IMPROVING MOLECULAR-LEVEL PROTEIN DOCKING AND INTERPRETING SYSTEM-LEVEL CANCER MECHANISMS THROUGH MACHINE LEARNING

A Thesis

by

HAORAN CHEN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Yang Shen
Committee Members,	Aniruddha Datta
	Nicholas Duffield
	Sing-Hoi Sze
Head of Department,	Miroslav M. Begovic

August 2018

Major Subject: Electrical Engineering

Copyright 2018 Haoran Chen

ABSTRACT

Protein-protein interactions (PPIs) are crucial to cellular function, yet researchers still have much to discover about their mechanisms. At the molecular level, two new computational tools are proposed in this study to facilitate protein docking: 1. A regression model for predicting both the direction and extent of protein conformational change, especially the extent in this study, provides a new approach for structural ensemble generation and conformational sampling. 2. A classifier for assessing whether a protein pair is suitable for rigidity docking provides a method for performing a sanity check before uniformly applying rigidity assumption in protein docking.

At the intra-cellular system level, PPIs participate intensively in the propagation of mutational effects of cancer, which is well-known as a complex disease often derived from "driver genes" containing pathogenic mutations. Here I propose a new machine learning framework with biologically meaningful features to identify driver genes with the help of PPI network topology. Further interpretation of the machine learning model can help us understand cancer mechanisms by explaining the reason why cancer would prefer to attack those positions in network.

DEDICATION

To my family, my advisors, and my friends.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Yang Shen [advisor] and Professor Aniruddha Datta of the Department of Electrical and Computer Engineering, Professor Nicholas Duffield of the Department of Electrical and Computer Engineering and Professor Sing-Hoi Sze of the Department of Computer Science and Engineering.

All other work conducted for the thesis was completed by the student independently.

Funding sources

Graduate study was supported by National Science Foundation (NSF) CCF: EAGER: Dimension Reduction and Optimization Methods for Flexible Refinement of Protein Docking PI: Yang Shen

NOMENCLATURE

AUC	Area Under Curve
CAPRI	Critical Assessment of Prediction of Interactions
CGC	Cancer Gene Census
cNMA	Complex-based Normal Mode Analysis
ENM	Elastic Network Model
SGL	Sparse Group Lasso
NMA	Normal Mode Analysis
PC	Precision Recall
PPI(s)	Protein-Protein Interaction(s)
RF	Random Forest
RMSD	Root-Mean-Square-Deviation
RMSE	Root-Mean-Square-Error
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Statement of Problem

Our lives depend on proteins, which usually have the form of protein-protein interactions (PPIs) instead of acting alone. PPIs play essential roles in many cellular activities and processes [8, 9, 10], indicating the necessity to investigate how two proteins interact and what the final structure of protein complex is. Even though experimentalists discover protein structures to certain extent by methods such as X-ray [11], NMR spectroscopy [12], and recent Nobel winning topic cryo electron microscopy (cryo-EM) [13, 14], PPIs on genome-wide scale is still not experimentally available due to the limitation of resolution in experiment approaches aforementioned. Computational methods (i.e. protein docking) thereby step in generating tangible results [15, 16, 17].

However, recent CAPRI competitions[18, 19, 20, 21] reveal that protein docking is still a challenging task with many failures. The main reason is that proteins not only have flexibility when it acts alone (i.e. intrinsic flexibility), but also often adjust the conformation induced by its binding partner from unbound to bound state (i.e. induced fit) in order to exhibit certain functions. Protein structure between unbound and bound states could be remarkably different, which increases the difficulty of correctly modeling the structure of protein complex. Unlike the unbound-to-bound protein docking with well-defined 3D structures, the homology-to-bound docking is derived based on the similarity between the sequences of query protein and proteins with known structure, which makes homology-to-bound docking more difficult to dock and less physical-related. However, it is as meaningful due to the emergence of protein sequence data from next-generation sequencing technologies.

A well-accepted solution is rigidity assumption. Under the consideration of saving computational time, rigidity assumption performs divide-and-conquer on protein docking as a two-step process: Rigid-body docking, namely the flexibility of the focal proteins is neglected, is firstly implemented to rapidly obtain rigid but acceptable structures. Then the optimization algorithm,

which is the main problem, takes over to optimize the complex structure with better accuracy by flexible refinement with energy minimization criterion.

In spite of the challenging protein docking optimization problem to solve, one caveat buried beneath rigidity assumption is noticeable. The focal proteins which are actually flexible could be mistakenly treated as rigid and cause inaccuracy in terms of structure prediction. It is thus foreseeable that the rigidity assumption is not uniformly applicable in protein docking and should be tested individually on each protein complex.

The functionality of PPIs are not just limited in protein pairs, but further extended to larger scale. By systematically mapping PPIs towards proteome-scale, a system-level PPIs network emerges and carries diverse and crucial biological processes and properties [22]. Cancer, as complex and systemic disease, would not be satisfied and fully-functional by affecting just one pair of PPIs, but poison the cellular system with the help of existing PPIs network. In particular, PPIs network frequently participate in propagating cancer mutational effect from 'driver genes', which contains pathogenic mutations as origins of cancer [23], during the mutagenesis. To understand how cancer works in PPIs network and further use the understanding to counterattack, it is imperative to not just identify the driver genes in the network, but ulteriorly dig up mechanisms operated by cancer. In other words, where in the network would be preferably attacked by cancer and why cancer would attack these position are two questions demanding answer.

1.2 Literature Review

1.2.1 Prediction of the extent of conformational change

Many physical-driven frameworks have been researching the protein unbound-to-bound conformation changes. Monte Carlo (MC) simulation was applied to simulate thermal conformational fluctuations with protein structure parameters [24]. Molecular dynamics (MD) is a well-studied method to simulate the movements of protein molecules through a long period of time [25]. Normal mode analysis (NMA) is a method that decomposes the fluctuation of protein into different normal modes via the Hessian matrix of displacement and use slowest modes to predict the direc-

tions of conformational changes [26, 27, 28]. These frameworks are not only applied on individual proteins, but also on protein complexes. Comparing to other methods having enriched physical details, normal mode analysis usually adopts coarse-grained models [29], to save computational burden tremendously while still assures good prediction accuracy.

1.2.2 Rigidity Assumption on Protein-Protein Interactions

The dilemma of docking accuracy vs computational complexity have formerly been attempted to address. AutoDock4 framework allows limited and preassigned flexible side-chain under rigidity assumption for both receptor and ligand prior to optimization [30]. Similarly, Wang et al. developed a docking method to incorporates explicit backbone flexibility [31]. Moreover, Gray et al. applied Monte Carlo search for rigid-body docking and Monte Carlo minimization for backbone displacement and side-chains conformation simultaneously [32]. This collection of solutions follow the idea that is essentially to embrace the protein flexibility to controllable extent based on the rigidity assumption, which could increase the accuracy and avoid computational complexity at the same time if the computational model is designed and tuned well. Here a novel angle for solving the dilemma is being offered: instead of universally applying certain algorithm or framework to all query protein with rigidity assumption, one should classify query proteins beforehand into two different categories: rigid and flexible, and then implement different algorithm accordingly. By this new direction, one could save much runtime by not deploying over-complicated algorithm on those rigid and easy protein complexes, and precisely recognize the flexible and difficult ones to ensure the accurate results with advanced docking methods.

1.2.3 Cancer Mutational Effects to Networks of Protein-Protein Interactions

There are two main directions by which researches are trying to understand cancer mechanisms: data-driven and principle-driven approaches. The data-driven approach often adopt statistical or prediction methods to identify cancer driver genes in PPI network. Pan-cancer analysis comprehensively pinpoints cancer driver genes based on statistical analysis by omics data [33, 34]. Standing on the should of pan-cancer analysis, network propagation methods well utilized PPI network as

the scaffold to propagate the mutational effects from the disease genes in order to identify driver genes with the extra help of network topology [35]. Whereas utilizing data from a wide range of cellular levels could provide more information and accuracy for identifying cancer driver genes, but it would hardly be able to explain why those genes are doomed to carry the original mutations of cancer due to the lack of interpretability. Centrality prediction [36, 37] further adopt network topological features to offer explanatory classifier, whereas global topological features are hard to be further interpreted. On the other hand, principle-driven approaches mainly represented by ordinary differential equation (ODE) [38, 39] applies stoichiometric analysis could precisely simulate how cancer propagates its mutational effects through local PPI network by monitoring the change of protein concentration over time. By simulating the dynamics of local PPI, ODE supposedly provides explanation of why cancer attacks certain positions in PPI network [40]. However, such approach inevitably requires massive amount of kinetic and concentration data as support of the framework of principle to ensure the simulation is numerically accurate and biologically meaningful. The larger the network is, the harder the experimentalist could economically obtain those data, not even mention the data from the mutant type. Therefore, ODE has a narrow range of applicability.

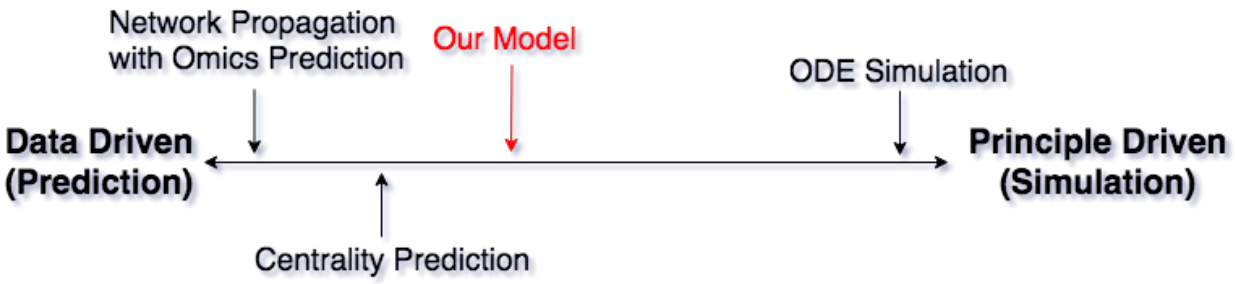


Figure 1.1: State-of-the-art Models for Cancer Mechanism Discovery.

1.3 Objective of Study

At protein docking level:

- Predicting the extent of protein conformational change.
- Examining and assessing rigidity assumption of protein docking.

At PPI network system level:

- Predicting cancer driver genes by network topology.
- Understanding cancer mechanisms by extracting rules from machine learning model.

Toward the goals above, a data-driven machine learning approach is developed to predict the extent of conformational change based on complex-based normal mode analysis (cNMA) framework [41]. A workable pipeline for examining and assessing rigidity assumption of protein docking is built through the conformational sampling, protein energy calculation, and classification model. An identification framework based on machine learning and network topology is constructed and a rule extraction scheme is designed. The common ground between two levels of work is mentionable. They both utilize network data (atomic network and PPI network, respectively) and subsequently try to perform reasoning in virtue of the power of machine learning.

1.4 Organization of Thesis

- Section I Introduction: This section investigates what problems we are solving, where these problems come from, and why they are significant and worthy to solve. Moreover, a detailed literature review with study objectives corresponding to each problem , respectively, is presented in this section.
- Section II Materials and Methodologies: This section explicitly introduces the utilized materials and adopted methodologies, especially the path of developing these methods, for solving the aforementioned problems.
- Section III Results and Discussion: All results from each developed models are laid out in this section. The discussion focuses on how these results answer the questions proposed in Section I.

- Section IV Conclusion and Future Directions: This section summarizes the conclusions derived from the results and discussion above and also points out the next step of research.

2. MATERIALS AND METHODOLOGIES

2.1 Predicting Protein Conformational Change *

2.1.1 Benchmark Set and Rigid-body Docking

Machine learning model, especially supervised learning, requires a dataset with true label for training. In this study, Protein-Protein Docking Benchmark Version 4.0 dataset [42] was used, which contains 176 representative PPIs with known unbound and bound structures. Due to little conformational change between unbound and bound states, 12 monomers, 6 PPIs were removed from the dataset, leaving 340 monomers, 170 PPIs in total remained.

Subsequently, 10 encounter complexes were generated for each PPI in remaining benchmark set by utilizing ZDOCK rigid-body docking software [43, 44]. ZDOCK algorithm is essentially a scoring function for ranking cluster centers by iterative search. Structures with the top 10 scores are regarded as acceptable rigid-body docking results. We further keep structures with complex interface root-mean-square-deviation less than 10 Å as reliable complexes for feature engineering and machine learning in the next step. Besides the benchmark set, we obtained CAPRI set data with 11 unbound and 19 homology structures for testing.

The "label" of each PPI is represented by the root-mean-square-deviation (RMSD) between two structures, which directly reflect the extent conformational change, derived as below:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

where δ_i stands for the distance of i^{th} heavy atom (i.e. C, N, or O) between unbound and bound states. Note that there are two kinds of RMSD utilized as label: the whole RMSD which is the geometric difference between two whole structures, and iRMSD which is the displacement between the interface of two structures.

*Reprinted with permission from [1].

2.1.2 cNMA Framework

Complex-based normal mode analysis (cNMA) [41] provides a model considering both intra- and intermolecular potentials based on elastic network model (ENM) [45]. All 170 PPIs are projected and processed in cNMA framework in order to extract physically meaningful features. Specifically, under normal mode analysis (NMA) framework (introduced in Section I), the prediction of conformational change is based on eigenvalue λ_i of Hessian matrix derived from ENM, where i is the i^{th} normal mode. The process of calculating eigenvalues and eigenvectors of Hessian matrix includes setting concurrent but differentiated intra- and intermolecular potentials in an ENM, projecting the complex-derived Hessian matrix H away from the space of rigid-body motions for the monomer protein under study, calculating eigenvalues and eigenvectors of the projected Hessian, extracting the components of eigenvectors (nontrivial normal modes) corresponding to the monomer under study,rescaling the eigenvalues accordingly, and re-ranking the rescaled eigenvalues (ascending positive scalars k^i... k^{3N-6} where N is the number of residues for the monomer) and corresponding eigenvectors (column vectors v_i of dimensionality $3N$) [1, 41].

2.1.3 Entropy-inspired Features

Conventionally, the extent of conformational change can be expressed as below [28]:

$$RMSD \approx \sqrt{\frac{1}{N}\frac{k_B T}{m}\sum_{i=1}^{3N-6}\frac{1}{\omega_i^2}} \propto \sqrt{\frac{1}{N}\sum_{i=1}^{3N-6}\frac{1}{\omega_i^2}} \propto \sqrt{\frac{1}{N}\sum_{i=1}^{3N-6}\frac{1}{\lambda_i}}$$

where k_B is the Boltzmann constant, T is the absolute temperature, m is the mass for all atoms, ω_i is the frequency of i^{th} mode, λ_i is the i^{th} eigenvalue. This RMSD approximation approach regards the squared magnitude of an atom fluctuating along a normal mode i is proportional to the inverse of i^{th} eigenvalue in linear combination "sum-type" manner [1], illustrated in Figure 2.1A. Therefore, the conventional "sum-type" feature can be expressed as below:

$$\Phi^{sum} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\frac{1}{\lambda_i}}$$

where K represents the first K^{th} slowest normal modes. The concern of computational speed is the reason that we did not use all normal modes.

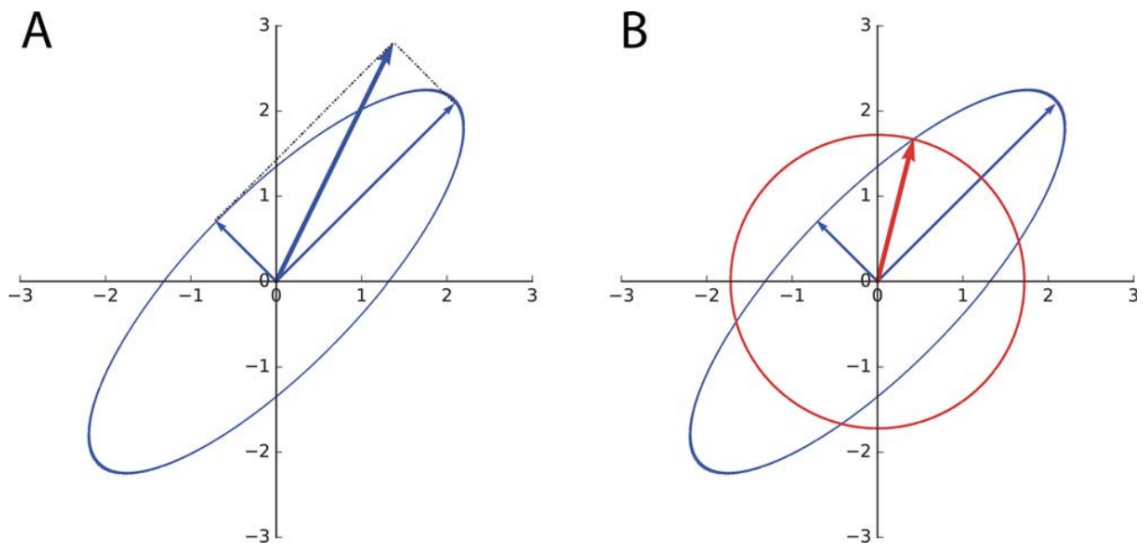


Figure 2.1: Geometric illustration of sum-type and novel physical-inspired product-type features calculated from normal mode frequencies (eigenvalues) in 2-dimensional space. The ellipsoid represents the space of all possible deformations along two normal modes to certain extent and its semi-axes are proportional to. A. When deformation extents along different normal modes are regarded independent, the extent of resulting combination (thicker blue arrow) would simply be a sum of two orthogonal vectors. B. When a linear combination of normal modes is constrained by preserving the volume of the conformational space (or vibrational entropy), its average extent (thicker red arrow) cannot go beyond the radius of a sphere of the same volume. (Reprinted from [1])

In this study, a novel "product-type" feature inspired by vibrational entropy was proposed as in Figure 2.1B. The ellipsoid represents the space of all possible deformations to certain extent along two normal modes and its semi-axes are proportional to $\frac{1}{\sqrt{\lambda_i}}$. When deformation extents along different normal modes are regarded independent, the extent of resulting combination (thicker blue arrow) would simply be a sum of two orthogonal vectors [1], which leads to the linear combination "sub-type" form [1]. However, when a linear combination of normal modes is constrained by preserving the volume of the conformational space (or vibrational entropy), its average extent

(thicker red arrow) cannot go beyond the radius of a sphere of the same volume [1], which leads to expression as following :

$$RMSD \propto \sqrt[3N-6]{\prod_{i=1}^{3N-6} \frac{1}{\sqrt{\lambda_i}}}$$

Therefore, the "product-type" features can be expressed as below:

$$\Phi^{prod} = \sqrt[K]{\prod_{i=1}^{K} \frac{1}{\sqrt{\lambda_i}}}$$

Based on the principle-driven "product-type" feature, we have derive three machine learning features by three different cutoffs in terms of the choice of K: absolute cutoff, rigidity-related cutoff, and size-related cutoff. The absolute cutoff adopts universal K across all encounter complexes. The rigidity-related cutoff uses fixed ratio between K^{th} eigenvalue and the first one across all encounter complexes. The size-related cutoff uses a fixed fraction of protein sizes across all encounter complexes. Additionally, the size of protein N is also used as feature.

2.1.4 Regression

Given principle-driven features and true conformational change labels above, a regression model should be trained accordingly. The benchmark set was divided into 80% training set and 20% held-out test set. A 4-fold cross-validation has performed on training set to avoid overfitting issue while grid search has been applied to optimize the hyper-parameters (e.g. regularization parameter). All features are standardized across all encounter complexes and all labels are centered before learning process. Both linear and RBF kernel versions of ridge, LASSO and elastic net regression (equal L_1 and L_2 penalty) are trained.

2.2 Examining Rigidity Assumption Protein Docking

2.2.1 Pipeline

A pipeline for examining and assessing rigidity assumption has been proposed in Figure 2.2. This pipeline is built on the model and results from last section, predicting the extent of protein

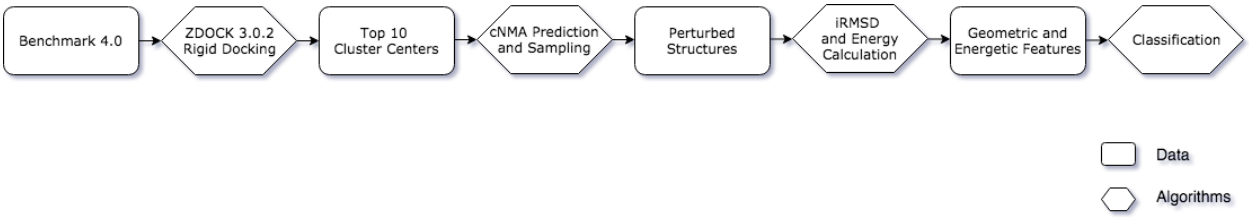


Figure 2.2: Pipeline of Assessing Rigidity Assumption.

conformational change. Therefore, the steps before "cNMA Prediction and Sampling" are exactly the same as in last section. The rest of steps, especially the algorithms, are proposed in Figure 2.2.

2.2.2 Sampling on Benchmark Set

The trained model for predicting the extent of protein conformational change has been applied on each monomer to obtain predictive conformational change. Accordingly, a random sampling function from ProDy [46, 47], a software for protein dynamic analysis built in Python, was applied to generate a ensemble conformations containing 50 perturbed structures for all 3400 monomers. The sampling strategy is based on Gaussian distribution which is for controlling the direction of protein conformational change. The extent, on the other hand, is set by the predicted RMSD.

2.2.3 iRMSD and Energy Calculation

By assuming every protein pair could be regarded as rigid body, the ones who are actually flexible would hypothetically endure energy permissiveness or interface distortion during docking, therefore the geometric and energetic features could be able to differentiate rigid and flexible proteins by the proof of contradiction. Accordingly, each structure will be calculated its energetic and geometric difference with original unbound structure before the sampling and perturbation to construct features.

Specifically, three enthalpy terms (electrostatics, Solvent Accessible Surface Area (SASA), Van Del Waals (VDWs)) were generated by CHARMM [48], a program for macromolecular energy, minimization, and dynamics calculations. The SASA and VDWs energy between two structures were easily calculated by CHARMM in one step. The electrostatics, however, are much more com-

plicated. One entropy term was obtained from the counting of the number of micro-states around +5 Kcal/mol of the lowest-energy structure among 50 perturbed structures for every encounter complexes.

One putative interface RMSD term will be estimated based on cNMA framework. The putative interface is defined for each perturbed structure based on the criterion of 6 Å, meaning that any heavy atoms are within 6 Å of its counterpart are regarded as interface. Then the putative interface RMSD can be calculated by aligning the perturbed and unbound structures.

2.2.4 Feature Compression

A feature compression procedure has been performed in order to obtain the final feature matrix from the massive feature space on the ensemble of perturbed structures. Each ensemble (e.g. 1A2K_1) has 50 receptor perturbations and 50 ligand perturbations. At "extend" step, The average, 1st quartile average, 2nd quartile average and 3rd quartile average are calculated for each feature across each ensemble (e.g. 1A2K_1) to generate only one row of features for one ensemble. At the "average" step, ensemble features are averaged to generate final row of feature for each protein complex (e.g. 1A2K). This compression process removes the redundancy from perturbations while extracting the useful information from the distribution of each ensemble.

2.2.5 Classification

Given pre-defined rigid/flexible classes from protein benchmark set 4.0 [42], multiple classifiers are trained with energetic and geometric features, including SVM, logistic regression and ensemble learning methods such as Random Forest and XGBoost. The preprocessing of feature and label, the partition of held-out, and the cross-validation and grid search are all identical to 2.1.4.

2.3 Predicting Cancer Driver Genes

2.3.1 Cancer Gene Census and PPI Network

Cancer Gene Census (CGC) database [49] is the biggest database which catalogues genes containing pathogenic mutations causally implicated in cancer. In other words, CGC could provide

Dataset	Source	Description
Network	Curated network (from Dr. Edwin Wang's Lab)	6302 proteins (nodes), 63000+ interactions (edges)
Driver Gene	Cancer Gene Census (CGC)	752 in total, 400 mapped to curated network

Table 2.1: Cancer Gene Census (CGC) Drive Gene Data and Curated PPI Network Data

the identification of known drive genes obtained from paper curation and experiments. There are 752 cancer driver genes in CGC database across 12 tumor types.

PPI data utilized in this study is the curated data from Edwin Wang's Lab. Unlike most of PPI databases containing predicted PPI data by high-throughput methods, curated data is better at enhance the reliability. Inevitably, there is a trade-off between the scale of network and reliability, namely a more reliable network must have smaller scale. To make sure the interpretation in the next step is biologically meaningful, we choose the network more reliable. Curated PPI data has 6302 nodes, which are proteins, and 63000+ edges, which are interactions. Three kinds of descriptions are annotated on each edge: positive, negative, and physical, which stand for activation, suppression, and binding between two proteins, respectively. The enrichment of physical details of curated PPI network further facilitate the interpretability. The details of CGC and PPI data are listed in Table 2.1

2.3.2 Topological Feature: Network Motif

In this study, there are two criteria for determining the choice of features: 1. The features should be able to differentiate driver genes from others. 2. The features should carry biological meaning for further interpretation based on machine learning model. In the light of these two criteria, we chose network motif as our features for machine learning. Network motif is recurrent and statistically significant sub-graphs or patterns in a specific network, which is important local property. We hypothesized that driver gene has distinguishable local environments represented by network motif comparing to other genes. Meanwhile, network motif not only possess statistical significance in network, but also equips the biological function. For example, Figure 2.3

and 2.4 altogether present the biological function of a three-node motif as sign-sensitive delay in PPI network [2], which could show no response under a short-term noise signal input but accurately response under a long-term actual input. Such biological function gives network motif with biological interpretability making it a solid choice of machine learning feature.

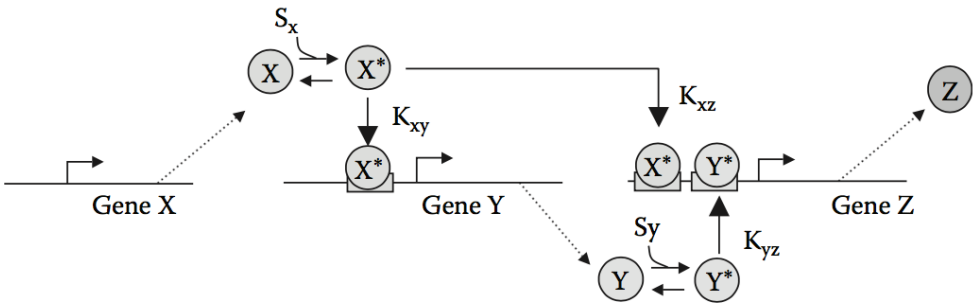


Figure 2.3: A example of network motif's biological function: gene X, Y, Z are all transcription factors which control the expression of genes by binding to the DNA of corresponding genes. Circle X, Y, Z is the protein product of gene X, Y, Z, respectively. Circle X* is the active state of X. S_x and S_y are upstream signal to convert protein X and Y from inactive state to active state, respectively. Protein X* can control the expression of gene Y to protein Y, whereas protein Z can be generated only if protein X* and Y* are binded to gene Z. K is the binding rate of its corresponding protein and DNA. (Reprinted from [2])

2.3.3 Network Motif Finding

MFINDER software [50] has been applied to find statistically significant subgraphs, namely network motif. MFINDER essentially employed enumeration to calculate the significance, which is slow but accurate. Specifically, MFINDER calculates and record all subgraphs in the real network. Each subgraph is annotated by a unique network ID as Figure 2.5 which is converted from the adjacency matrix of subgraph. The final decimal ID is reversely converted by the binary sequence lay out from the adjacency matrix. Then MFINDER will generate randomized network by switching edges in the real network (Figure 2.6 [3]). The number of generated randomized networks can be controlled by input parameter of software. The P-value and Z-score are accordingly calculated based on the counting difference between real network and randomized network

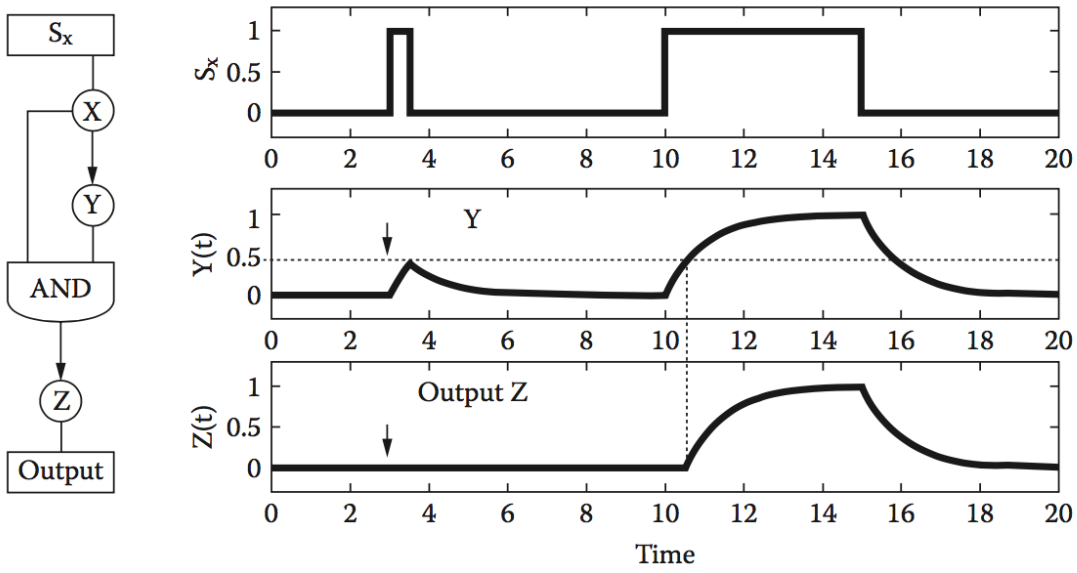


Figure 2.4: A example of network motif's biological function (cont.): Figure 2.3 can be compactly illustrated by the left figure. The right figure demonstrates the biological function of this three-node motif. If the input signal S_x is short-term noise. The concentration of protein Y, Y(t) has not reached the threshold to produce protein Z, so there is no output from this system. If the input is long-term actual command from upstream, Y(t) will reach the threshold and start generating Z. If the signal stops, the output will also vanish immediately without delay. Therefore, this three-node motif can filter out noise and also be sensitive to the switch of input signal. (Reprinted from [2])

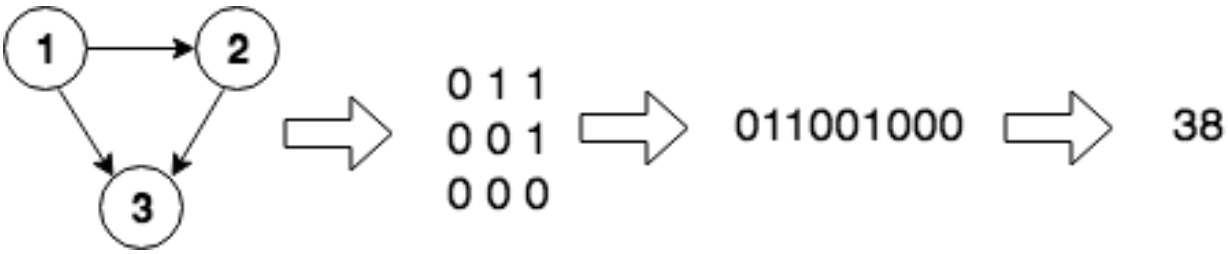


Figure 2.5: The Conversion of Motif ID: The first arrow leads to the adjacency matrix of this three-node motif. The second arrow layouts the adjacency matrix row-by-row. The third arrow represents the transform between binary sequence and decimal code. Note that the binary sequence should be read reversely. In this case, it should be 100110 as 38.

for each subgraph as below:

$$P(m) = \frac{1}{N}\sum_{n=1}^{N} c(n)$$

$$Z(m) = \frac{f_G(m) - average(f_R(m))}{std(f_R(m))}$$

where m is m^{th} subgraph, N is the total number of randomized network, $f()$ is the counting of corresponding subgraph in the network, G represents the real network and R is the randomized network. The criteria for a subgraph being network motif is $P(m) < 0.01$ or $Z(m) > 2$.

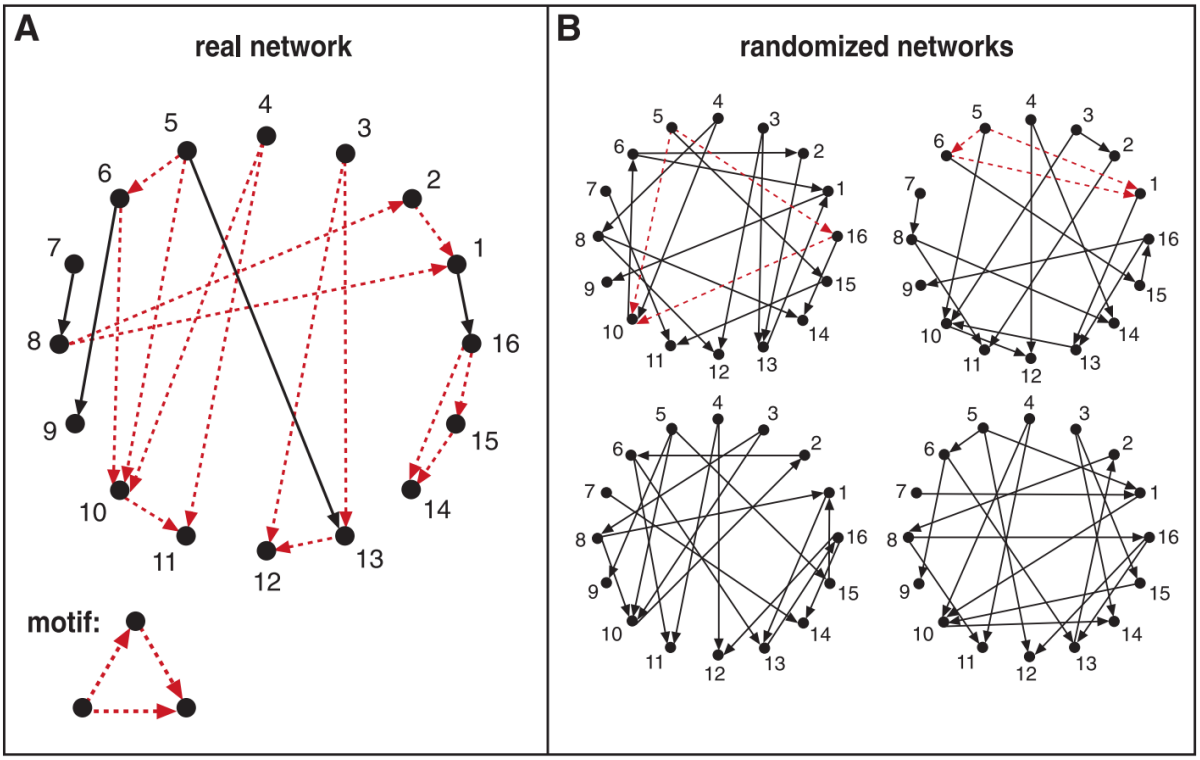


Figure 2.6: Finding Network Motif by MFINDER algorithm: Panel A illustrates the real network with its subgraphs. The red dashed lines constitute the network motif. Panel B illustrates the randomized network. It is obvious that the frequency of network motifs in randomized networks are much less than real network, which is why this specific subgraph is qualified as network motif. (Reprinted from [3])

One limitation of MFINDER is that it cannot handle edges with different physical meaning such as activation, suppression, and physical binding provided in curated PPI network. We have to regard activation and suppression as the same and identify them afterwards. The physical binding is converted into bi-directed edge and mixed-up with very few real bi-directed edges. The output

motifs are further processed as in Figure 2.7 to illustrate the specific biological function of each motif, namely the motif subtype.

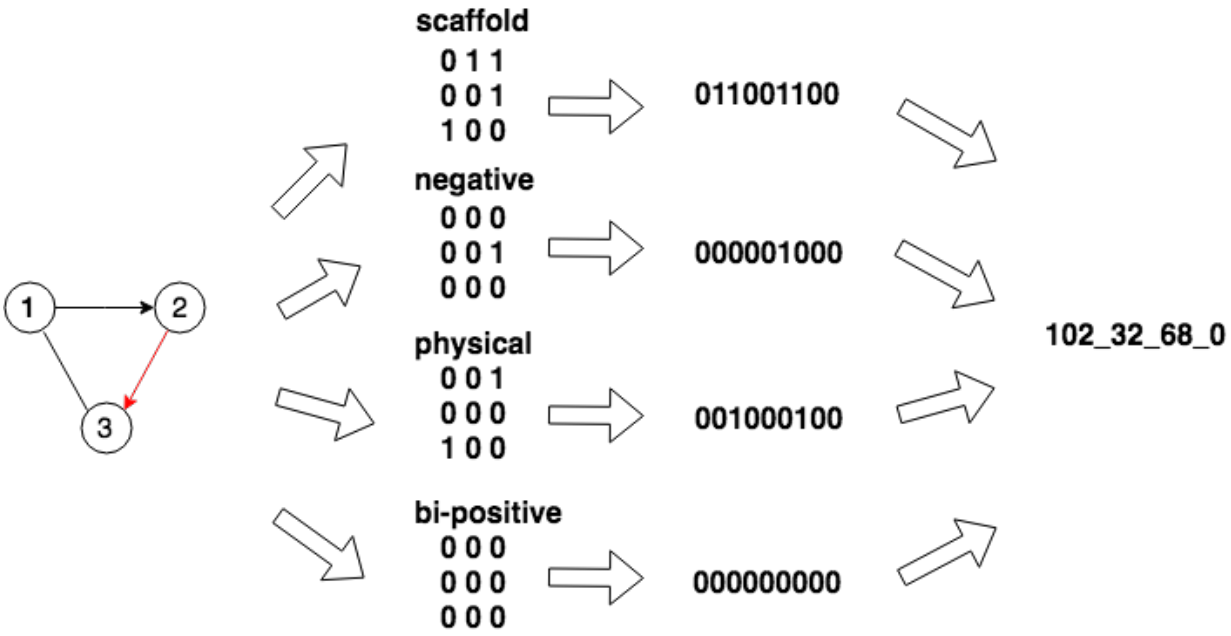


Figure 2.7: The Conversion of Motif subtype ID: The red arrow from node 2 to node 3 stands for the negative arc (suppression). The edge between node 1 and 3 stands for the physical interaction. In 38_32_68_0, the first digit represents the scaffold of motif. The second digit represents the negative arcs in the motif. The third digit represents the physical interactions in the motif. The four digit represents the bi-positive reactions in the motif.

After identifying and locating all network motifs in PPI network, a feature matrix with one-hot encoding features as Figure 2.8 is extracted. Each row is a gene with Entrez Gene ID. Each column is a network motif with subtype annotation. The subtype annotation express the way to take advantages of the physical details provided by curated PPI dataset, which is presented in the rightest of Figure 2.7. As an example, the first digit ”38” in motif ID ”38_32_0_0” is the motif ID presenting the basic interaction of motif. The combined rest of digits is called "subtype_ID". The second digit is to separate the negative arc, namely the suppression, in this kind of motif. The third digit encodes physical interactions in bi-directed edge. And the fourth digit is positive edges in bi-directed edges. The basic scaffold of motif is neglected and only corresponding kind of edges

are considered when subtype_ID is being calculated. Each element is a binary value recording the existence or not of corresponding gene in corresponding network motif subtype.

	row.names	38_0_0_0	38_32_0_0	38_2_0_0	38_4_0_0	38_6_0_0
1	7157	1	1	1	1	1
2	940	1	1	1	1	0
3	2006	0	0	0	0	0
4	90	1	0	0	0	0

Figure 2.8: One-hot Encoding Feature Matrix: Each row is gene in Entrez ID, each column is a motif subtype, each element is whether this gene is in this motif ever

2.3.4 Statistical Analysis for Motif Hotspots

Network motifs with biological function is an interpretable extension of driver genes. Here we try to find out the "driver motifs", namely the motif hotspots that would be cancer's favorable attacking points by dint of statistical analysis. The algorithm is inspired by the hotspot finding in amino acid sequence [51], which basically identifies the hotspot by comparing real mutations and the distribution of randomized mutations. The pseudocode of statistical analysis for motif hotspots is in Algorithm 1. The example of the feature matrix calculated from function FEATURE_CALC is in Figure 2.9.

	row.names	38_0_0_0	38_32_0_0	38_2_0_0	38_4_0_0	38_6_0_0
1	6774	424	54	58	33	18
2	3725	392	99	52	67	25
3	5579	209	27	24	6	3
4	6714	445	52	104	40	24
5	4851	22	9	14	4	9

Figure 2.9: Number of Occurrence Feature Matrix: Each row is gene in Entrez ID, each column is a motif subtype, each element is how many times a gene is in a motif subtype

Algorithm 1 Statistical Analysis for Motif Hotspots

function RANDOMIZATION
    num_driver ← number of driver genes
    network ← RANDOM_ASSIGNMENT(*num_driver*) ▷ Randomly assign same number of nodes to be driver genes as random background
    return $network$
function FEATURE_CALC
    for `i = 1 to all genes` **do**
        for `j = 1 to all motifs` **do**
            $feature\_matrix[i, j]$ ← number of occurrence gene i in motif j.
    return $feature\_matrix$
procedure MAIN
loop:
    real_features ← **Feature_calc**(*real_network*)
    num_driver ← number of driver genes
    randomized_network ← RANDOMIZATION(*num_driver*)
    randomized_features ← FEATURE_CALC(*randomized_network*)
    distribution ← APPEND_ROW(*randomized_features*) ▷ compress *randomized_features* to a row and append it to *distribution* matrix
end of loop
    for `j = 1 to all motifs` **do**
        if *real_network[j]* > 99% of *distribution*[:, j] **then**
            motif_hotspot[j] ← $True$
        else
            motif_hotspot[j] ← $False$

2.3.5 Classification

One major issue appeared during the classification is the imbalanced data. There are 6302 nodes in the whole curated PPI network and 3875 remained after MFINDER motif finding process. Only 400 of them are catalogued as driver gene by CGC database. In machine learning perspective, the positive class has only 400 samples but the negative class has more than 3000 samples. This skew distribution of label will cause the model hard to distinguish the minority class. To solve this imbalanced data issue, we applied SMOTE algorithm [52] to generate synthetic minority data. Specifically, SMOTE algorithm adopts K-nearest-neighbor (KNN) method to oversampling the minority class and use sampling with replacement to downsampling the majority class. The detail

process is illustrated in Figure.

We randomly selected 20% of dataset as held-out test data. The remaining 80% are as training set processed by SMOTE algorithm. The new generated dataset contains class one samples and class zero samples. A 4-hold classification with grid search was performed on the training set similar to 2.1.4 and 2.2.5.

2.4 Understanding Mutational Propagation

2.4.1 Rule Extraction from Random Forest

Random Forest (RF) [53] is an ensemble algorithm taking majority vote from each decision tree to determine the final category for each sample, which is naturally suitable for interpreting crucial rules from the tree ensemble. Mashayekhi et al. [4] recently proposed a framework combining RF and sparse group LASSO to pinpoint the important rules. The gist of this framework is to take advantages of the power of SGL's internal feature selection to filter out rules that contribute none or little during the RF learning process. The pseudocode is presented in Figure 2.10. In the first step, all rules in trained model have been collected. Note that each leaf of each decision tree corresponds to one and only one rule. Step 2 basically performs scanning procedure to check if a sample is satisfied a rule. This step projects the original feature space to new rule feature space. After this step, each row of feature matrix is still a gene but each column is a rule. The grouping function at the end of step 2 is essential. It naturally groups rules from the same decision tree together for the sparse group LASSO due to the fact that all rules from the same tree are in the same randomized feature space. SGL are employed in step 3 for removing non-relevant rules under model training. The last step is to test the performance of the model. Note that there would be worse performance in testing than the original Random Forest due to using fewer but important rules than original trivial rule set.

The motifs that participate in the remaining rules are the candidate motifs as hotspots. After cross-validating with the results from statistical analysis, we could find the final motif hotspots to answer "where" question.

Algorithm 2 RF+SGL

```
Input: trainingSet, testSet, treeNo
Step 1: // construct Random Forest
RF = trainRF(trainingSet, treeNo);
Rs = generateRules(RF);

Step 2: //compute rules coverage
m = size(trainingSet);
n = size(Rs);
RsCoverage=zeros(m, n);
foreach sample in trainingSet
      foreach rule in Rs
            if match(rule, sample)
                  RsCoverage(sample, rule) = 1;
            end if
      end for
end for
Grouping = group(RsCoverage);

Step 3: //use sparse group lasso
Fit = cvSGL(RsCoverage, Grouping)
bestLambda = min(cvFit.lldiff)
Fit = SGL(RsCoverage, Grouping, bestLambda)

Step 4: // evaluate the performance on test set
pred = predictSGL(Fit, testSet)

return errorRate, numberOfSelectedRules
```

Figure 2.10: Sparse Group LASSO for Rule Extraction from Random Forest: Step 1: Training RF model and collecting all rules. Step 2: Screening all rules over all samples and recording satisfaction. Step 3: Fit new sample-rule features matrix with groups. Step 4: Testing. (Reprinted from [4])

2.4.2 Interpretation of Cancer Mechanisms

It is obviously not enough to just answer "where" with motif hotspots, we still need to answer "why". We suspect that the reason why cancer preferably attack driver genes is because they are in network motifs with specific function which cancer need to hijack or damage. But what function exactly? Can it be validated? We firstly design scoring function to rank the importance of each motif subtype. Specifically, the importance of a motif subtype in a rule can be measured by the position its appearance in the rule since the RF algorithm should choose the variable which can maximumly decrease the gini coefficient at each split point. We discretize the length of each rule based on the quartile. Motifs at 1st quartile have 4 points, 2nd quartile have 3 point and so forth. At the end, we sum each motif across all rules to get the final score. Then we drew motifs with top 50 score and use specific biologically functional motif to validate the correctness of motif hotspots.

```
[1] "102_0_68_0 > 0.00125430582556873 ; 110_40_78_0 < 0.003015969065018 ; 108_12_
68_0 < 0.00153328431770205 ; 110_0_68_14 < 0.000652368762530385 ; 238_0_78_160 >
0.00125430582556873 ; 102_64_68_0 > 0.00125430582556873"
[2] "1"
```

Figure 2.11: Example of Rule: First entity is the specific rule. The second entity is the class of this rule. The reason why the split points are not integer 0 or 1 is because the synthetic data generated by SMOTE is float number.

3. RESULTS AND DISCUSSION

3.1 Prediction of the Extent of Protein Conformational Change*

3.1.1 Assessment of Features

Before performing the machine learning process, we firstly assess the quality of the product-type features we developed, and benchmarked them with previously published sum-type features. The results are shown in Figure 3.1. The top panel utilizes the eigenvalues data from canonical NMA analysis for all monomer proteins in benchmark set, whereas the bottom panel is based on cNMA framework for protein complexes which have iRMSD < 10 Å. Three cutoff methods introduced in 2.1.3 correspond to left, middle, and right column, respectively. By increasing absolute number K, size-related fraction η of each protein, and rigidity-related threshold M which is the multiple of the smallest eigenvalue corresponding to each encounter complex, the Pearson correlation coefficients between the extent of protein conformational changes RMSD and are drawn. Our novel product-type features are presented by solid lines and sum-type features by dashed lines.

Under canonical NMA platform, there is no clear differentiation between two types of features. However, in cNMA framework for encounter complex, the product-type features obviously outperform sum-type features in terms of the correlation with RMSD. Although it is not as better as other two methods for rigidity-related cutoff, but it still higher before 500. Based on the quality assessment, we chose $K = 100$, $\eta = 60\%$, and $M = 100$ as official cutoffs to derive product-type features.

3.1.2 Extent of Protein Conformational Changes

By entering product-type features which are derived based on the cutoff obtained from last section, we test the performance of various machine learning models. The details of regression models are introduced in 2.1.4.

Figure 3.2 and Figure 3.3 present the results from machine learning models for RMSD predic-

*Reprinted with permission from [1].

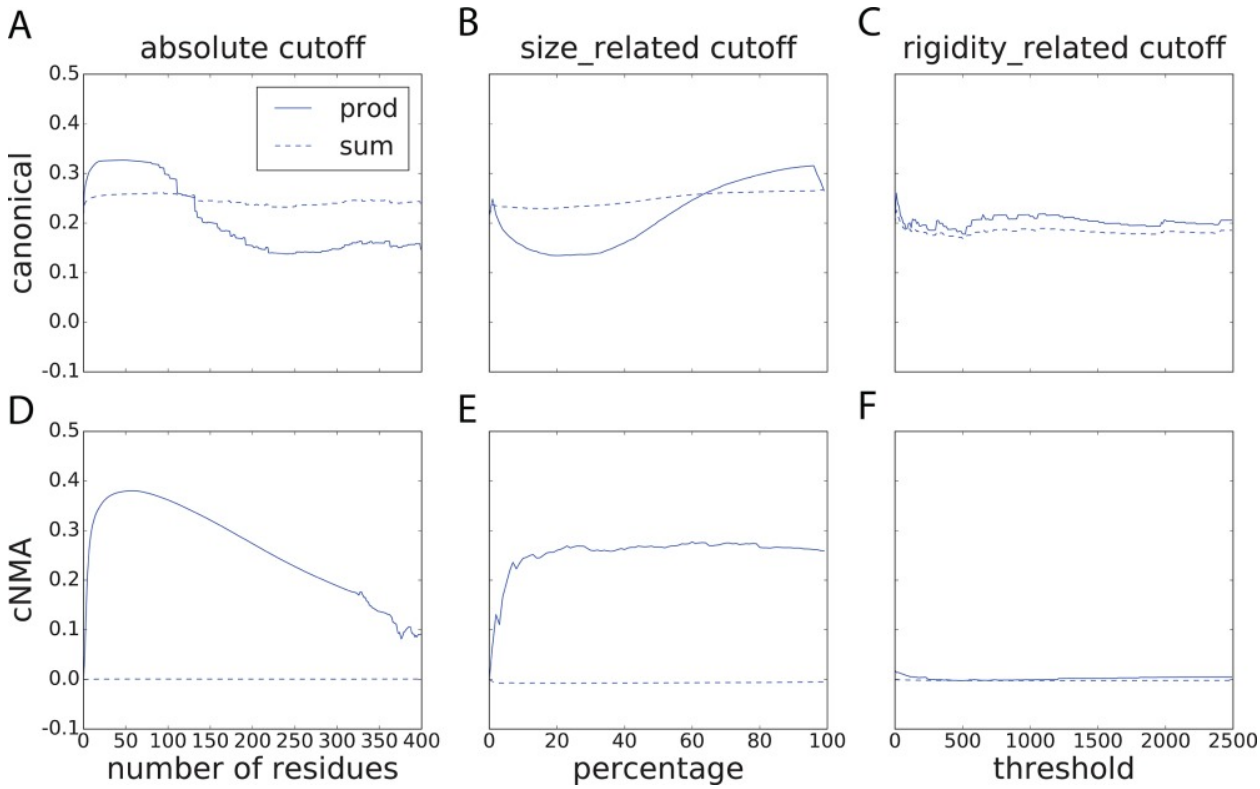


Figure 3.1: Pearson correlation coefficients between the extent of conformational changes and features calculated from conventional NMA on individual monomers (top panel) or those calculated from cNMA on monomers in encounter complexes (bottom panel). The cutoffs used were absolute value K, size-related fraction g, and rigidity-related cutoff M, respectively, from left to right. Sum-type features are shown in dashed lines and product-type ones in solid lines. (Reprinted from [1])

tion and iRMSD prediction, respectively. LASSO, Ridge, and Elastic Net regression were applied to both prediction tasks with linear or RBF kernel. ρ represents the Pearson correlation coefficient, RMSE is the root-mean-square-deviation between overall predicted RMSDs and real RMSDs, α and γ are regularization parameter and kernel parameter, respectively, σ is the standard deviation of predicted RMSDs. Based on these measures, the results show that the performance of RBF kernel for all prediction are better than corresponding linear ones, which reveals that the features are nonlinear in original feature space. Across three regression methods with nonlinear RBF kernel, ridge regression outperformed others. The reason is that either LASSO or Elastic Net would eliminate features based on $L1$ penalty, given the fact that we only have four features. Therefore, the

ridge regression utilizing all four features with more information undoubtedly has the best performance. The results from CAPRI test set are in Figure 3.4. The performance from each sub-figure is below 1 Å , except the bottom left figure, RMSD on homology-to-bound prediction, with very bad starting points (close to 8Å) from rigidity docking.

3.2 Examining Rigidity Assumption

We examine rigidity assumption by applying two types of kernelize classifier: kernel logistic regression and kernel SVM, both with RBF kernel, and two ensemble learning methods: Random Forest and popular gradient boosting method XGBoost. The results is presented in Figure 3.5. Random Forest clearly shows the better performance with AUCROC = 0.7667, which is a promising classification to differentiate rigid and flexible proteins.

3.3 Prediction of Driver Genes

Given one-hot encoding gene-motif feature matrix with SMOTE-generated synthetic data (introduced in 2.3.2), Random Forest algorithm has been applied to train the classifier with 4-fold cross-validation. In order to benchmark the performance of our features, centrality features used by Cui et al. was also obtained from the exact same training set. Since our main task is to identify the driver gene, it is apparently more important to recognize the positive samples than differentiate the two classes. That is why we drew precision-recall (PC) curve besides ROC curve. The only different between PC and ROC curve is that PC use precision, namely positive predictive value (PPV), instead of false positive rate (FPR). The PPC describes how many selected items are relevant so that it could directly indicate the ability of a classifier to recognize the positive class. The results of ROC and PC curves from RF are shown in Figure 3.6. The classifier based on network motif features (blue curves) performs much better than classifier based on centrality features (yellow) curves in either ROC or PC measurement. The better performance from our novel features validate the choice of network motifs as promising features and also ensure the meaningful interpretation, while providing a reliable tool to identify cancer driver genes based on PPI network topology data.

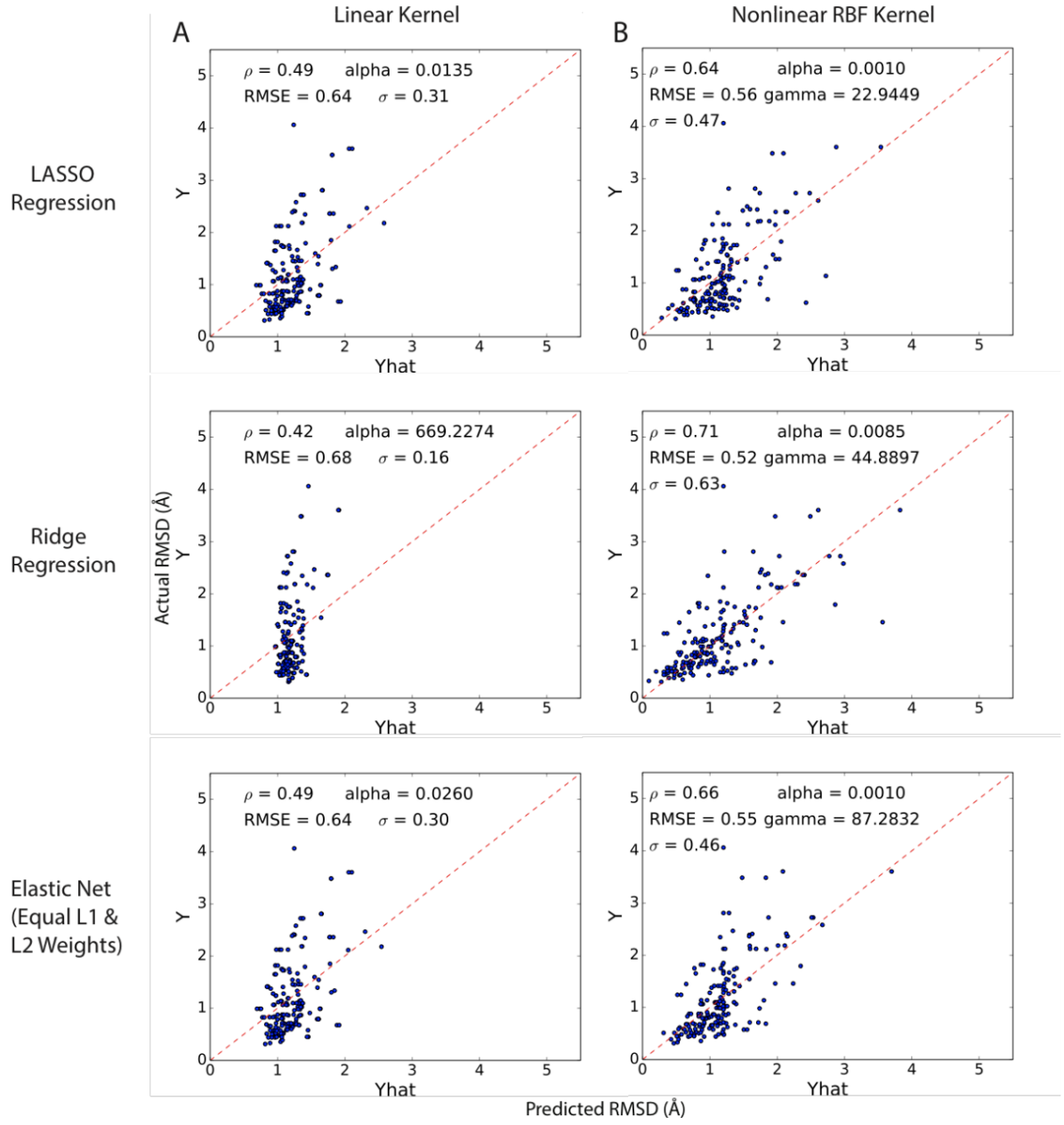


Figure 3.2: Extent Prediction for RMSD on the Held-out Portion of Benchmark Set: Left column: linear kernel. Right column: nonlinear RBF kernel. Top panel: LASSO regression. Middle panel: Ridge regression. Bottom panel: Elastic Net. Actual versus predicted RMSD. (Reprinted from [1])

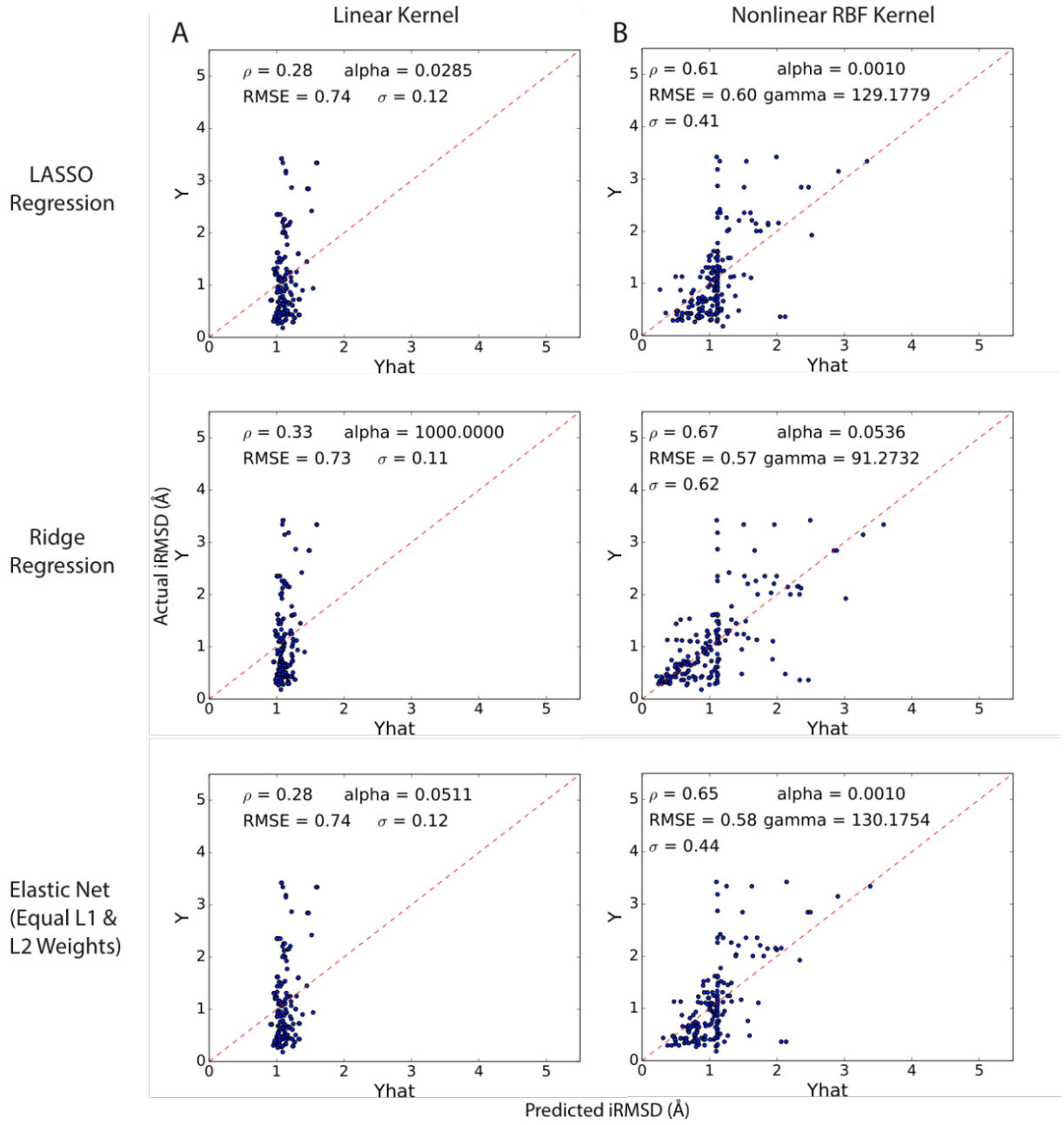


Figure 3.3: Extent Prediction for iRMSD on the Held-out Portion of Benchmark Set: Left column: linear kernel. Right column: nonlinear RBF kernel. Top panel: LASSO regression. Middle panel: Ridge regression. Bottom panel: Elastic Net. Actual versus predicted iRMSD. (Reprinted from [1])

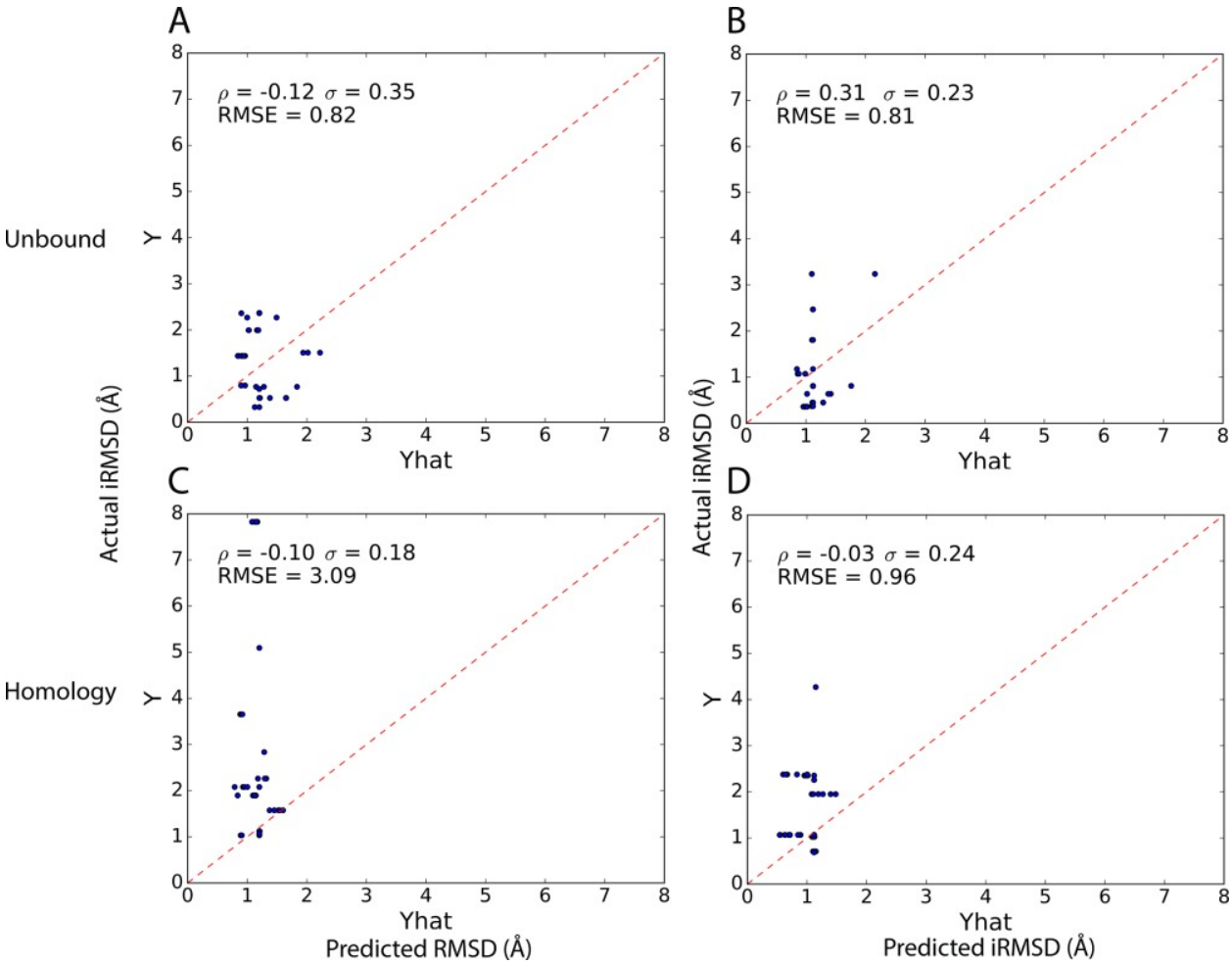


Figure 3.4: Extent prediction for RMSD and iRMSD on CAPRI Set: Top panel: Unbound-to-bound prediction. Bottom panel: Homology-to-bound prediction. Left panel: RMSD label. Right panel: iRMSD panel. (Reprinted from [1])

3.4 Statistical Analysis with Rule Extraction

The statistical analysis provides 319 significant network motifs with P-values < 0.001 as motif hotspot candidates. There are 12100 rules in original RF model. The SGL algorithm keeps 330 rules remained for classifying positive class which stands for driver genes. Among 330 rules, 327 network motifs participate in the process of decision as split points in decision trees. There are 256 network motifs in the intersection between 319 found in statistical analysis and 327 extracted from rules for positive class. These 256 are validated motif hotspots which would preferably attacked by cancer. The statistical analysis echoes the results from machine learning prediction, answering

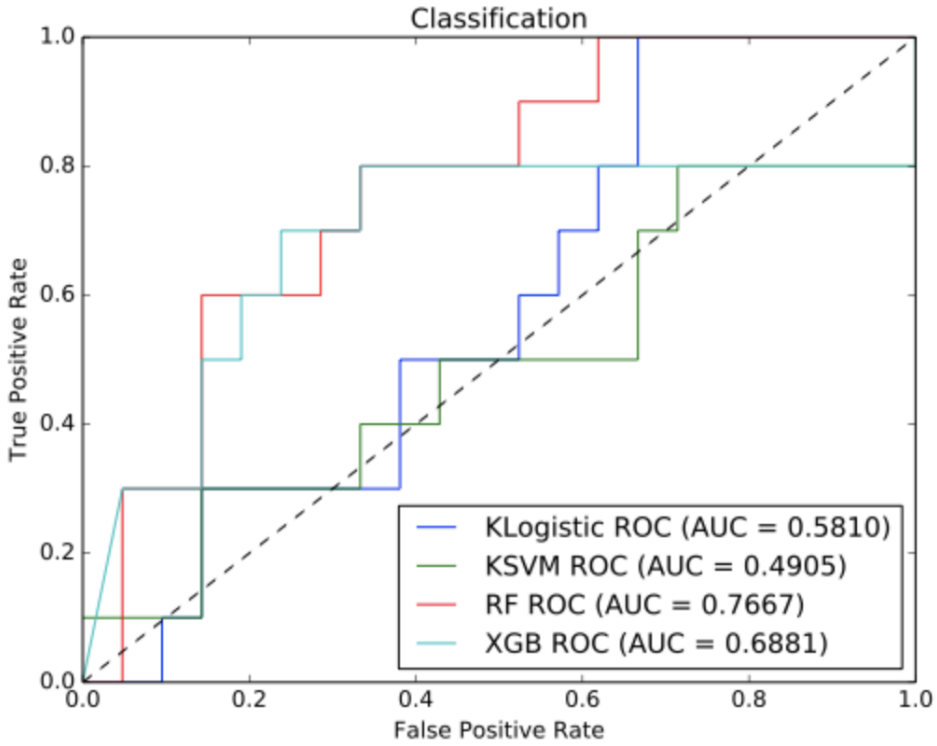


Figure 3.5: The Classification Results: The performance of various of machine learning models including kernel logistic classification with RBF kernel, kernel SVM with RBF kernel, gradient boosting method XGBoost, and Random Forest which outperforms others with AUCROC = 0.76.

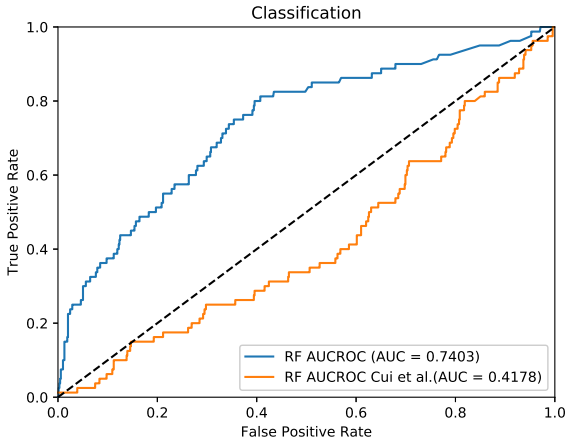
the question "where" with 256 motif hotspots. In other words, these motif hotspots are preferably attacked by cancer.

3.5 Understanding Cancer Mechanisms

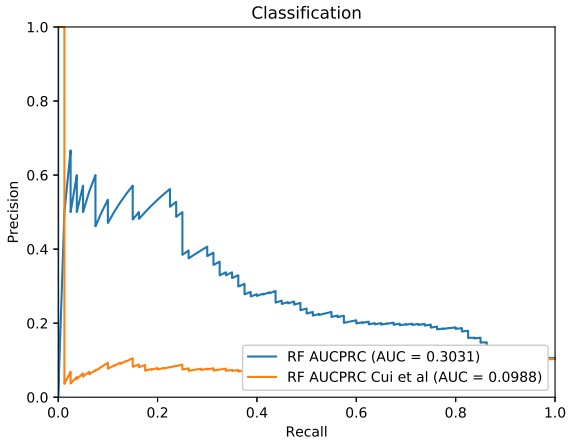
As we can see in Figure 3.8, there are 3 motif subtypes having well-known biological function and participate in cancer can validate our conjecture that cancer need to disrupt motifs to propagate its mutational effects.

Motif 38_4_0_0 in Figure 3.9 is a transient response device in PPIs system. Upstream activation signal through EFGR sent to downstream protein immediately, but balanced by the inhibition signal sent from intermediate regulator ZFP36. τ represents the delayed function for slower signal transduction. Downstream protein can only be active for a certain amount of time despite upstream kinase could be keep ON. Cancer will damage this device by attacking EGFR [5].

[!ht]



(a) Receiver Operating Characteristic Curves with AUCROC



(b) Precision-Recall Curves with AUCPRC

Figure 3.6: ROC and PC Curves Based on Network Motif Features and Centrality Features: Blue curves presents network motif classifier whereas yellow ones stands for centrality classifier. The network motif classifier obtains AUCROC = 0.74 and AUCPRC =0.30 comparing to the centrality classifier with AUCROC = 0.41 and AUCPRC = 0.098. The random AUCROC is 0.5 and random AUCPRC is the portion of positive class within the whole samples which is about 0.1 in this study

Figure 3.7: The Results of Statistical Analysis and Rule Extraction for Finding Motif Hotspots

Motif 98_34_0_0 in Figure 3.10 is a bi-stability switch. Once the phosphorylated concentration of upstream CCNE1 is above certain level, the downstream E2F will be switched to and locked on

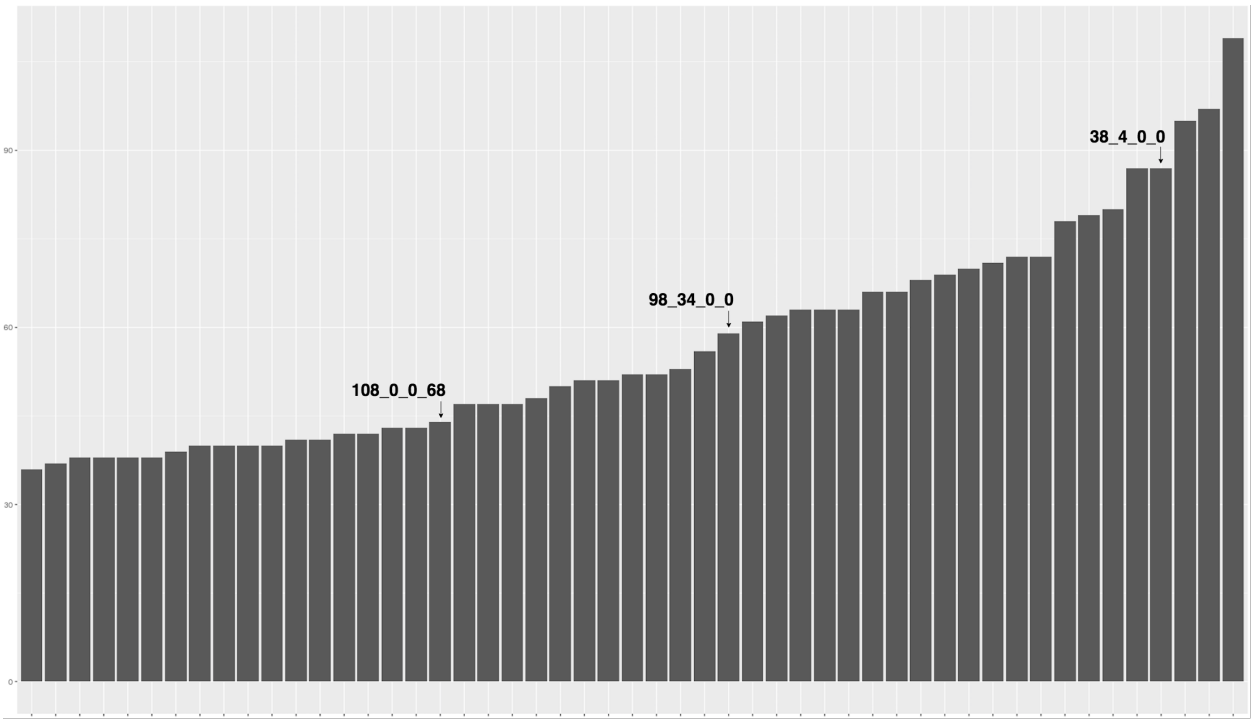


Figure 3.8: Motif Subtypes with Top 50 Scores: Three motifs, 38_4_0_0, 98_34_0_0, 108_0_0_68 marked in the figure are proven having direct relation with cancer with their biological function.

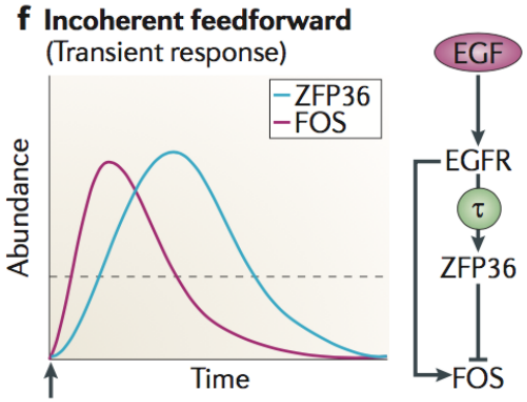


Figure 3.9: Motif 38_4_0_0: EGFR transient response, mutations on EGFR. (Reprinted from [5])

active state. Cancer would keep the switch at ON state by attacking upstream CCNE1 and RB1 [6].

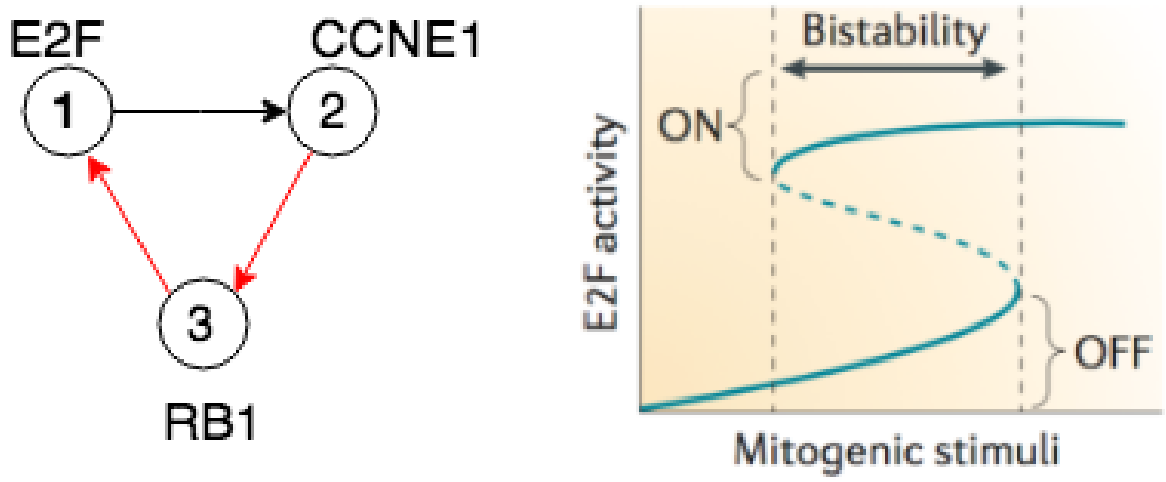


Figure 3.10: Motif 98_34_0_0: Bistability switch, mutations on CCNE1 and RB1. (Reprinted from [6])

Motif 108_0_0_68 Figure 3.11 is a mediator for protein degradation. When Z (upstream kinase) send an activation signal, X and Y (CDC proteins) can remain locked at active state. This device mediating ubiquitination and subsequent degradation of target proteins. Cancer would attack CDC proteins to maintain cell division [2, 7].

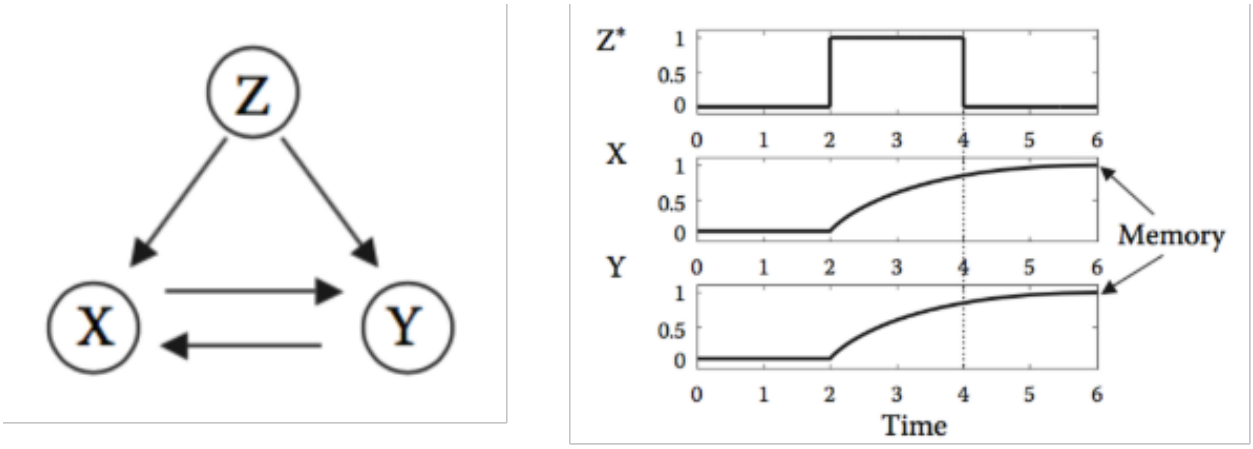


Figure 3.11: Motif 108_0_0_68: Cell Division Control (CDC) mediator, mutations on CDCs. (Reprinted from [2, 7])

To sum up, we answer "where" question by cross-validating 256 motif hotspots from statistical analysis and machine learning. We then use 3 motifs with specific biological function to validate that the reason why cancer would preferably attack those hotspots is because it needs to disrupt the normal biological function of motifs to create chaos in the PPIs system.

4. SUMMARY AND FUTURE STUDY

In this study, we investigate protein-protein interaction at two different cellular levels: molecular level and system level. At molecular level, we are eager to obtain the structure of protein complex. The barrier, however, in front of us is the induced-fit flexibility while proteins approach to each other from unbound to bound state. After applying rigidity assumption to simplify the problem to structural optimization question, we apply machine learning based on cNMA framework to predict the extent of conformational changes. The regression results show RMSE $= 0.52$ for RMSD and RMSE $= 0.57$ for interface RMSD. The performance on challenging CAPRI test set is about below 1 Å and insensitive to bad starting points from rigidity docking. A better training set containing more homology-to-bound samples should be soon constructed.

We subsequently assess the validity of rigidity assumption based on proof of contradiction. The idea is to assume all proteins as rigid and apply the machine learning model above to predict the extent of unbound-to-bound conformational change. The prediction results provide the scale to generating ensemble structures. If a flexible protein is falsely treated as rigid, it must endure energetic permissiveness and geometric distortion after the aforementioned process. The classifier has AUCROC = 0.76. The next step of this study is finding a ground truth to benchmark our performance.

As for the research at system-level PPI network, technically, we need to apply our model along with the state-of-the-art centrality model on the data used by the centrality model paper to validate the robustness of our model. Conceptually, after answering "where" and "why", it is imperative for use to answer "how" in the next step. Our plan is to higher the resolution to mutation level and discover the specific mechanism on each motif subtype. Meanwhile, we figure that the driver genes should not only be determined by the local topology, but also the its global environment. We plan to look for global features besides centrality to extend the interpretation from local network motif to global "super-motif".

Other than 3-node motifs, 4-node motifs have also been found and discovered as features in

our machine learning model. The results, however, did not improve. So we remove 4-node motifs features based on Occam's Razor.

The limitation of network motif idea is related to its definition. Similar to the mutation hotspots study only focused on mutations with high frequency across samples, the rare but also favorable-to-cancer subgraphs would be falsely ignored since it only considers the significant subgraphs in network.

REFERENCES

[1] H. Chen, Y. Sun, and Y. Shen, "Predicting protein conformational changes for unbound and homology docking: learning from intrinsic and induced flexibility," *Proteins: Structure, Function, and Bioinformatics*, vol. 85, no. 3, pp. 544–556, 2017.

[2] U. Alon, *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[4] M. Mashayekhi and R. Gras, "Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods," *International Journal of Information Technology & Decision Making*, vol. 16, no. 06, pp. 1707–1727, 2017.

[5] R. Avraham and Y. Yarden, "Feedback regulation of egfr signalling: decision making by early and delayed loops," *Nature reviews Molecular cell biology*, vol. 12, no. 2, p. 104, 2011.

[6] W. Kolch, M. Halasz, M. Granovskaya, and B. N. Kholodenko, "The dynamic control of signal transduction networks in cancer cells," *Nature Reviews Cancer*, vol. 15, no. 9, p. 515, 2015.

[7] E. A. Nigg, "Cell division: mitotic kinases as regulators of cell division and its checkpoints," *Nature reviews Molecular cell biology*, vol. 2, no. 1, p. 21, 2001.

[8] H.-F. Yang-Yen, J.-C. Chambard, Y.-L. Sun, T. Smeal, T. J. Schmidt, J. Drouin, and M. Karin, "Transcriptional interference between c-jun and the glucocorticoid receptor: mutual inhibition of dna binding due to direct protein-protein interaction," *Cell*, vol. 62, no. 6, pp. 1205–1215, 1990.

[9] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," *Science*,

vol. 285, no. 5428, pp. 751–753, 1999.

[10] S. Shangary and S. Wang, "Small-molecule inhibitors of the mdm2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy," *Annual review of pharmacology and toxicology*, vol. 49, pp. 223–241, 2009.

[11] R. A. Engh and R. Huber, "Accurate bond and angle parameters for x-ray protein structure refinement," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 47, no. 4, pp. 392–400, 1991.

[12] D. S. Wishart, B. D. Sykes, and F. M. Richards, "The chemical shift index: a fast and simple method for the assignment of protein secondary structure through nmr spectroscopy," *Biochemistry*, vol. 31, no. 6, pp. 1647–1651, 1992.

[13] J. Dubochet, M. Adrian, J.-J. Chang, J.-C. Homo, J. Lepault, A. W. McDowall, and P. Schultz, "Cryo-electron microscopy of vitrified specimens," *Quarterly reviews of biophysics*, vol. 21, no. 2, pp. 129–228, 1988.

[14] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall, "Cryo-electron microscopy of viruses," *Nature*, vol. 308, no. 5954, p. 32, 1984.

[15] D. W. Ritchie and G. J. Kemp, "Protein docking using spherical polar fourier correlations," *Proteins: Structure, Function, and Bioinformatics*, vol. 39, no. 2, pp. 178–194, 2000.

[16] G. R. Smith and M. J. Sternberg, "Prediction of protein–protein interactions by docking methods," *Current opinion in structural biology*, vol. 12, no. 1, pp. 28–35, 2002.

[17] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "Cluspro: a fully automated algorithm for protein–protein docking," *Nucleic acids research*, vol. 32, no. suppl_2, pp. W96–W99, 2004.

[18] R. Méndez, R. Leplae, M. F. Lensink, and S. J. Wodak, "Assessment of capri predictions in rounds 3–5 shows progress in docking procedures," *Proteins: Structure, Function, and Bioinformatics*, vol. 60, no. 2, pp. 150–169, 2005.

[19] M. F. Lensink, R. Méndez, and S. J. Wodak, "Docking and scoring protein complexes: Capri 3rd edition," *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. 4, pp. 704–718, 2007.

[20] M. F. Lensink and S. J. Wodak, "Docking and scoring protein interactions: Capri 2009," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, pp. 3073–3084, 2010.

[21] M. F. Lensink and S. J. Wodak, "Docking, scoring, and affinity prediction in capri," *Proteins: Structure, Function, and Bioinformatics*, vol. 81, no. 12, pp. 2082–2095, 2013.

[22] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, *et al.*, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, vol. 437, no. 7062, p. 1173, 2005.

[23] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[24] T. Noguti and N. Gō, "Efficient monte carlo method for simulation of fluctuating conformations of native proteins," *Biopolymers*, vol. 24, no. 3, pp. 527–546, 1985.

[25] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature Structural and Molecular Biology*, vol. 9, no. 9, p. 646, 2002.

[26] A. Bakan and I. Bahar, "The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding," *Proceedings of the National Academy of Sciences*, vol. 106, no. 34, pp. 14349–14354, 2009.

[27] E. Frezza and R. Lavery, "Internal normal mode analysis (inma) applied to protein conformational flexibility," *Journal of chemical theory and computation*, vol. 11, no. 11, pp. 5503–5512, 2015.

[28] S. E. Dobbins, V. I. Lesk, and M. J. Sternberg, "Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking," *Proceedings of the National Academy of Sciences*, vol. 105, no. 30, pp. 10390–10395, 2008.

[29] V. Tozzini, "Coarse-grained models for proteins," *Current opinion in structural biology*, vol. 15, no. 2, pp. 144–150, 2005.

[30] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *Journal of computational chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.

[31] C. Wang, P. Bradley, and D. Baker, "Protein–protein docking with backbone flexibility," *Journal of molecular biology*, vol. 373, no. 2, pp. 503–519, 2007.

[32] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, "Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations," *Journal of molecular biology*, vol. 331, no. 1, pp. 281–299, 2003.

[33] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding, *et al.*, "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Scientific reports*, vol. 3, p. 2650, 2013.

[34] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, p. 1113, 2013.

[35] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, p. 551, 2017.

[36] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein–protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

[37] Y. Cui, M. Cai, and H. E. Stanley, "Discovering disease-associated genes in weighted protein-protein interaction networks," *Physica A: Statistical Mechanics and its Applications*, 2017.

[38] E. C. Stites, P. C. Trampont, Z. Ma, and K. S. Ravichandran, "Network analysis of oncogenic ras activation in cancer," *Science*, vol. 318, no. 5849, pp. 463–467, 2007.

[39] J. A. McCammon, B. M. Pettitt, and L. R. Scott, "Ordinary differential equations of molecular dynamics," *Computers & Mathematics with Applications*, vol. 28, no. 10-12, pp. 319–326, 1994.

[40] L. Zhao, T. Sun, J. Pei, and Q. Ouyang, "Mutation-induced protein interaction kinetics changes affect apoptotic network dynamic properties and facilitate oncogenesis," *Proceedings of the National Academy of Sciences*, vol. 112, no. 30, pp. E4046–E4054, 2015.

[41] T. Oliwa and Y. Shen, "cnma: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions," *Bioinformatics*, vol. 31, no. 12, pp. i151–i160, 2015.

[42] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "Protein–protein docking benchmark version 4.0," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, pp. 3111–3114, 2010.

[43] R. Chen, L. Li, and Z. Weng, "Zdock: an initial-stage protein-docking algorithm," *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 1, pp. 80–87, 2003.

[44] B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng, "Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers," *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, 2014.

[45] A. R. Atilgan, S. Durell, R. L. Jernigan, M. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical journal*, vol. 80, no. 1, pp. 505–515, 2001.

[46] A. Bakan, L. M. Meireles, and I. Bahar, "Prody: protein dynamics inferred from theory and experiments," *Bioinformatics*, vol. 27, no. 11, pp. 1575–1577, 2011.

[47] A. Bakan, A. Dutta, W. Mao, Y. Liu, C. Chennubhotla, T. R. Lezon, and I. Bahar, "Evol and prody for bridging protein sequence evolution and structural dynamics," *Bioinformatics*, vol. 30, no. 18, pp. 2681–2683, 2014.

[48] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan, and M. Karplus, "Charmm: a program for macromolecular energy, minimization, and dynamics calculations," *Journal of computational chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[49] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, p. 177, 2004.

[50] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Mfinder tool guide," *Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot Israel, Tech Rep*, 2002.

[51] M. L. Miller, E. Reznik, N. P. Gauthier, B. A. Aksoy, A. Korkut, J. Gao, G. Ciriello, N. Schultz, and C. Sander, "Pan-cancer analysis of mutation hotspots in protein domains," *Cell systems*, vol. 1, no. 3, pp. 197–209, 2015.

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[53] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.