost equally evolutionary conserved [Adzhubei et al., 2013]. No Disease missense variants are rather more evolutionary conserved substitutions than Cancer and Non-Cancer variants (Figure 5.4G).

Delta PSIC distributions also show a clear distinction between neutral and deleterious variants. Delta PSIC was the most important feature for predicting whether a missense variant is neutral or deleterious (Figure 5.2B). Small values of PASE implies that the physicochemical properties between the wild type and mutant amino acids are similar, and vice versa. In Figure 5.4H, we observe that No Disease missense variants are substitutions with more physiochemically conserved than Cancer and Non-Cancer variants.

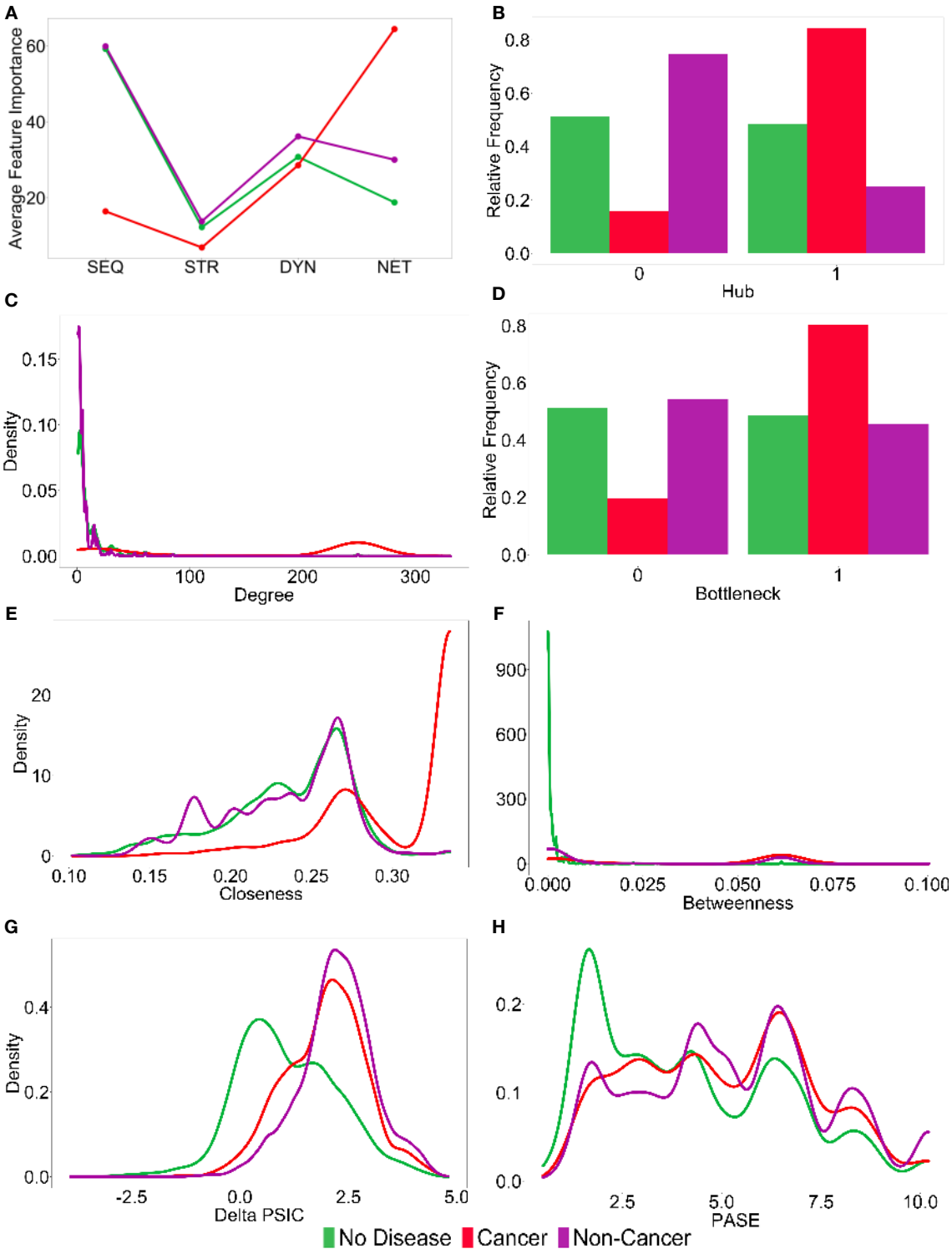

Figure 5.4: Distributions of all NET and two SEQ features. (A) The average feature importance for predicting disease category type. The importance levels for each feature group were calculated by taking the average of each feature importance level within group. (B) The relative frequency of being a non-Hub protein and Hub protein which are indicated by 0 and 1, respectively. (C) The density distribution of feature Degree. (D) The relative frequency of being a non-Bottleneck protein and Bottleneck protein which are indicated by 0 and 1, respectively. The density distribution of Closeness (E), Betweenness (F), Delta PSIC (G), and PASE (H).

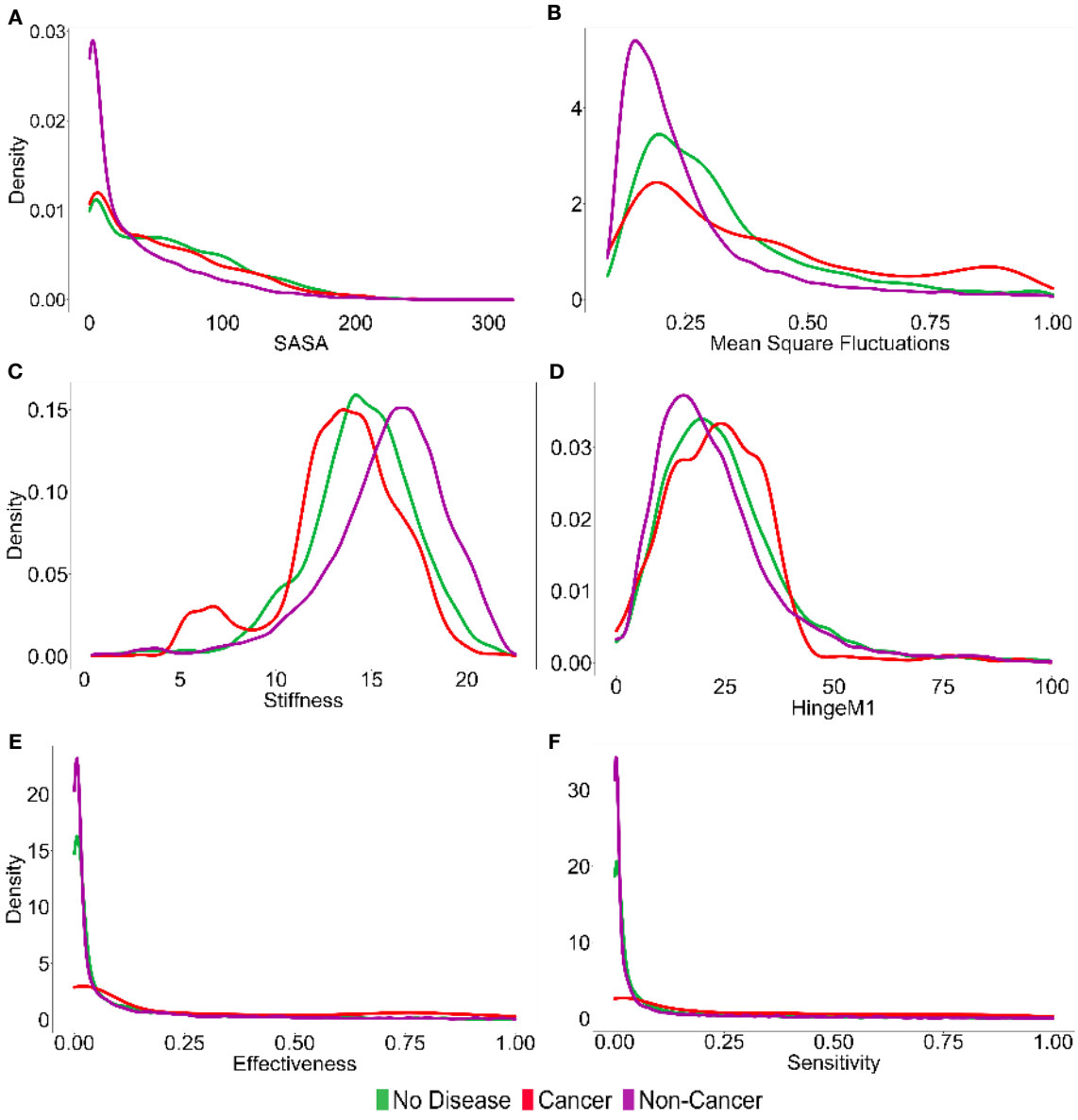

Figure 5.5: Distributions of one STR and DYN features. The density distributions of feature Solvent Accessible Surface Area (SASA) (A), Mean Square Fluctuations (B), Stiffness (C), HingeM1 (D), Effectiveness (E), and Sensitivity (F).

5.3 Characterizing Missense Variants by STR and DYN Features for Disease Category Types

Solvent Accessible Surface Area (SASA) distributions show that Non-Cancer missense variants are concentrated on residues with very low SASA (Figure 5.5A). This shows that Non-Cancer missense variants tend to be on buried residues; however, No Disease and Cancer missense variants tend to be on the surface residues. Previously, we observed that Cancer missense variants prefer Hub or Bottleneck proteins. Thus, with the fact that Cancer missense variants enrich surface residues, this implies that these variants may disturb protein-protein interactions. It was also shown that the protein-protein interaction interfaces of oncoproteins and tumor suppressors are enriched by cancer somatic missense variants [Engin et al., 2016].

The fact that Non-Cancer missense variants prefer proteins with less network centrality and buried residues imply that these variants may disrupt protein stability. Since buried residues are essential both in the early folding stages and in the stability of the folded structure [Ancien et al., 2018]. Moreover, Non-Cancer missense variants occur more frequently on residues with low levels of Mean Square Fluctuations (Figure 5.5B) and high level of Stiffness (Figure 5.5C) which indicates that these residues have low conformational flexibility and high resistance to uniaxial tension, respectively [Ponzoni and Bahar, 2018]. Therefore, evolutionary and physiochemically unconserved Non-Cancer missense variants on buried residues may deform protein stability.

In addition, Figure 5.5D illustrates that, in less than 10 Å neighbor-hood, Cancer missense variants concentrate on the 5 Å distance to hinge residues. We know that Cancer missense variants tend to be on surface residues. Also, it was shown that hinge residues tend to be on surface residues (Flores et al., 2007). Furthermore, we showed that 33% of the deleterious and 7% of the neutral missense variants in the Rhapsody dataset cluster around hinge residues (Pearson's Chi-squared test, p-value $<$ 1e-5) [Sayılgan et al., 2021]. Also, the effect of ef-fectiveness and sensitivity describe allosteric regulation or/and in sub-strate recognition [Ponzoni and Bahar,

2018]. Figure 5.5E and Figure 5.5F demon-strate that low level of Effectiveness and Sensitivity more frequent in Non-Cancer than in Cancer missense variants.

5.4 A Test Case with De Novo PTEN Missense Variants

We considered 3,230 missense variants for the tumor suppressor protein PTEN, which were able to be annotated to a PDB structure and had an abundance class label, to additionally test PathDis. The pathogenicity of these PTEN variants was measured by Variant Abundance by Massively Parallel Sequencing (VAMP-seq) (Table A.7.). Low-abundance variants are enriched in deleterious variants and correlated with low protein thermodynamic stability (Matreyek et al., 2018). Thus, we labeled the abundance classes 'low-abundance' and 'possibly low-abundance' as deleterious, and 'wild type-like-abundance' and 'possibly wild type-like-abundance' as neutral. Before testing for these PTEN variants, we retrained our model without any PTEN variants (56 deleterious and 3 neutral) from the Rhapsody dataset.

Besides testing on PathDis, we also obtained pathogenicity prediction results for these PTEN variants from other pathogenicity prediction models (i.e. Rhapsody [Ponzoni et al., 2020], Provean [Choi and Chan, 2015], PolyPhen2 [Adzhubei et al., 2010] and MutPred2 [Pejaver et al., 2020]). All predictions for these models were attained from their webservers. The area under the ROC curve (AUROC) result for PathDis, Rhapsody, Provean, PolyPhen2 and MutPred2 are 0.732, 0.706, 0.684, 0.657 and 0.630 (Figure 5.6A). The raw predictions for each can be found in the Table A.7.

We computed the Spearman's correlation with the abundance classes of the PTEN variants and the deleterious and neutral predictions of each model (Figure 5.6B). The highest correlation between the PTEN variants was with PathDis scoring 0.34. The next is with Rhapsody scoring 0.32. Also, as expected the highest correlations among the models are between PathDis and Rhapsody scoring 0.59, and between Rhapsody and PolyPhen2 scoring 0.57. This is expected since Rhapsody extracted the important SEQ features DeltaPSIC and wtPSIC from PolyPhen2. Moreover, PathDis extracts both these SEQ features and also some DYN features

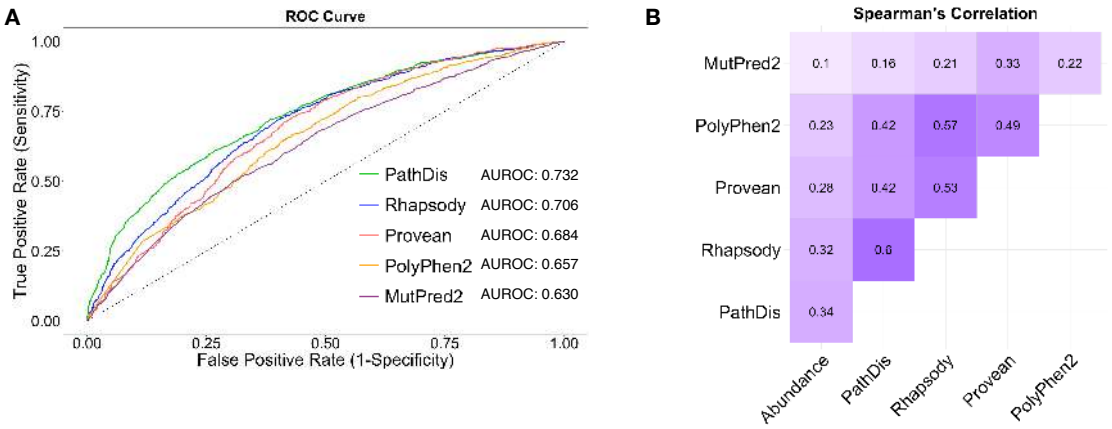

Figure 5.6: Evaluation of pathogenicity predictions for de novo PTEN missense variants. (A) ROC curves obtained from testing the de novo PTEN missense variants for models PathDis, Rhapsody, PolyPhen2, and MutPred2. The AUROC scores for each model are also listed. (B) The Spearman's correlation of the prediction classes (i.e. deleterious or neutral) between all paired models and the abundance class of the PTEN missense variants.

from Rhapsody. The pathogenicity class (i.e. deleterious or neutral) was assigned based on a score cutoff to maximize the Youden's index in the corresponding ROC curve. We set the cutoff to 0.77 for PathDis.

Chapter 6

CONCLUSION

Using the COSMIC dataset, we showed that cancer-related genes have statistically significant enrichment in hinge-neighboring mutations (Table 3.2). MAPK1 is one of the selected 69 genes whose missense mutations were not enriched on hinge residues in all cut-off thresholds (Table A.1.). This might be due to the fact that MAPK1 has a single mutational hotspot, E322K, which is not near to any of its hinge residues. E322K is an activating mutation that enhances EGFR phosphorylation [Van Allen et al., 2015]. Nonetheless, rare mutations on hinge-neighboring residues may play an important role in cancer. For instance, it was demonstrated that missense mutations on the hinge residues M108 and E109 of MAPK1 result in an allosteric effect, which mimics the activation of the kinase caused by phosphorylation [Sours et al., 2014]. Therefore, even though hinge-neighboring mutations of a gene are not enriched, rare mutations on hinge-neighboring residues may have harmful effects.

Missense mutations can have diverse effects on the properties of proteins such as stability, interaction, and conformational dynamics [Stefl et al., 2013, Vihinen, 2015]. Considering the effects of mutations on hinge-neighboring residues, the most related property is conformational dynamics but at the same time possibly uncoupled from other effects. As in the example above, a hinge mutation on MAP kinase 1 could bring the protein into its active conformation by affecting the most cooperative mode of motion that facilitate inactive/active conformational transition. Another study showed that hinge mutations have destabilizing effects on conformational dynamics of the maltose-binding protein (MBP) [Seo et al., 2014]. The mechanism was explained by the increased energy level of the mutant MBP, which facilitated a conformational change from open to close, but hindered a change from

close to open. Furthermore, hinge-neighboring mutations have an allosteric effect on function through conformational change [Sumbul et al., 2015, Clarke et al., 2016].

Furthermore, hinge residues might regulate DNA-binding properties of proteins. Somatic missense mutations in DNA-binding proteins are significantly enriched in cancer; however, the missense mutations rarely locate in the DNA-binding interfaces [Gress et al., 2017]. Apart from these interfaces, mutations in other regions such as hinge residues can disrupt DNA-binding properties of proteins as shown for Androgen Receptor and SMC proteins [Haelens et al., 2007, Hirano and Hirano, 2006]. In our study, only four PDB structures contain DNA (Table A.4.). These structures represent the genes MSH2, DDB2, TOP1, and H3F3A. Clearly, four samples are insufficient for interpretation. Nonetheless, none of the DNA-binding proteins had statistically significantly enriched hinge-neighboring mutations (Table A.1.). Since mutations on DNA-binding proteins do not statistically significantly locate either in the DNA-binding interfaces or on hinge-neighboring residues, future studies need to be performed to decipher plausible dynamic roles of missense mutations on DNA-binding proteins. It may also be the case that DNA-binding proteins mainly suffer from rare driver mutations more than other protein classes do. Furthermore, DNA-binding proteins mostly exist in high-oligomeric states, which is the case with four proteins here. Considering the biological complex structures may disclose functional hinge residues that are enriched by missense mutations.

On the other hand, hinge-neighboring mutations might not always lead to abnormality. For example, shaker K channels in *Drosophila melanogaster* open and close their pores in a process called gating. Some amino acid substitutions on hinges of the shaker channel prevented pore opening. However, when a substitution of glycine was introduced at an adjacent position, opening ability was regained [Magidovich and Yifrach, 2004b]. A mutation neutralized the effect of another. This example presents the difficulty of discovering the effects of missense mutations on hinge-neighboring residues when proteins harbor multiple missense mutations.

Computational analyses on PDB structures have certain limitations. First, not all proteins have PDB structures, and if they have, the sequence of the PDB structure

might not cover the full sequence of the protein. Moreover, the coverage of the PDB structure might change the hinge residues detected by GNM. For instance, GNM might detect different hinge residues for a PDB structure with 70% and a PDB structure with 100% coverage of the same protein. Although the residues may differ, the two sets of hinge residues may have overlapping or adjacent residues depending on the extent of the missing parts of the PDB structure. Therefore, we used PDB structures with at least 70% sequence coverage to be able to obtain robust results (Table A.4.).

The missing parts of the PDB structures are often intrinsically disordered regions of the proteins. For example, the PDB structures of TP53 cover small segments of its sequence. This is due to the high composition of intrinsically disordered regions in TP53. In addition, a study predicted that approximately 79% of the human cancer proteins contain intrinsically disordered regions [Iakoucheva et al., 2002]. Interestingly, cancer-associated mutations are not enriched in disordered regions [Pajkos et al., 2012]. Therefore, the PDB structures of the ordered regions for proteins such as TP53 thus still might contain useful information to investigate the relationship between hinge residues and missense mutations.

In addition to the coverage of the PDB structures, it is important to point out that the asymmetric unit of the PDB structures, and the accommodation of ligand, metal, drug, and modified residues in the PDB structure might affect the process of detecting hinge residues. We filtered out PDB structures with drugs and modified residues. A further limitation was the biological unit of the proteins. To simplify GNM calculations, we treated proteins as monomers.

Also, we systemically searched for spatial clustering of cancer somatic missense mutations around hinge residues from a perspective of protein dynamics using a large-scale TCGA dataset. The approach we used to incorporate protein dynamics consists of basically two steps. We found the statistically significantly enriched missense mutations that cluster around these hinge residues. Overall, we identified 71 hinge-mutation clusters on 54 out of 1,969 proteins. As anticipated, some of these are known cancer proteins such as TCL1A, CDK4, MAPK1, and CDKN2A, whereas

others such as PLA2G4A, DDB1, and PYGL can be considered as candidate cancer proteins that require experimental follow-up. In this study, we did not only suggest candidate driver genes, but also suggest candidate missense mutations that may play driver roles in known and candidate cancer genes. In addition, we demonstrated that hinge-based features statistically significantly improved AUROC scores after adding them to an existing model.

Next, we devised a multilabel classification method for predicting the pathogenicity and the associated disease category type of given missense variants. We utilized the missense variants that can be annotated to a PDB structure. So, with this dataset, we had to group the disease labels into three groups (i.e. No Disease, Cancer and Non-Cancer). However, in the future, with the availability of more PDB structures, the increased dataset can perhaps allow a more refined grouping of disease types. Our model PathDis scored 0.87 in the none-stratification case. This is an improvement from the Rhapsody model's AUROC score 0.84. Also, we acquired 79% accuracy for predicting the disease category type. These high scores were obtained by introducing the features we described above, but also by utilizing some important features were introduced in the literature. Furthermore, it was shown that deleterious missense variants are located at singlet hot spots at protein–protein interaction interfaces [Ozdemir et al., 2018]. However, since not all of the PDB structures have protein-protein interaction information, we cannot utilize the interaction residue information without reducing the size of the Rhapsody dataset. In addition, combining models with paired low Spearman's correlation could improve prediction results.

Besides predicting pathogenicity, we also characterized the missense variants depending on their disease category types based on features. We observed that No Disease missense variants tend to be evolutionary and physiochemically conserved mutations, Cancer missense variants on Hub and Bottleneck proteins' surface residues with being near to hinge residues which may disrupt protein-protein interactions by causing conformational perturbations, and Non-Cancer missense variants on non-Hub but highly between proteins' buried residues which may deform protein

stability with being mutated on residues with high stiffness and low mean square fluctuations. As a result, sequence-based, network-based and structure-/dynamics-based features characterize No Disease, Cancer and Non-Cancer missense variants, respectively.

BIBLIOGRAPHY

[Adzhubei et al., 2013] Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, (SUPPL.76):1–41.

[Adzhubei et al., 2010] Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A., and Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 4(7):248–249.

[Ainscough et al., 2016] Ainscough, B. J., Griffith, M., Coffman, A. C., Wagner, A. H., Kunisaki, J., Choudhary, M. N., McMichael, J. F., Fulton, R. S., Wilson, R. K., Griffith, O. L., and Mardis, E. R. (2016). DoCM: A database of curated mutations in cancer. *Nature Methods*, 13(10):806–807.

[Ancien et al., 2018] Ancien, F., Pucci, F., Godfroid, M., and Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Scientific Reports*, 8(1):1–11.

[Aoki et al., 2011] Aoki, K., Yamada, M., Kunida, K., Yasuda, S., and Matsuda, M. (2011). Processive phosphorylation of ERK MAP kinase in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(31):12675–12680.

[Bahar et al., 1997a] Bahar, I., Atilgan, A. R., Demirel, M. C., and Erman, B. (1997a). Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys Rev Lett*, 80:2733–2736.

[Bahar et al., 1997b] Bahar, I., Atilgan, A. R., and Erman, B. (1997b). Direct

evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2:173–181.

[Bailey et al., 2018] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.

[Bakan et al., 2011] Bakan, A., Meireles, L. M., and Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577.

[Baker and Reddy, 2012] Baker, S. J. and Reddy, E. P. (2012). CDK4: A key player in the cell cycle, development, and cancer. *Genes & Cancer*, 3(11-12):658–669.

[Bendl et al., 2014] Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Computational Biology*, 10(1):1–11.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28:235–242.

[Blain, 2008] Blain, S. W. (2008). Switching cyclin D-Cdk4 kinase activity on and off. *Cell Cycle*, 7(7):892–898.

[Canagarajah et al., 1997] Canagarajah, B. J., Khokhlatchev, A., Cobb, M. H., and Goldsmith, E. J. (1997). Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell*, 90(5):859–869.

[Cancer Genome Atlas Research Network et al., 2013] Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

[Capriotti and Altman, 2011] Capriotti, E. and Altman, R. B. (2011). Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics*, 12(SUPPL. 4).

[Carter et al., 2009] Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Research*, 69(16):6660–6667.

[Casar et al., 2008] Casar, B., Pinto, A., and Crespo, P. (2008). Essential role of ERK dimers in the activation of cytoplasmic but not nuclear substrates by erk-scaffold complexes. *Molecular Cell*, 31(5):708–721.

[Chakravarty et al., 2017] Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., et al. (2017). OncoKB: A precision oncology knowledge case. *JCO Precision Oncology*, 1:1–16.

[Chell et al., 2006] Chell, S. D., Witherden, I. R., Dobson, R. R., Moorghen, M., Herman, A. A., et al. (2006). Increased EP4 receptor expression in colorectal cancer progression promotes cell growth and anchorage independence. *Cancer Research*, 66(6):3106–3113.

[Chen et al., 2019] Chen, Y. A., Tripathi, L. P., Fujiwara, T., Kameyama, T., Itoh, M. N., and Mizuguchi, K. (2019). The TargetMine Data Warehouse: Enhancement and Updates. *Frontiers in Genetics*, 10(October):1–9.

[Choi and Chan, 2015] Choi, Y. and Chan, A. P. (2015). PROVEAN web server:

A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16):2745–2747.

[Choi et al., 2012] Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7(10):e46688.

[Clarke et al., 2016] Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W., Chen, J., and Gerstein, M. (2016). Identifying allosteric hotspots with dynamics: Application to inter- and intra-species conservation. *Structure*, 24:826–837.

[Das and Cho, 2002] Das, S. and Cho, W. (2002). Roles of catalytic domain residues in interfacial binding and activation of group IV cytosolic phospholipase A2. *Journal of Biological Chemistry*, 277(26):23838–23846.

[Denis et al., 2005] Denis, F., Gilleron, R., and Letouzey, F. (2005). Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32:1792–1797.

[Emekli et al., 2008] Emekli, U., Schneidman-Duhovny, D., Wolfson, H. J., Nussinov, R., and Haliloglu, T. (2008). HingeProt: Automated prediction of hinges in protein structures. *Proteins*, 70:1219–1227.

[Engin et al., 2016] Engin, H. B., Kreisberg, J. F., and Carter, H. (2016). Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS One*, 11:e0152929.

[Flores et al., 2008] Flores, S. C., Keating, K. S., Painter, J., Morcos, F., Nguyen, K., Merritt, E. A., Kuhn, L. A., and Gerstein, M. B. (2008). Hingemaster: normal mode hinge prediction approach and integration of complementary predictors. *Proteins*, 73:299–319.

[Forbes et al., 2017] Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., and Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, 45:D777–D783.

[Freeman et al., 2011] Freeman, A. M., Mole, B. M., Silversmith, R. E., and Bourret, R. B. (2011). Action at a distance: Amino acid substitutions that affect binding of the phosphorylated CheY response regulator and catalysis of dephosphorylation can be far from the CheZ phosphatase active site. *Journal of Bacteriology*, 193(18):4709–4718.

[Frishman and Argos, 1995] Frishman, D. and Argos, P. (1995). Knowledge-based secondary structure assignment.

[Fujino et al., 2002] Fujino, H., West, K. A., and Regan, J. W. (2002). Phosphorylation of glycogen synthase kinase-3 and stimulation of T-cell factor signaling following activation of EP2 and EP4 prostanoid receptors by prostaglandin E2. *Journal of Biological Chemistry*, 277(4):2614–2619.

[Futreal et al., 2004] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer*, 4:177–183.

[Gerstein et al., 1994] Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry*, 33:6739–6749.

[Golubeva et al., 2019] Golubeva, V. A., Nepomuceno, T. C., and Monteiro, A. N. (2019). Germline missense variants in brca1: New trends and challenges for clinical annotation. *Cancers*, 11(4).

[Grant et al., 2006] Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon,

J. A., and Caves, L. S. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, 22:2696–2696.

[Gray et al., 2018] Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018). Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*, 6(1):116–124.e3.

[Gress et al., 2017] Gress, A., Ramensky, V., and Kalinina, O. V. (2017). Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes. *Oncogenesis*, 6:e380.

[Grimm et al., 2015] Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., Macarthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., and Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, 36(5):513–523.

[Haelens et al., 2007] Haelens, A., Tanner, T., Denayer, S., Callewaert, L., and Claessens, F. (2007). The hinge region regulates DNA binding, nuclear translocation, and transactivation of the androgen receptor. *Cancer Res*, 67:4514–4523.

[Haliloglu and Bahar, 2015] Haliloglu, T. and Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr Opin Struct Biol*, 35:17–23.

[Haliloglu et al., 1997] Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.*, 79(16):3090–3093.

[Haliloglu et al., 1998] Haliloglu, T., Bahar, I., and Erman, B. (1998). Gaussian dynamics of folded proteins. *Phys Rev Lett*, 79:3090–3093.

[Henikoff and Henikoff, 2000] Henikoff, S. and Henikoff, J. G. (2000). Amino acid substitution matrices. *Advances in Protein Chemistry*, 54(November):73–97.

[Hess et al., 2019] Hess, J. M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N. J., Lawrence, M. S., and Getz, G. (2019). Passenger Hotspot Mutations in Cancer. *Cancer Cell*, 36(3):288–301.e14.

[Hirano and Hirano, 2006] Hirano, M. and Hirano, T. (2006). Opening closed arms: long-distance activation of SMC ATPase by hinge-DNA interactions. *Mol Cell*, 21:175–186.

[Hoofnagle et al., 2001] Hoofnagle, A. N., Resing, K. A., Goldsmith, E. J., and Ahn, N. G. (2001). Changes in protein conformational mobility upon activation of extracellular regulated protein kinase-2 as detected by hydrogen exchange. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3):956–961.

[Hopf et al., 2017] Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135.

[Huot et al., 2002] Huot, T. J., Rowe, J., Harland, M., Drayton, S., Brookes, S., et al. (2002). Biallelic mutations in p16(INK4a) confer resistance to Ras- and Ets-induced senescence in human diploid fibroblasts. *Molecular and Cellular Biology*, 22(23):8135–8143.

[Iakoucheva et al., 2002] Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*, 323:573–584.

[Ilsley et al., 2005] Ilsley, J. N. M., Nakanishi, M., Flynn, C., Belinsky, G. S., De Guise, S., et al. (2005). Cytoplasmic phospholipase A2 deletion enhances colon tumorigenesis. *Cancer Research*, 65(7):2636–2643.

[Kamburov et al., 2015] Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., and Getz, G. (2015). Comprehensive

assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci USA*, 112:E5486–E5495.

[Kato et al., 1994] Kato, J. Y., Matsuoka, M., Strom, D. K., and Sherr, C. J. (1994). Regulation of cyclin D-dependent kinase 4 (cdk4) by cdk4-activating kinase. *Molecular and Cellular Biology*, 14(4):2713–2721.

[Kawashima et al., 2008] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(SUPPL. 1):202–205.

[Khokhlatchev et al., 1998] Khokhlatchev, A. V., Canagarajah, B., Wilsbacher, J., Robinson, M., Atkinson, M., et al. (1998). Phosphorylation of the MAP kinase ERK2 promotes its homodimerization and nuclear translocation. *Cell*, 93(4):605–615.

[Kuhn, 2008] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

[Kumar et al., 2019] Kumar, S., Clarke, D., and Gerstein, M. B. (2019). Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. *Proceedings of the National Academy of Sciences of the United States of America*, 116(38):18962–18970.

[Kundu et al., 2004] Kundu, S., Sorensen, D. C., and Phillips, G. N., J. (2004). Automatic domain decomposition of proteins by a gaussian network model. *Proteins*, 57:725–733.

[Landrum et al., 2016] Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868.

[Laskowski et al., 1997] Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997). PDBsum: A web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci*, 22:488–490.

[Li et al., 2017] Li, H., Chang, Y.-Y., Lee, J. Y., Bahar, I., and Yang, L.-W. (2017). Dynomics: dynamics of structural proteome and beyond. *Nucleic Acids Research*, 45(W1):W374–W380.

[Li et al., 2010] Li, J., Duncan, D. T., and Zhang, B. (2010). CanProVar: A human cancer proteome variation database. *Human Mutation*, 31(3):219–228.

[Li et al., 2013a] Li, M., Kwan, J., Bao, S. Y., Yang, W., Ho, S. L., Song, Y. Q., and Sham, P. C. (2013a). Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genetics*, 9(1):1–11.

[Li et al., 2013b] Li, X., Kierczak, M., Shen, X., Ahsan, M., Carlborg, Ö., and Marklund, S. (2013b). PASE: A novel method for functional prediction of amino acid substitutions based on physicochemical properties. *Frontiers in Genetics*, 4(MAR):1–6.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *RNews*, 2(3):18–22.

[Magidovich and Yifrach, 2004a] Magidovich, E. and Yifrach, O. (2004a). Conserved gating hinge in ligand- and voltage-dependent K+ channels. *Biochemistry*, 43(42):13242–13247.

[Magidovich and Yifrach, 2004b] Magidovich, E. and Yifrach, O. (2004b). Conserved gating hinge in ligand- and voltage-dependent K+ channels. *Biochemistry*, 43:13242–13247.

[Majer et al., 2012] Majer, C. R., Jin, L., Scott, M. P., Knutson, S. K., Kuntz, K. W., Keilhack, H., Smith, J. J., Moyer, M. P., Richon, V. M., Copeland, R. A., and Wigle, T. J. (2012). A687V EZH2 is a gain-of-function mutation found in lymphoma patients. *FEBS Lett*, 586:3448–3451.

[Matreyek et al., 2018] Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Melissa, A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., and Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, 50(6):874–882.

[McCabe et al., 2012] McCabe, M. T., Graves, A. P., Ganji, G., Diaz, E., Halsey, W. S., Jiang, Y., Smitheman, K. N., Ott, H. M., Pappalardi, M. B., Allen, K. E., Chen, S. B., Della Pietra, A., r., Dul, E., Hughes, A. M., Gilbert, S. A., Thrall, S. H., Tummino, P. J., Kruger, R. G., Brandt, M., Schwartz, B., and Creasy, C. L. (2012). Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). *Proc Natl Acad Sci USA*, 109:2989–2994.

[Meyer et al., 2004] Meyer, A. M., Dwyer-Nield, L. D., Hurteau, G. J., Keith, R. L., O'Leary, E., et al. (2004). Decreased lung tumorigenesis in mice genetically deficient in cytosolic phospholipase A2. *Carcinogenesis*, 25(8):1517–1524.

[Morgan, 1997] Morgan, D. O. (1997). Cyclin-dependent kinases: Engines, clocks, and microprocessors. *Annual Review of Cell and Developmental Biology*, 13(1):261–291.

[Mottaz et al., 2010] Mottaz, A., David, F. P., Veuthey, A. L., and Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, 26(6):851–852.

[Ng and Henikoff, 2003] Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino

acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814.

[Niroula and Vihinen, 2016] Niroula, A. and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation*, 37(6):579–597.

[Ozdemir et al., 2018] Ozdemir, E. S., Gursoy, A., and Keskin, O. (2018). Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. *Bioinformatics*, 34(17):i795–i801.

[Pajkos et al., 2012] Pajkos, M., Meszaros, B., Simon, I., and Dosztanyi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst*, 8:296–307.

[Patel et al., 2008] Patel, M. I., Singh, J., Niknami, M., Kurek, C., Yao, M., et al. (2008). Cytosolic phospholipase A_2-α: A potential therapeutic target for prostate cancer. *Clinical Cancer Research*, 14(24):8070–8079.

[Pejaver et al., 2020] Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., and Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, 11(1):5918.

[Pettersen et al., 2004] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., et al. (2004). UCSF Chimera–A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

[Ponzoni and Bahar, 2018] Ponzoni, L. and Bahar, I. (2018). Structural dynamics is a determinant of the functional significance of missense variants. *Proc Natl Acad Sci USA*, in press.

[Ponzoni et al., 2020] Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N., and Bahar, I. (2020). Rhapsody: Predicting the pathogenicity of human missense variants. *Bioinformatics*, 36(10):3084–3092.

[Rentzsch et al., 2019] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894.

[Rivolli and de Carvalho, 2018] Rivolli, A. and de Carvalho, A. C. (2018). The utiml Package: Multi-label Classification in R. *The R Journal*, 10(2):24–37.

[Robin et al., 2011] Robin, X., Turck, N., Hainard, A., Tiberti, N., Sanchez, J.-C., Lisacek, F., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

[Ryslik et al., 2016] Ryslik, G. A., Cheng, Y., Modis, Y., and Zhao, H. (2016). Leveraging protein quaternary structure to identify oncogenic driver mutations. *BMC Bioinformatics*, 17:137.

[Sayılgan et al., 2019] Sayılgan, J. F., Haliloğlu, T., and Gönen, M. (2019). Protein dynamics analysis reveals that missense mutations in cancer-related genes appear frequently on hinge-neighboring residues. *Proteins: Structure, Function, and Bioinformatics*, 87(6):512–519.

[Sayılgan et al., 2021] Sayılgan, J. F., Haliloğlu, T., and Gönen, M. (2021). Protein dynamics analysis identifies candidate cancer driver genes and mutations in TCGA data. *Proteins: Structure, Function and Bioinformatics*, page in press.

[Seo et al., 2014] Seo, M. H., Park, J., Kim, E., Hohng, S., and Kim, H. S. (2014). Protein conformational dynamics dictate the binding affinity for a ligand. *Nat Commun*, 5:3724.

[Serrano et al., 1993] Serrano, M., Hannon, G. J., and Beach, D. (1993). A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature*, 366(6456):704–707.

[Shatsky et al., 2004] Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol*, 11:83–106.

[Sing et al., 2005] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881.

[Sneeringer et al., 2010] Sneeringer, C. J., Scott, M. P., Kuntz, K. W., Knutson, S. K., Pollock, R. M., Richon, V. M., and Copeland, R. A. (2010). Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proc Natl Acad Sci USA*, 107:20980–20985.

[Sours et al., 2014] Sours, K. M., Xiao, Y., and Ahn, N. G. (2014). Extracellular-regulated kinase 2 is activated by the enhancement of hinge flexibility. *J Mol Biol*, 426:1925–1935.

[Stefl et al., 2013] Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. *J Mol Biol*, 425:3919–3936.

[Stehr et al., 2011] Stehr, H., Jang, S. H., Duarte, J. M., Wierling, C., Lehrach, H., Lappe, M., and Lange, B. M. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer*, 10:54.

[Stenson et al., 2017] Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., and Cooper, D. N. (2017). The Human

Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 136(6):665–677.

[Sumbul et al., 2015] Sumbul, F., Acuner-Ozbabacan, S. E., and Haliloglu, T. (2015). Allosteric dynamic control of binding. *Biophysical Journal*, 109:1190–1201.

[Sunyaev et al., 1999] Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering*, 12(5):387–394.

[Tanoue et al., 2000] Tanoue, T., Adachi, M., Moriguchi, T., and Nishida, E. (2000). A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nature Cell Biology*, 2(2):110–116.

[Tate et al., 2019] Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., et al. (2019). COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947.

[Taylor et al., 2019] Taylor, C. A., Cormier, K. W., Keenan, S. E., Earnest, S., Stippec, S., et al. (2019). Functional divergence caused by mutations in an energetic hotspot in ERK2. *Proceedings of the National Academy of Sciences of the United States of America*, 116(31):15514–15523.

[Tevelev et al., 1996] Tevelev, A., Byeon, I. J. L., Selby, T., Ericson, K., Kraynov, V., et al. (1996). Tumor suppressor p16INK4A: Structural characterization of wild-type and mutant proteins by NMR and circular dichroism. *Biochemistry*, 35(29):9475–9487.

[The Gene Ontology Consortium, 2017] The Gene Ontology Consortium (2017).

Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338.

[The ICGC/TCGA Pan-Cancer Analysis and of Whole Genomes Consortium, 2020] The ICGC/TCGA Pan-Cancer Analysis and of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578:82–93.

[The UniProt Consortium, 2017] The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 45:D158–D169.

[Thusberg et al., 2011] Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, 32(4):358–368.

[Tokheim and Karchin, 2019a] Tokheim, C. and Karchin, R. (2019a). CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. *Cell Systems*, 9(1):9–23.e8.

[Tokheim and Karchin, 2019b] Tokheim, C. and Karchin, R. (2019b). Chasmplus reveals the scope of somatic missense mutations driving human cancers. *Cell Systems*, 9(1):9 – 23.e8.

[Touw et al., 2015] Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., and Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368.

[Van Allen et al., 2015] Van Allen, E. M., Lui, V. W., Egloff, A. M., Goetz, E. M., Li, H., Johnson, J. T., Duvvuri, U., Bauman, J. E., Stransky, N., Zeng, Y., Gilbert, B. R., Pendleton, K. P., Wang, L., Chiosea, S., Sougnez, C., Wagle, N., Zhang, F., Du, Y., Close, D., Johnston, P. A., McKenna, A., Carter, S. L., Golub, T. R., Getz, G., Mills, G. B., Garraway, L. A., and Grandis, J. R. (2015). Genomic correlate of exceptional erlotinib response in head and neck squamous cell carcinoma. *JAMA Oncol*, 1:238–244.

[van Buuren and Groothuis-Oudshoorn, 2011] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67.

[Velankar et al., 2012] Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., et al. (2012). SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(Database issue):D483–D489.

[Velankar et al., 2013] Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M. J., and Kleywegt, G. J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res*, 41:D483–D489.

[Vihinen, 2015] Vihinen, M. (2015). Types and effects of protein variations. *Hum Genet*, 134:405–421.

[Vogelstein et al., 2013] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., J., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339:1546–1558.

[Wen et al., 2016] Wen, Y., Li, H., Zeng, Y., Wen, W., Pendleton, K. P., et al. (2016). MAPK1E322K mutation increases head and neck squamous cell carcinoma sensitivity to erlotinib through enhanced secretion of amphiregulin. *Oncotarget*, 7(17):23300–23311.

[Yamaguchi and Hung, 2014] Yamaguchi, H. and Hung, M. C. (2014). Regulation and role of EZH2 in cancer. *Cancer Res Treat*, 46:209–222.

[Yang and Chen, 2008] Yang, H. and Chen, C. (2008). Cyclooxygenase-2 in synaptic signaling. *Current Pharmaceutical Design*, 14(14):1443–1451.

[Yu et al., 2007] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):713–720.

[Zhang et al., 2018] Zhang, W., Wang, X., Zhang, L., Geng, D., Wang, Y., et al. (2018). Inhibition of PLA2G4A reduces the expression of lung cancer-related cytokines. *DNA and Cell Biology*, 37(12):1076–1081.

[Zhang et al., 2013] Zhang, X., Wu, Q., Gan, L., Yu, G.-Z., Wang, R., et al. (2013). Reduced group IVA phospholipase A2 expression is associated with unfavorable outcome for patients with gastric cancer. *Medical Oncology*, 30(1):454.

Appendix A

Table A.1. Detailed list of genes with statistically significantly enriched hinge-neighboring mutations in the COSMIC dataset.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fprot.25673&file=prot25673-sup-0001-AppendixS1.xlsx

Table A.2. List of hinge residues for each protein in the COSMIC dataset.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fprot.25673&file=prot25673-sup-0002-AppendixS2.xlsx

Table A.3. The proportions of the number of hinge residues to the sequence lengths in the COSMIC dataset.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fprot.25673&file=prot25673-sup-0003-AppendixS3.xlsx

Table A.4. List of PDB coverage and interaction information for each protein in the COSMIC dataset.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fprot.25673&file=prot25673-sup-0004-AppendixS4.xlsx

Table A.5. Detailed list of genes with statistically significantly enriched hinge-mutation clusters in the TCGA dataset.

https://drive.google.com/file/d/1srykn-YIWLzXgq0t74Yrud9nVgRUGG54/view?usp=sharing

Table A.6. List of genes with statistically significantly enriched hinge-mutation clusters in the Rhapsody dataset

https://drive.google.com/file/d/1IcrudYwvvv3Z3Y7yPL0jTWDz-JRbwmjk/view?usp=sharing

Table A.7. Predictions on PTEN missense variants

https://drive.google.com/file/d/1eQB7D5-BPpTHkCz4LObY718HgTclkQJF/view?usp=sharing

Table A.8. Dataset used for PathDis

https://drive.google.com/file/d/1sk8OyAY5DbkOoRHwf0KdDSNYc-Izw3pH/view?usp=sharing