Chapman University

Chapman University Digital Commons

Computational and Data Sciences (PhD) Dissertations

Dissertations and Theses

Fall 1-2020

Development of Integrated Machine Learning and Data Science Approaches for the Prediction of Cancer Mutation and Autonomous Drug Discovery of Anti-Cancer Therapeutic Agents

Steven Agajanian
Chapman University, agaja102@mail.chapman.edu

Follow this and additional works at: https://digitalcommons.chapman.edu/cads_dissertations

Recommended Citation

S. Agajanian, "Development of integrated machine learning and data science approaches for the prediction of cancer mutation and autonomous drug discovery of anti-cancer therapeutic agents," Ph.D. dissertation, Chapman University, Orange, CA, 2021. https://doi.org/10.36837/chapman.000220

This Dissertation is brought to you for free and open access by the Dissertations and Theses at Chapman University Digital Commons. It has been accepted for inclusion in Computational and Data Sciences (PhD) Dissertations by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

Development of Integrated Machine Learning and Data Science Approaches for the Prediction of Cancer Mutation and Autonomous Drug Discovery of Anti-Cancer Therapeutic Agents

A Dissertation by

Steven Agajanian

Chapman University

Orange, CA

Schmid College of Science and Technology

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Computational and Data Sciences

January 2021

Committee in charge:

Gennady M. Verkhivker, Ph.D., Committee Chair

Hesham El-Askary, Ph.D.

Erik Linstead, Ph.D.

Cyril Rakovski, Ph.D.

The Dissertation of Steven Agajanian is approved

Gennady Verkhivker

Gennady Verkhivker Ph.D Committee Chair

Hesham El-Askary

Hesham El-Askary Ph.D

Erik Linstead

Digitally signed by Erik Linstead
Date: 2021.01.20 10:04:07 -08'00'

Erik Linstead Ph.D

Cyril Rakovski Ph.D

December 2020

Development of Integrated Machine Learning and Data Science Approaches for the Prediction of Cancer Mutation and Autonomous Drug Discovery of Anti-Cancer Therapeutic Agents

Copyright © 2021

by Steven Agajanian

ACKNOWLEDGEMENTS

I would like to thank Dr. Gennady Verkhivker, Dr. Hesham El-Askary, Dr. Erik Linstead, Dr. Cyril Rakovski, Dr. Oluyemi Odeyemi, Natalia Stewart, Hamilton Pitlik, and everyone who helped me along the way. I would not have succeeded without you. I also wanted to express additional appreciation for everyone in the Chapman University Computational and Data Science Program for providing the support necessary for me to accomplish everything included in this dissertation.

ABSTRACT

Development of Integrated Machine Learning and Data Science Approaches for the Prediction of Cancer Mutation and Autonomous Drug Discovery of Anti-Cancer Therapeutic Agents

by Steven Agajanian

Few technological ideas have captivated the minds of biochemical researchers to the degree that machine learning (ML) and artificial intelligence (AI) have. Over the last few years, advances in the ML field have driven the design of new computational systems that improve with experience and are able to model increasingly complex chemical and biological phenomena. In this dissertation, we capitalize on these achievements and use machine learning to study drug receptor sites and design drugs to target these sites. First, we analyze the significance of various single nucleotide variations and assess their rate of contribution to cancer. Following that, we used a portfolio of machine learning and data science approaches to design new drugs to target protein kinase inhibitors. We show that these techniques exhibit strong promise in aiding cancer research and drug discovery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Machine Learning and Biochemical Applications

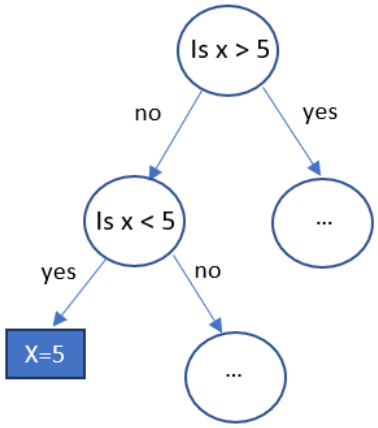
1.1 Introduction

Few technological ideas have captivated the minds of biochemical researchers to the degree that machine learning (ML) and artificial intelligence (AI) have. Over the last few years, advances in the ML field have driven the design of new computational systems that improve with experience and are able to model increasingly complex chemical and biological phenomena (Chen et al., 2018; Dimitrov et al., 2019; Goh et al., 2017; Korotcov et al., 2017; Mater and Coote, 2019; Popova et al., 2018). ML techniques have been successfully applied to various computational chemistry challenges (Husic and Pande 2018), pharmaceutical data analysis, (Burbidge et al. 2001) protein–ligand binding affinity prediction problems (Ballester and Mitchell 2010, Decherchi et al. 2015), dissecting molecular determinants of protein mechanisms and biochemical reactions (Li et al., 2015, Cortina and Kasson 2018, Shcherbinin and Veselovsky 2019). Data-intensive ML modeling can be also applied for detection and classification of allosteric protein states. The integration of Markov modeling, simulations and ML approaches into robust and reproducible computational pipelines with the experimental feedback can be explored for atomistic modeling and classification of allosteric states. Two key factors were necessary for them to see so much use. First, large amounts of rich data. We are generating more data today than ever before and the biochemistry field is no exception. Computational tools for molecular modeling (Cao & Kipf, 2018; Kadurin, Nikolenko, Khrabrov, Aliper, & Zhavoronkov, 2017), protein folding simulation (Hespenheide, Rader, Thorpe, & Kuhn, 2002), or mutation analysis (Adzhubei, et al., 2010) have started to generate more data than can even be stored. Second, powerful computational tools like GPUs that augment our abilities to perform parallel processing allowing ML models to ingest these large datasets. This has allowed medicine to become more personalized, with current research catering solutions to specific genetic profiles rather than taking a one size fits all approach (Vogenberg, Barash, & Pursel, 2010). Much of the benefit of these

methods comes from their versatility: not only do they both generate and analyze data, but often enhancements to ML techniques in one domain can be readily applied to techniques in any domain. For example, techniques designed for the image processing domain have been applied to molecular design (Maziarka, et al., 2020). There are three main machine learning techniques used in the following projects, Random forests, neural networks, and reinforcement learning techniques.

1.2 Random Forests

Random Forests are a type of machine learning model that have been proven to be very robust in a variety of applications. First proposed in 2001 by Leo Breiman (Breiman, Random Forests, 2001), Random Forests attempt to improve on the shortcomings of the decision tree model using that statistical technique "bagging" which is short for bootstrap aggregation. Decision trees attempt to learn about the data by creating binary split points with yes/no answers at every point, like below.

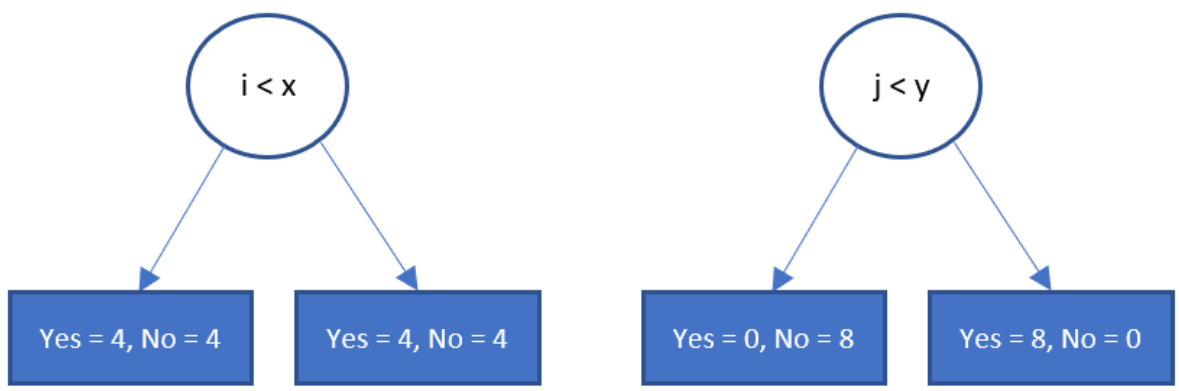


the modern form of decision tree learning was also proposed by Breiman et al in 1984 (Breiman, Friedman, Stone, & Olshen, 1984), and performed well on a variety of machine learning tasks. These models can learn to perform both classification and regression.

For classification tasks, decision trees will attempt to minimize or maximize a particular metric at every node in the tree. This metric is often the Gini impurity coefficient defined as

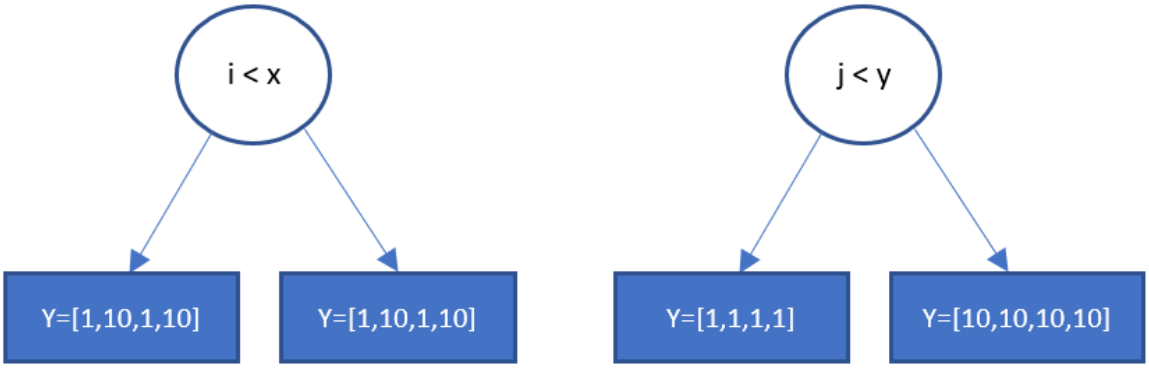
$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Where C is the total number of classes and p(i) is the probability of picking a data point with class i. In order to optimize this metric, decision trees take a brute force approach to the problemand observe what the G would like like with every possible split value for every possible column. Once they have calculated the target metric for every possible point, they choose the split value/column pair that optimizes most.



In this example, the tree on the left would have the worst possible split point, with a 50-50 chance to get the right answer based on this column/pair combination. This system represents an entropy of 1, and would be actively avoided by the decision tree. The right tree on the other hand has the best possible split point that would get the correct answer every time. This system represents an entropy of 0. This is the scenario that the decision tree tries to arrive at every time. Decision trees will grow until they reach a 'pure' leaf node. Leaf nodes are the ending point of any decision tree, and are required to hold the answer to the classfication problem. A pure leaf node is one such that all samples belong to one class. So, the tree on the right above has two pure leaf nodes and would stop growing.

Regression trees also try to optimize a metric, though they use a metric that allows for continuous data. This is typically the mean squared error or the mean absolute error. These trees also take the same brute force approach to optimizing their target. However, instead of a class distribution at each node, regression trees compute the average of all the y values that belong to each node.



Once again, the left tree represents the worst possible split point. The model would output a regression value of (10+1+10+1)/4 = 5.5 regardless of whether the target was 1 or 10, for a mean squared error of 20.25. The right tree is the perfect split point, predicting 1 when the target is 1 and 10 when it is 10, for a mean squared error of 0. These will grow until pure leaf nodes are obtained.

These models have been proven to overfit the training set, since often pure leaf nodes can only be obtained by creating enough split points where leaf nodes only contain one sample. Techniques like pruning, where trees would be forced to stop growing early, alleviated the problem but weren't a perfect solution. Random Forests were the next solution, and though not perfect, they increased the performance immensely. The most important of the Random Forest's improvements is "bagging". Bagging works by randomly exposing a predetermined number of decision trees to random samples of the dataset, and then aggregating their predictions with either majority vote (classification trees) or another layer of averaging (regression trees). Not only are they exposed to random subset of the rows however, they are also only shown random subsets of the columns at each split point. This forces the decision trees to learn to use other columns rather than relying on any dominant column which leads to overfitting.

1.3 Neural Networks and Deep Learning

The remarkable rise of deep learning (DL) relying on the robust function approximations and representation properties of deep neural networks has provided us with new tools to automatically find compact low-dimensional representations (features) of high-dimensional data (LeCun et al., 2015). DL models have achieved outstanding predictive performance making dramatic breakthroughs in a wide range of applications, including automatic speech processing and image recognition (Hey et al., 2020; Kim et al., 2019; Toledano et al., 2018; Wu et al., 2020). In the words of Geoffrey Hinton who is the founder of DL technologies "Deep Learning is an algorithm which has no theoretical limitations on what it can learn; the more data you give and the more computational time you provide the better it is" (LeCun et al., 2015). Deep neural network methods were successfully applied to predict intrinsic molecular properties such as atomization energy based on simple molecular geometry and element types (Rupp et al., 2012). DL models were recently used for structure-functional prediction of cancer mutations and functional hotspots of ligand binding in cancer-associated genes (Agajanian et al., 2018). The developed models can capture ~90% of experimentally validated mutational hotspots and yield novel information about molecular signatures of driver mutations. In the recent studies, we have proposed novel DL architectures capable of predicting functional protein hotspots directly from raw nucleotide sequence information (Agajanian et al., 2019). These studies have shown that DL models can learn high importance features from raw genomic information and produce reliable recognition and classification of functionally significant cancer mutation hotspots. Moreover, these DL models can often outperform computational predictors of cancer mutations that are based on protein sequence and structure features (Agajanian et al., 2019). The success of DL tools in deciphering important functional phenotypes directly from primary sequence information is encouraging as these models can bypass the need for a large number of empirically derived functional and structural features. However, ML methods often result in "black box" models with limited interpretability. There has been an explosion of interest in transparent and interpretable ML models to enable more efficient data mining and scientific knowledge discovery (Holzinger et al., 2014). Our investigations have also provided a roadmap how to combine DL predictions of functional sites with subsequent biophysical analysis to aid

in the interpretability of ML models and facilitate their applications in biological problems (Agajanian et al., 2018; Agajanian et al., 2019).

Neural networks are some of the most powerful machine learning models due to their ability to approximate any function (Csáji, 2001). First proposed in 1943, they have become a juggernaut in the machine learning world having benefited immensely from computational advancements like the GPU. Inspired by the brain's architecture for decision making, these models stack layers of nodes called perceptrons together. These perceptrons learn a set of weights to linearly combine with their inputs and feed the output into a nonlinear activation function. For a given set of predictors x_i with activation function f and weights w the perceptron's output is defined as

$$P(x_i) = f(\sum_{j=0}^{n} x_{ij} w_j)$$

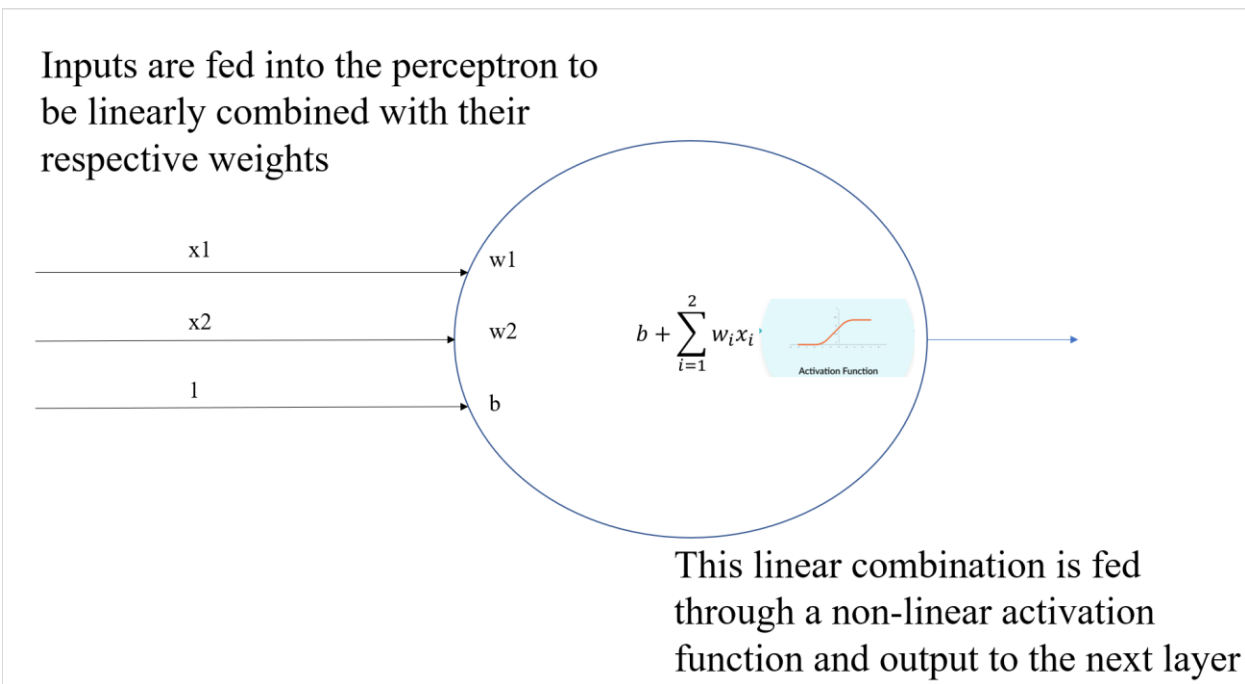


Figure 1. Inner Workings of a Perceptron

This value is either fed into a subsequent layer or used as the output of the entire network. Typically, sigmoidal functions or functions such as rectified linear units (ReLU) are used. This allows smooth gradients to be obtained that can be used to improve the model's weights and fit more closely to the data.

Essentially, this model can be thought of as a network of logistic regressions that increase in level of abstraction with each subsequent layer. The output layer can then perform a final logistic regression that will classify/regress the input data point. In order to update the weights of a neural network, the backpropagation algorithm is used. This algorithm computes the error with respect to each of the perceptrons in the network so that they can alter their weights in the right direction and is comprised of four equations. In the first step of backpropagation, the weights of the output layer are updated with respect to the cost function of the neural network. This cost function is an estimate of how far off a prediction was from the actual answer. So, the error for the output layer δ^L with respect to cost function C is

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

Where $\nabla_a C$ is the vector of partial derivatives of cost with respect to activations, and $\odot$ is the Hadamard product operator. After the error is calculated for the output layers, we can calculate the error for each preceding layer in terms of the output of the following layer with transposed weights matrix $\left(w^{l+1}\right)^T$ and error δ^{l+1}.

$$\delta^l = \left(\left(w^{l+1}\right)^T \delta^{l+1}\right) \odot \sigma'(z^l)$$

These two equations allow us to calculate the error in any layer even though we don't know what the correct activation is for any of the nodes in that layer. Once we can assign error to the nodes in any layer, we can calculate the gradient with respect to the biases b^l in layer l as

$$\frac{\partial C}{\partial b^l} = \delta^l$$

And the gradient with respect to any weight in layer l w_{jk}^l as

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l$$

These four equations give all the tools necessary to obtain optimal weights for any given neural network. Once the gradients are obtained for the weights and biases, they are simply nudged in the direction of the gradient to reduce the error. Alternate types of neural networks are widely used such as recurrent neural networks (RNN) or convolutional neural networks (CNN). These networks approach the learning task a little differently, tailoring the network to excel in different scenarios.

RNNs alter the neural network framework by enforcing that not only do nodes feed forward to the next layer, they send their output back into themselves as an input for the next sample. This allows the network to exhibit "memory" like properties since it can start to gain some context for its predictions. These types of models have excelled in sequence prediction tasks due to this context addition. Notably, they have been proven to have limits to how informative the context can be as they have trouble remembering contextual details for very long sequences. This is known as the vanishing gradient problem and was combatted by the long short-term memory (LSTM) and gated recurrent unit (GRU), which have additional parameters that dictate which information carries on to the next sample and which doesn't. However, this additional functionality comes at the cost of training time, as RNN architectures are notoriously expensive to train. CNNs on the other hand lend themselves to problems that have a spatial component to them, like in vision. When the arrangement of the inputs with respect to themselves matters, like in object detection or any other vision related task, CNNs have excelled. This is due to their convolving filter framework, where a filter is convolved around the input and features are extracted in order to reduce dimensionality. No matter which neural network architecture is used, backpropagation will be used in some form to train the model.

Autoencoders are a neural network architecture used to encode and decode data (Hinton & Salakhutdinov, 2006). Typically, these models are applied to compress or reduce the dimensionality of data resulting in an output variable referred to as the latent variable or code. The coupled encoder/decoder architecture provides a way to encode data into a latent variable and decode that latent variable back to the original input. In order to do this, a neural network is instantiated with decreasing numbers of nodes per layer until the bottleneck layer is reached. The bottleneck layer is where the latent variable is created. The following layers increase

in size similarly to the beginning layers, culminating in an output layer with the same dimensionality as the input layer. In essence, the model attempts to take an input, reduce it down to the size of the bottleneck layer, and then recreate it with as much similarity to the original input as possible. This means that this bottleneck layer must contain an informative representation of the input otherwise there would be no chance at success in recreation. These models have a wide variety of applications, from the computer vision domain to the natural language processing domain. Variational autoencoders enhance autoencoders by learning a probabilistic mapping rather than a deterministic one (Welling & Kingma, 2013). In other words, in an autoencoder framework a latent variable will have only one decoded representation, in a variational autoencoder, the latent variable will have a probabilistic distribution of possible decoded representations. This probabilistic process more accurately represents most underlying data distributions because it is more often that problems are stochastic in nature.

1.4 Reinforcement Learning

One of the primary goals of artificial intelligence (AI) is to produce fully autonomous agents that interact with their environments to learn optimal behavior, improving over time through trial and error. An important mathematical framework for experience-driven autonomous learning through interactions with the environment is reinforcement learning (RL) (Sutton and Barto, 1981; Barto, 1994; Botvinick, 2012; Hassabis et al., 2017). RL is similar to supervised learning, however rather than making a prediction with a binary correct/incorrect outcome, reinforcement learning models suggest an action that receives a reward value. These techniques add a layer of complexity on the learning task due to the lack of a “right” answer, which is replaced by the reward value. Furthermore, classification tasks typically have a reasonable number of classes that can be predicted whereas RL tasks have no limit. As such, these models typically have a multitude of actions that could yield a positive reward value, and they learn a policy that dictates what action to take when given a state representation. For a given state representation s, RL models maintain an internal probability estimate of taking any action a at time t with policy π:

$$\pi(\mathrm{a}, \mathrm{s}) = \Pr(a_t = a | \, s_t = s)$$

While previous RL approaches lacked scalability and were limited to fairly low-dimensional problems, a marriage between deep neural networks and RL resulted in the new rapidly evolving field of deep reinforcement learning (DRL) that has achieved remarkable success in game-oriented and various scientific applications, attaining a wide popularity and celebrity-like following among researchers (Botvinick et al., 2019; Jaderberg et al., 2019; Mnih et al., 2015; Senior et al., 2019; Silver et al., 2017). DRL concepts leverage and symbiotically combine neural network modeling with reinforcement learning, in which optimization strategies are crafted based on the trade-offs and competition between rewards and punishments rather than conventional deterministic or stochastic exploration methods. After years of serving as a mere inspiration rather than a practical tool, DRL techniques have taken off overcoming scalability and data limitation issues and exploding into one of the most intense areas of AI research. Recent years have witnessed the expansion of DRL applications into biomedical research including but not limited to biomedical informatics, drug discovery (Baskin, 2020; Grebner et al., 2020), and toxicology (Chary et al., 2020).

The rationale for employing DRL technologies in studies of allosteric regulation is to capitalize on conceptual and algorithmic similarity between Markov decisions processes (MDPs) which are at the core of RL methods and Markovian modeling of allosteric states in proteins. Several methods adopted RL-based conceptualization to develop MDP algorithms for conformational mapping of the protein landscapes and detection of functional allosteric states. REinforcement learning based Adaptive samPling (REAP) algorithm has shown a considerable promise by adopting RL principles in which an agent (or learning algorithm) takes actions in an environment (conformational protein landscape) to maximize a reward function (Shamsi et al., 2018). In this study, the action is associated with launching a pool of simulations along different collective variables (reaction coordinates), with the reward function proportional to the efficiency of a reaction coordinate to sample space and detect unknown states, and the agent selecting the directions which are most rewarding ultimately leading to the optimal adaptive strategy (Shamsi et al., 2018). Similar concepts were used to develop a goal-oriented sampling method, termed fluctuation

amplification of specific traits (FAST) for rapid search of conformational space and identification of distinct functional states by balancing search near promising solutions (exploitation) and attempts to find novel solutions (exploration). Inspired by the RL ideas, this methods runs pools of simulations from starting points chosen based on the reward functions that encourages discovery of new conformations with selected physical properties (Zimmerman and Bowman, 2015; Zimmerman et al., 2018). Generative neural networks have been recently proposed as a tool for the discovery of efficient collective variables that are fundamental for adaptive exploration of the conformational landscapes and finding functional states and hidden allosteric states by guiding sampling towards poorly explored regions (Chen et al., 2018; Chiavazzo et al., 2017; Hernandez et al., 2018; Mardt et al., 2018). MD simulations were combined with DL approach to train an autoencoder (Hinton and Salakhutdinov, 2006) in order to generate new protein conformations and mine conformational space of the bound state from an ensemble of unbound protein structures (Degiacomi, 2019). Another interesting study employed autoencoder-based detection algorithm to explore dynamic allostery induced by ligand binding based on the comparison of time fluctuations of distance matrices obtained from MD simulations in ligand-bound and unbound forms (Tsuchiya et al., 2019). In this method, the autoencoder neural network is first trained on the time fluctuations of protein motions in the apo form, and the trained autoencoder is then applied to analyze patterns of fluctuations in the holo form. Using this elegant implementation of RL approach, the authors mapped allosteric communication networks of the dynamically coupled residues and ligand pockets in the PDZ2 domain induced by binding (Tsuchiya et al., 2019). Allosteric pocket crosstalk defined as a temporal exchange of atoms between adjacent pockets in the MD trajectories can produce a fingerprint of hidden allosteric communication networks (La Sala et al., 2017). The recent RL-inspired studies of allosteric systems suggested that simulation-driven ML modeling and analysis of conformational landscapes may uncover rarely populated functional states and shed the light on the key features of allosteric communications between visible and hidden binding pockets in proteins.

DRL is a continuous trial-and-error based sampling-learning process where the agent tries to apply different combination of actions on a state to find the highest cumulative reward. Although DRL methods can tackle

a wide range of learning problems with a rigorous mathematical formulation, the challenges posed by the properly crafted interplay between rich data acquisition and delayed rewards remains a significant impediment to the widespread of RL methods in many application domains, including prediction of allosteric protein states and mechanisms. The challenges of DRL approaches often lie in the art of designing robust reward function. The hybrid reward functions with a weighted combination of topological, dynamic and network-based rewards describing different characteristics of allosteric states may represent a potentially interesting strategy to mitigate the inherent deficiencies of RL and DRL methods. For this, the rewards may be treated as neural networks trained on the set of known allosteric states. A new saga in the rapidly evolving DRL field was the development of episodic-based DRL algorithms that estimate the value of actions and states using episodic memories where the agent stores each encountered state along with the sum of rewards obtained during the n time steps (Botvinick et al., 2019). The episodic memory-based models can be extended to develop curiosity reward bonus functions for efficient exploration of the environment and detecting states in the poorly accessible regions (Han et al.; 2020). In this context, DRL framework that iterates episodes of DRL and community decomposition of the dynamic network flows on the conformational landscapes may enhance the interplay between sampling and learning, thus facilitating identification of hidden allosteric states. Different from traditional DRL approaches, this strategy can consider communities of states as intermediate stages in the learning process, rather than unique states, which could potentially lead to a more robust and versatile learning procedure.

Deep neural network (DNN) models, most notably autoencoders and variational autoencoders (VAE) (Gomez-Bombarelli et al., 2018) and generative adversarial networks (GANs) (Sorin et al., 2020; Zhong et al., 2020) have proven fruitful in chemical discovery and molecular design of novel synthesizable chemical probes. Automated chemical design approaches employed VAE neural networks for a data-driven continuous representation of molecules (Gomez-Bombarelli et al., 2018).

GAN models are often considered as one of the most significant advances in the field of machine learning, and their success has generated a considerable momentum with growing number of applications including molecular design of novel chemical probes and materials (Olivecrona et al., 2017; Yu et al., 2017; Gupta

et al., 2018; Kadurin, Aliper, et al., 2017; Kadurin et al., 2017; Polykovskiy et al., 2018; Putin, et al., 2018a,b). By leveraging sequence data generation (SeqGAN) approach (Yu et al., 2017); Objective-Reinforced Generative Adversarial Networks (ORGAN) (Guimaraes et al., 2017) combines GANs and RL to apply the GAN framework to molecular design with domain-specific rewards and feedback. Of particular importance is MolGAN, an implicit, generative model for small molecular graphs that circumvents the need for expensive graph matching procedures and adapts GAN approach to operate directly on graph-structured data (Cao et al., 2018). CycleGAN provides unpaired image-to-image translation using Cycle-Consistent Adversarial Networks (Zhu et al., 20128). MolCycleGAN, which extended the CycleGAN framework with an added loss and extra encoding network, maps from distribution to distribution on unpaired samples, so it can amplify the size of our dataset in the process by taking all of the pairing combinations rather than relying on a training dataset of predefined molecule-inhibitor pairs (Maziarka et al., 2019). The advantage of MolCycleGAN is the ability to learn transformation rules from the sets of compounds with desired and undesired values of the considered property. The methodological and algorithmic progress in GAN applications to molecular discovery has been further catalyzed by the development of several comprehensive benchmarking sets embedded into a sophisticated cheminformatics infrastructure supporting open-source implementations of molecular features and learning algorithms (Olson et al., 2017; Racz et al., 2019; Polykovskiy et al., 2018). Despite recent developments in GANs models, the applicability of these tools for molecular design continues to present a promise rather than a validated strategy, lacking systematic and comprehensive tools to target specific protein families and interrogate molecular mechanisms. There is also growing interest in generative models which can incorporate both chemical and structural information about small molecules and their interactions with protein targets.

GANs pit two competing models, the generator and discriminator, against each other in a minimax game of counterfeiting and detection (Goodfellow, et al., 2014). The generator attempts to sample from a learned distribution that maximizes the probability of fooling its adversary, the discriminator. The discriminator, in

turn, attempts to estimate the probability that a sample was created by the generator. With this, the generator minimizes the equation (Goodfellow, et al., 2014).

$$\nabla_{\theta_\theta} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)$$

However, minimizing this equation proves to saturate in training (Goodfellow, et al., 2014). To combat this the authors suggested flipping the equation to instead maximize

$$\nabla_{\theta_\theta} \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log\left(D\left(G\left(z^{(i)}\right)\right)\right).$$

What this means is that the generator is trying to maximize the probability of the discriminator being incorrect rather than minimizing the probability of the discriminator being correct. Simultaneously, the discriminator is trying to improve its discerning abilities by optimizing the equation

$$\nabla_{\theta_d} \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right].$$

This sets up a two-player minimax game defined by the following adversarial loss equation

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log\left(1 - D\left(G(z)\right)\right)].$$

Essentially, this function creates an environment whereas the discriminator gets better at telling real samples from generated samples, the generator must produce better fakes to receive a reward. When executed correctly, this results in the two models learning together and the generator developing fakes that the discriminator thinks are real (Goodfellow, et al., 2014).

Chapter 2: Machine Learning Classification and Structure-Functional Analysis of Cancer Mutations Reveal Unique Dynamic and Network Signatures of Driver Sites in Oncogenes and Tumor Suppressor Genes

2.1 Introduction

A central goal of cancer research is to discover and characterize functional effects of mutated genes that contribute to tumorigenesis. The Cancer Genome Atlas and related DNA sequencing initiatives have motivated sequencing studies of tumors, producing invaluable insights into the underlying genomic basis of tumorigenesis (Davies, et al., 2002; Bardelli, et al., 2003; Wang, et al., 2004; Samuels, et al., 2004; Stephens, et al., 2004; Futreal, et al., 2004; Stephens, et al., 2005; Sjoblom, et al., 2006; Wood, et al., 2007). Cancer genome landscapes of somatic mutations have been extensively characterized through deep-sequencing analyses of the coding exomes and whole genomes in a variety of cancer types, showing that there are ~140 genes whose intragenic mutations contribute to cancer, with a relatively small fraction of recurrent somatic variants (termed "driver" mutations) providing growth advantage to cancer cells and often detected based on the mutational frequency in high-throughput studies (Greenman, et al., 2007; Watson, Takahashi, Futreal, & Chin, 2013; Vogelstein, et al., 2013). Most of the somatic mutations are "passengers" that occur stochastically as a result of mutagenesis, without a measurable functional impact (Vogelstein, et al., 2013; Lawrence, et al., 2013). By examining the alterations driving cancer formation in more than 7,600 tumors, molecular evolution-informed studies have revealed that a relatively consistent small number of mutated genes is required to convert a single normal cell into a cancer cell across cancer types (Martincorena, et al., 2017). Recent approaches have also focused on discovery of putative driver mutations within the non-coding regions of the genome (Piraino & Furney, 2016; Rheinbay, et al., 2017). Although 'major drivers' can strongly promote tumor growth, some passenger mutations can be individually weak yet collectively deleterious, suggesting that disease progression is often difficult to rationalize on the basis of a binary driver-passenger classical model (De & Ganesan, 2017). Cancer-associated mutations that are

less essential for tumor growth and present at low frequency in cancer cohorts can form a group of so-called "mini-drivers", indicating that mutational patterns in cancer genomes are highly heterogeneous spanning a continuum of phenotypic impacts (Castro-Giner, Ratcliffe, & Thomlinson, 2015). The advances in high-throughput genome analysis and next-generation sequencing (NGS), have led to the initiation and development of multi-centered cancer genomic projects and major data portals, such as The Cancer Genome Atlas (TCGA) hosted at the Genomics Data Commons Portal (Weinstein, et al., 2013), COSMIC database (Forbes, et al., 2015), and the International Cancer Genome Consortium (ICGC) cancer genome projects (Hudson, et al., 2010; Zhang, et al., 2011). TCGA data include information about 40 cancer projects from > 20,000 genes and 3,142,246 mutations (Jensen, Ferretti, Grossman, & Staudt, 2017). The ICGC data portal include 84 cancer projects of 22 cancer primary sites with 77,462,290 annotated simple somatic mutations (Klonowska, et al., 2016; Hinkson, Davidsen, Klemm, Kerlavage, & Kibbe, 2017). A highly detailed gene and tumor entity centric analysis of vast TCGA data is now readily accessible (Deng, Bragelmann, Schultze, & Perner, 2016). The cBio Cancer Genomics Portal provides access to cancer genomics data sets from > 5,000 tumor samples, 215 cancer studies and 981 genes (Cerami, et al., 2012; Gao, et al., 2013). The development of oncogenomic resources enabled cancer genomics to join the big data revolution, facilitating data-driven analyses of genetic alterations in multiple tumor types (Poulos & Wong, 2018).

A decade-long monumental cancer genomics efforts have recently culminated in the completion of PanCancer Atlas project, describing an unprecedented in its depth and scope analysis of molecular and clinical information from > 10,000 tumors representing 33 types of cancer (Ding, et al., 2018; Ellrott, et al., 2018). By combining a battery of functional and computational approaches, the Multi-Center Mutation Calling project has generated a comprehensive collection of somatic mutation calls and classified 751,876 unique missense mutations across 299 cancer driver genes, leading to 9,919 predicted cancer driver mutations (Bailey, et al., 2018). This herculean effort has produced a dataset of 3,442 predicted driver mutations that were validated through consensus of functional analysis, sequence-based and structure-based

approaches. Another systematic functional analysis and annotation of somatic mutations in cancer leveraged a genomic-based platform sensitive to weak drivers, producing a dataset of 1,049 experimentally tested somatic mutations (Ng, et al., 2018). A number of computational tools that measure the functional impact of a given single nucleotide variant (SNV) has been developed in a recent decade for functional annotation of somatic mutations and predictions of putative driver mutations (Cheng, Zhao, & Zhao, 2017; Ding, Wendl, McMichael, & Raphael, 2014; Raphael, Dobson, Oesper, & Vandin, 2014). Computational prediction methods were focused on mutations in the protein-coding regions (Sim, et al., 2012; Adzhubei, et al., 2010; Chun & Fay, 2009; Reva, Antipin, & Sander, 2011; Schwarz, Rodelsperger, Schuelke, & Seelow, 2010; Gonzalez-Perez & Lopez-Bigas, 2011; Choi, Sims, Murphy, Miller, & Chan, 2012; Shihab, et al., 2013) and several sequence-based scores (Davydov, et al., 2010; Garber, et al., 2009) were successfully used for prediction of cancer driver mutations in non-coding regions. Combined Annotation-Dependent Depletion (CADD) (Kircher, et al., 2014) and Genome-Wide Annotation of Variants (GWAVA) (Ritchie, Dunham, Zeggini, & Flicek, 2014) methods supported characterization and classification of non-coding variants by combining various genomic annotations into integrated score measures with the aid of support vector machine models. Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) (Carter, et al., 2009) and Cancer Driver Annotation (CanDrA) (Mao, et al., 2013) are cancer-specific machine learning approaches utilizing structural, evolutionary and genetic features computed by multiple prediction algorithms. Cancer-Related Analysis of VAriants Toolkit (CRAVAT) is a web-based CHASM application for prioritization of genes and variants important for specific cancer tissue types (Douville, et al., 2013; Masica, et al., 2017).

The efforts to consolidate and maintain a comprehensive functional annotation for SNVs discovered in exome sequencing studies, a database of human nonsynonymous SNVs (dbNSFP) was developed as one-stop resource for analysis of disease-causing mutations (Liu, Jian, & Boerwinkle, 2011; Liu, Jian, & Boerwinkle, 2013; Liu, Wu, Li, & Boerwinkle, 2016). Based on a detailed comparison and machine learning-based integration of 18 prediction scoring method for nonsynonymous SNVs, the two ensemble

scores RadialSVM and LR were developed that outperformed their 10 component scores (Dong, et al., 2015). The latest database dbWGFP of functional predictions for SNVs collected nearly 8.58 billion possible human whole-genome SNVs, with a capability to compute a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches, 15 conservation features from 4 different tools including ensemble-based predictors RadialSVM, LR and MSRV scores (Wu, et al., 2016).

Detecting missense mutation hotspot regions in 3D protein structures represented a fruitful approach for identifying driver mutations. Structurally conserved mutational hotspots can be shared by multiple kinase genes and are often enriched by cancer driver mutations with high oncogenic activity (Dixit, et al., 2009). A statistically rigorous algorithm HotMAPS finds clusters of amino acid residues with significantly increased local mutation density in structural space, enabling identification of hotspot regions and suggesting the increased sensitivity to hotspot regions in tumor suppressor genes (Tokheim, et al., 2016). Structural analysis of somatic missense mutations across 32,445 protein structures from 7390 genes has identified and characterized mutational hotspot clusters, showing that they may represent robust spatial signatures of cancer driver mutations (Gao, et al., 2017). A computational tool, HotSpot3D was applied to >4,400 TCGA tumors across 19 cancer types, discovering >6,000 intra- and intermolecular clusters and 369 rare mutations all mapping within clusters having potential functional implications (Niu, et al., 2016). Recent studies have indicated that cancer missense mutations can target protein interaction interfaces and structural mutational hotspots may be enriched at the important protein binding interfaces, pointing and point to functional sites and interactions potentially perturbed in cancer genes (Kamburov, et al., 2015; Engin, Kresiberg, & Carter, 2016).

The machine learning integrated scores have indicated that sequence and structure-based scores can frequently provide orthogonal information for specific types of genes, but the underlying molecular reasons for a weak consensus between functional features remains poorly understood and hidden in feature selection process. Modern deep learning methods can leverage large data sets for finding hidden patterns and making robust predictions in cancer genomics, and drug discovery applications (Angermueller, Parnamaa, Parts, &

Stegle, 2016; Zhang, Tan, Han, & Zhu, 2017; Min, Lee, & Yoon, 2017; Jing, Bian, Hu, Wang, & Xie, 2018; Zhou & Troyanskaya, 2015; Yuan, et al., 2016). However, it is often overlooked that the performance and interpretability of machine learning models are equally important for predictions and understanding of cancer mutation signatures.

In this study, we developed two machine learning classifiers, random forest and logistic regression by training on cancer-specific "golden" sets of functionally validated mutations (Mao, et al., 2013; Martelotto, et al., 2014) and using a set of diverse feature scores that included computations of 48 functional scores using dbWGFP server (Wu, et al., 2016). By examining sequence, structure-based and ensemble-based integrated features, we show that evolutionary conservation scores play a more significant role in classification of cancer drivers and provide the strongest signal for the machine learning prediction. We apply the developed RF and LR models for prediction of driver mutations in Cbioportal cancer genomics dataset by considering all different cancer subtypes and 145,601 mutations including duplicates from multiple samples from 310 genes. We assess the prediction performance through a comparative analysis against functional experiments and multi-center mutational calling data from Pan Cancer Atlas studies (Ellrott, et al., 2018; Bailey, et al., 2018; Ng, et al., 2018). To address interpretability of machine learning approaches, we map our cancer driver predictions against the catalog of 3D cluster mutations (Gao, et al., 2017) and positions of activating mutations hotspots. This analysis reveals an enrichment of predicted tumor suppressor driver mutations in structural clusters and suggests novel hotspot clusters of potential driver mutations, while classified oncogene driver mutations are primarily aligned with gain-of-function activating mutations. By using machine learning results, we examine conformational mobility and structure-based network properties of residue positions enriched by predicted driver mutations. We show that the greater flexibility of specific functional regions targeted by driver mutations in oncogenes may facilitate activating conformational changes and acquisition of constitutively active oncogenic states, while loss-of-function driver mutations in tumor suppressor genes can preferentially target structurally rigid and

network-centric positions that are responsible for mediating protein stability and modulation of protein binding interfaces.

2.2 Mutational Datasets

In the initial stage of training and validating classifier models, we employed several 'gold standard' benchmarking data sets of manually curated, functionally- validated mutations. The first set included a total of 3,591 SNVs from several oncogenes (BRAF, KIT, PIK3CA, KRAS, EGFR, and ERRB2), recently described cancer genes (DICER1, ESR1, IDH1, IDH2, MYOD, and SF3B1), and major tumor suppressor genes (TP53, BRCA1, and BRCA2) (Martelotto, et al., 2014). In the original study, these SNVs were initially classified as non-neutral, neutral or uncertain, but due to binary classification model employed here (drivers and passengers), neutral or uncertain SNVs were assigned as passengers. The machine learning model was trained and validated only on missense mutations. The first benchmarking contained 3,706 mutations, with only 3,591 SNVs that included 140 neutral (assigned as passengers), 849 non-neutral (assigned as drivers) and 2,602 of uncertain function (assigned as passengers). The second employed data set was taken from the original CanDrA study (Mao, et al., 2013) with 1,550 SNVs. The CanDra data set (Mao, et al., 2013) was initially obtained by combining glioblastoma multiforme (GBM) and ovarian carcinoma (OVC) mutational data extracted from TCGA19 and COSMIC repositories (Forbes, et al., 2015). In the GBM sets, 134 SNVs were drivers and 585 SNVs were passenger mutations, while in the OVC sets 122 SNVs were driver mutations and 709 were passengers (Mao, et al., 2013). After the CanDra datasets (Mao, et al., 2013) and benchmarking datasets (Martelotto, et al., 2014) were gathered and processed, the two datasets were consolidated and combined to create one master training set used for the construction of machine learning models (Supplemental Tables S1,S2).

Machine learning models are used for large scale prediction of cancer driver mutations in the Cbioportal cancer genomics dataset by considering 17 cancer subtypes and a total of 80,081 unique missense mutations after excluding 65,520 duplicate mutations in patients with multiple samples from 310 cancer genes (Cerami, et al., 2012; Gao, et al., 2013). We considered all missense mutations from the following cancer

subtypes that are collected in Cbioportal database : glioblastoma, ovarian cancer, prostate cancer, cell cycle control, p53 signaling, notch signaling, DNA damage response, other growth/proliferation signaling, survival/cell death regulation signaling, telomere maintenance, RTK signaling family, PI3K-AKT-mTOR signaling, Ras-Raf-MEK-Erk/JNK signaling, regulation of ribosomal protein synthesis and cell growth, angiogenesis, folate transport, and invasion and metastasis (Supplemental Tables S1,S2). A total of 56,634 unique missense mutations were finally examined and classified by the machine learning models.

2.3 Mutational Predictor Scores: Feature Selection and Feature Importance Analysis

The initially selected features were obtained by computing functional scores using a database and web server dbWGFP of functional predictions for human whole-genome single nucleotide variants that provided 32 functional prediction scores and 15 conservation features (Wu, et al., 2016). Some of the score features (SIFT (Sim, et al., 2012), PolyPhen (Adzhubei, et al., 2010), LRT (Chun & Fay, 2009), Mutation Assessor (Reva, Antipin, & Sander, 2011), MutationTaster (Schwarz, Rodelsperger, Schuelke, & Seelow, 2010), FATHMM (Shihab, et al., 2013), RadialSVM (Dong, et al., 2015), LR (Dong, et al., 2015), MSRV (Wu, et al., 2016) and SinBaD) can be applied only to SNVs in protein coding regions, while other scores (GERP++ (Garber, et al., 2009), SiPhy (Garber, et al., 2009), PhyloP (Garber, et al., 2009), Grantham, CADD (Kircher, et al., 2014) and GWAVA (Ritchie, Dunham, Zeggini, & Flicek, 2014)) can evaluate SNVs spreading over the whole genome. The ensemble-based prediction scores RadialSVM and LR are machine-learning integrated features based on 10 component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP), and the maximum frequency observed in the 1000 genomes populations (Liu, Wu, Li, & Boerwinkle, 2016). We computed 48 feature scores for each SNV in our combined data set (Wu, et al., 2016). The processed dataset was split into training and test sets. The test set contained 20% of the samples from the original dataset, ensuring that the distribution of drivers and passengers was equivalent to that of the original dataset. The training set was subjected to recursive feature elimination process, where one by one a feature is removed, and the model

trained on the resulting dataset. If the accuracy of the model stays above a predefined threshold, the feature is removed permanently, and the process is repeated. If removal of any feature increases the accuracy, the threshold increases also. Single value elimination was used. We set a threshold of 0.001 less accuracy to declare a feature important to prediction, resulting in a final dataset of 32 features. The test set contained 20% of the samples from the original dataset, ensuring that the distribution of drivers and passengers was equivalent to that of the original dataset.

2.4 Protein Structure Networks and Network Centrality Analysis

For network-based analysis, a graph-based representation of protein structures is employed in which residues are treated as network nodes and inter-residue edges represent residue interactions (Chakrabarty & Parkekh, 2016; del Sol, Fujihashi, Amoros, & Nussinov, 2006; del Sol, Fujihashi, Amoros, & Nussinov, 2006; Vijayabaskar & Visheshwara, 2010; Stetz & Verkhivker, 2015; Stetz & Verkhivker, 2016). We used NAPS approach (Chakrabarty & Parkekh, 2016) that allows for rapid construction of residue interaction networks with unweighted or weighted edges, and subsequent residue-based network centrality analysis. For our analysis, an interaction strength-based graph representation of protein structures was used in which a residue is considered as node in the network and an edge is constructed if the interaction strength between two residues is more than the threshold of 4%. The interaction strength between two amino acid side chains is evaluated as follows:

$$I_{ij} = \frac{n_{ij}}{\sqrt{(N_i \times N_j)}} \times 100$$

where n_{ij} is number of distinct atom pairs between the side chains of amino acid residues i and j that lie within a distance of 4.5 Å. N_i and N_j and are the normalization factors for residues and respectively (Vijayabaskar & Visheshwara, 2010). The normalization factors take into account the differences in the sizes of the side chains of the different residue types and their propensity to make the maximum number of contacts with other amino acid residues in protein structures (Chakrabarty & Parkekh, 2016). The pair of

residues with the interaction I_{ij} greater than a user-defined cut-off (I_{min}) are connected by edges and produce a protein structure network graph for a given interaction cutoff I_{min}. The interaction strength I_{ij} is considered as edge weight. Noteworthy, owing to a large number of analyzed crystal structures of cancer genes, the edges in the residue interaction networks were weighted only based on the defined interaction strength (Chakrabarty & Parkekh, 2016) and did not employ a more detailed model with coevolutionary mutual information (Stetz & Verkhivker, 2017) and dynamic residue correlations couplings from molecular dynamics simulations (Sethi, Eargle, Black, & Luthey-Schulten, 2009).

Using the constructed protein structure networks, the residue-based betweenness parameters were also computed with the NAPS server (Chakrabarty & Parkekh, 2016). The betweenness of residue is defined to be the sum of the fraction of shortest paths between all pairs of residues that pass-through residue:

$$C_b(n_i) = \sum_{j<k}^{N} \frac{g_{jk}(i)}{g_{jk}}$$

g_{jk} denotes the number of shortest geodesics paths connecting j and k , and $g_{jk}(i)$ is the number of shortest paths between residues j and k passing through the node n_i. Residues with high occurrence in the shortest paths connecting all residue pairs have a higher betweenness values. For each node n, the betweenness value is normalized by the number of node pairs excluding n given as $(N-1)(N-2)/2$, where N is the total number of nodes in the connected component that node n belongs to.

2.5 Machine Learning Classification of Cancer Driver Mutations on Canonical Datasets: Ensemble-Based and Sequence Conservation Features Consistently Outperform Structural Prediction Scores

We first trained random forest (RF) and logistic regression (LR) machine learning models by considering a combination of two canonical cancer-specific sets of functionally validated mutations (Mao, et al., 2013; Martelotto, et al., 2014) using a set of diverse features that included functional scores obtained from dbWGFP server (Wu, et al., 2016). In this analysis, we compared the performance of two classifiers and

focused on identifying a group of shared dominant features that drive machine learning predictions of these models (Figure 3). Consistent with previous studies (Dong, et al., 2015), the integrated ensemble-based scores LR and RadialSVM dominated the feature importance distribution in both models, outweighing the contributions of other features (Figure 3A,B). In the RF model, the LR and RadialSVM scores were the top ranked features with the information value scores of 0.27 and 0.23 respectively, followed by a group of sequence-based evolutionary conservation features (Figure 3A). Some of the evolutionary conservation scores derived from multiple sequence alignments and reflecting functional specificity, such as Mutation Assessor (Reva, Antipin, & Sander, 2011) GERP++ (Davydov, et al., 2010), GerpRS (Davydov, et al., 2010), SiPhy (Garber, et al., 2009), and PhyloP (Garber, et al., 2009) also showed appreciable information score values (Figure 3A). Noteworthy, RadialSVM and LR ensemble features are based on integrating 10 component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP), with some of them also contributing individually to the model performance. Although these ensemble-based scores reversed their ranking in the LR model, they remained the top 2 features with the information scores of 0.441 and 0.332 for RadialSVM and LR scores respectively (Figure 3B). For the LR model, top ranking features also included MSRV, SinBAD_HVAR and SinBAD_HGMD scores. Of notice, SinBaD scores were originally derived by a logistic regression model with 90 binary features obtained from multiple sequence alignment designed for evaluation of mutational effects in protein coding and promoter regions (Lehmann & Chen, 2013). MSRV is another integrated feature developed by machine learning from a set of 24 physiochemical properties and several conservation scores to prioritize disease-causing nonsynonymous SNV mutations (Jiang, et al., 2007). Since both models

revealed the RadialSVM and LR scores as the two most important ranked features, we carried out a Wilcoxon signed-rank statistical test and confirmed that there is a statistical difference (p-value < 2.2e-16) between these two mutational scores. To assess the ability of the model to recapitulate classification performance in the absence of two ensemble-based metrics, we repeated our experiments by removing these features and ranked the importance of the remaining mutational scores (Figure 3C, D). Interestingly, the top ranked features corresponded primarily to functional scores based on evolutionary conservation patterns, such as Mutation Assessor derived using combinatorial entropy formalism (Reva, Antipin, & Sander, 2011), GerpN neutral evolution score and Gerp element scores (Davydov, et al., 2010; Garber, et al., 2009) that are identified by quantifying position-specific substitution deficits in multiple alignments. Another highly ranked feature was also likelihood ratio test (LRT) score that adopted the log-likelihood ratio of the conserved relative to the neutral model to predict functional significance of mutations (Chun & Fay, 2009). Interestingly, some of the highly ranked features (PhyloP, SiPhy, Grantham) were designed to provide prediction scores for variants spreading over the whole genome. At the same time, feature scores that assess mutational variants in protein coding regions (SIFT, PolyPhen2_HVAR, LRT, MutationTaster,)

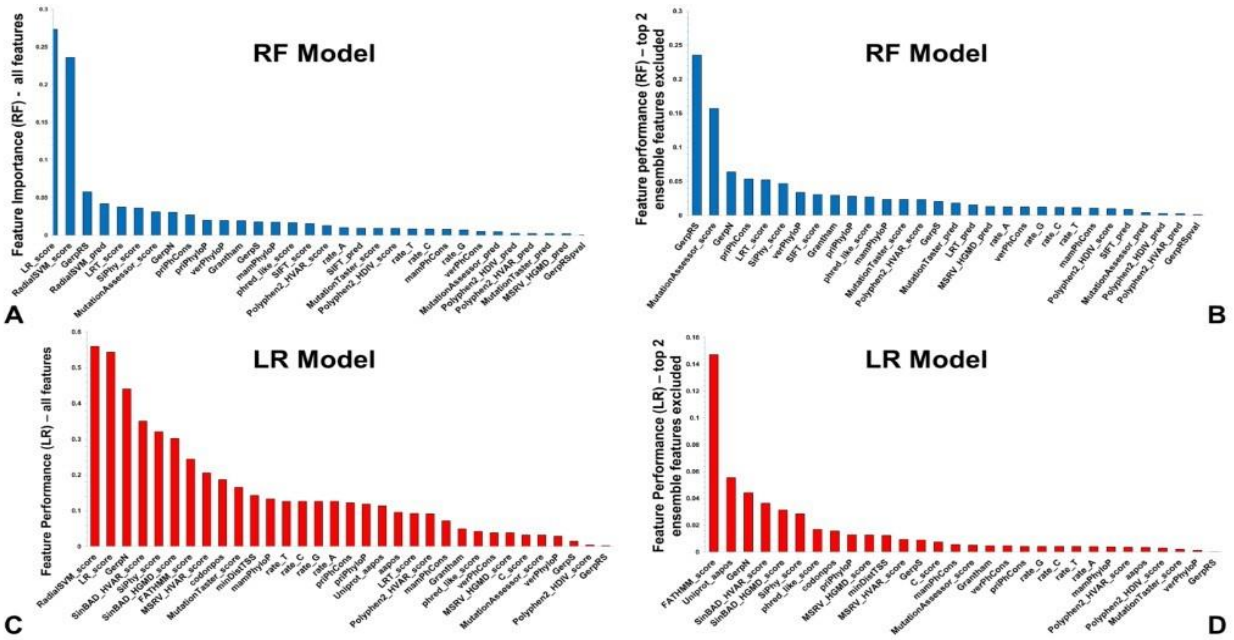


Figure 2. Feature Importance Analysis of the RF and LR Machine Learning Models on the Canonical Cancer-Specific Dataset of Functionally Validated Mutations

contributed only moderately to the feature performance (Figure 3C,D). These findings are consistent with previous analysis of deleterious SNV mutations (Dong, et al., 2015), indicating that ensemble-based scores and functional features based primarily on evolutionary conservation measures and statistical descriptors of substitution patterns may allow for robust classification of cancer driver mutations. Our results suggested that cancer-specific conservation features can outperform those designed for detection of pathogenic variants in coding regions. Despite differences in the prediction scores, the obtained highly important features reflected a common fundamental signature that is mutations of evolutionarily conserved residues in functional regions are likely to be deleterious. As a result, the probabilistic evaluation of deleterious mutations that underlies most of the dominant features appeared to be sufficient for robust classification of driver mutations in the canonical cancer datasets (Figure 3).

Collinearity is often a strong indicator of redundancy of prediction feature pairs. To evaluate possible redundancies, we next computed and analyzed pairwise correlations between different prediction scores with the Spearman's rank correlation coefficient (Figure 4). According to this analysis, two dominant feature scores RadialSVM and LR are highly correlated only with one of its original component scores,

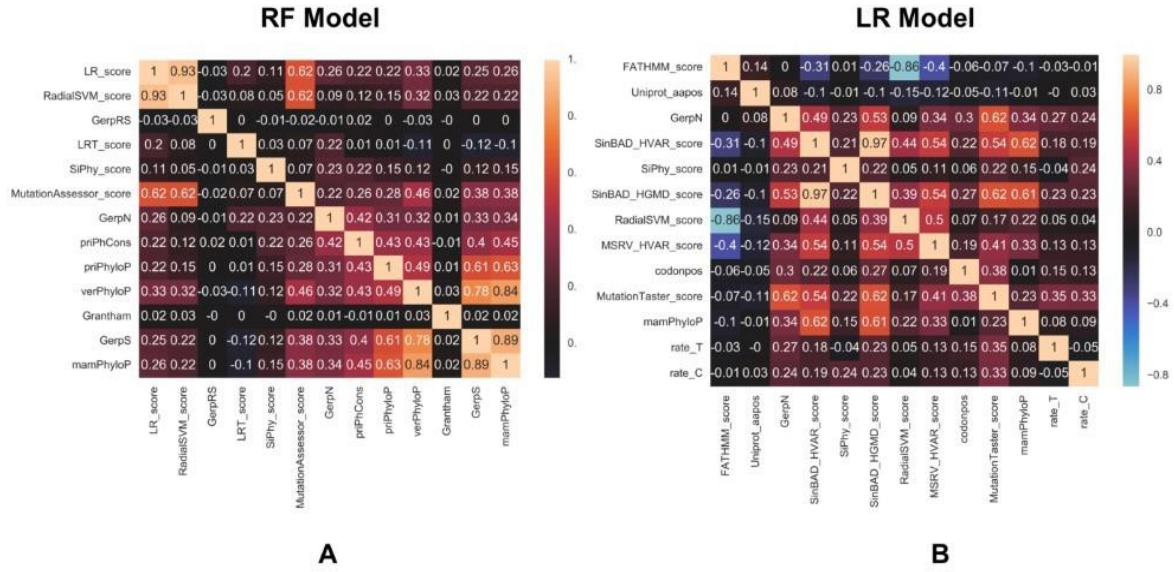


Figure 3. The Pairwise Spearman's Rank Correlation OCefficients Between Different Prediction Scores

Mutation Assessor, and moderately correlated with other top performing sequence-based features (PhyloP, PhCons, GerpN, GerpRS, and GerpS). At the same time, the ensemble scores are relatively weakly correlated with some of their integral components (LRT, SiPhy, and PhyloP). In addition, GerpRS, LRT, and Grantham scores have a fairly low correlation with other scores in the RF model (Figure 4A). Evolutionary-based conservation scores PhyloP, PhCons, GerpN, GerpRS, and GerpS have only moderate correlation. In the LR model, we noted a significant correlation between RadialSVM and other ensemble-based scores MSRV, SinBAD_HVAR, and SinBAD_HGMD scores (Figure 4B). Some of the other important sequence-based scores GerpN, SiPhy have only a moderate correlation with other top features. In general, the ensemble scores and sequence-based features were superior according to feature importance ranking. Additionally, and perhaps not surprisingly, so-called population-based scores that distinguish pathogenic missense variants from common polymorphisms (SIFT, PolyPhen2, Mutation Assessor) were

found to be less important in our classification models that cancer-specific features (FATHMM) designed to differentiate somatic driver mutations (Figure 4).

We then compared the predictive performance of RF and LR models as well the contribution of individual features using area under the curve (AUC) from the receiver operating characteristic (ROC) plots, in which sensitivity (or true positive rate TPR) is plotted as a function of 1-specificity, where specificity is true negative rate TNR (Figure 5). These graphs showed the improved performance of the RF model after feature selection process was complete (AUC=0.97) as compared to the original AUC=0.85 for the initial set of features (Figure 5A). By examining the contribution of individual features to the performance of RF model, we found that sequence-based Gerp element score GerpRS achieved as high performance in the testing dataset (AUC=0.91) as the ensemble-based LR score (AUC=0.88) and RadialSVM score (AUC=0.87) (Figure 5B).

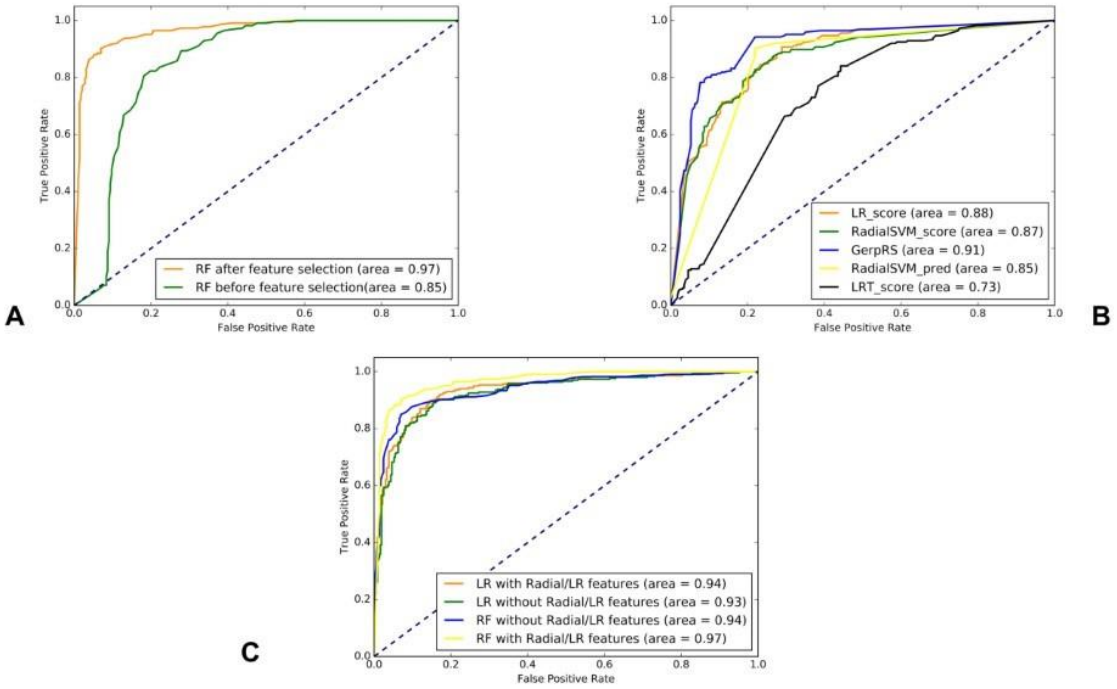


Figure 4. The ROC plots of Sensitivity (TPR) as a Function of Specificity (TNR).

These observations supported our conclusions that evolutionary conservation scores can often outperform other features including ensemble based LR and RadialSVM scores. A comparative AUC analysis in the presence and absence of top ensemble scores showed only a minor effect on performance of both models

(Figure 5C). Noteworthy is a strong and similar performance of both models (AUC > 0.9) on the testing set when the top two ensemble features were excluded. This analysis confirmed that group sequence-based scores based on evolutionary conservation and statistical descriptors of substitution patterns may allow for robust machine learning classification of cancer driver mutations in canonical datasets.

The performance of classification models was also assessed using accuracy, recall, specificity, PPV, NPV and F-score values (Tables 1,2). The RF model achieved a higher classification accuracy of 0.906, correctly classifying 465 out of 513 held out samples (Table 1). While sensitivity (recall) values were similar for both models, RF model achieved a higher specificity, PPV and NPV values. A comparative analysis of individual top features in the RF model showed the highest accuracy for GerpRS, LR, and RadialSVM scores (Table 1), while in the LR model the most accurate predictors were RadialSVM, LR and FATHMM scores (Table 2).

METHOD/SCORE	***ACCURACY***	***TPR***	***TNR***	***PPV***	***NPV***	***F - SCORE***
RF model	0.9064	0.8672	0.9373	0.9159	0.8997	0.8909
LR_score	0.7973	0.8000	0.7460	0.8000	0.8159	0.8000
RadialSVM_score	0.7992	0.8000	0.7573	0.8000	0.8009	0.8000
GerpRS	0.8363	0.8400	0.8087	0.8400	0.8230	0.8400
RadialSVM_pred	0.8343	0.8300	0.7640	0.8500	0.7640	0.8300
LRT_score	0.6803	0.6800	0.6099	0.7000	0.7611	0.6800
SiPhy_score	0.6823	0.6800	0.6300	0.6900	0.6770	0.6800
MutationAssessor_score	0.7232	0.7200	0.7019	0.7200	0.6489	0.7200
GerpN	0.7797	0.7800	0.7703	0.7800	0.7124	0.7800
priPhCons	0.5906	0.5900	0.5325	0.5900	0.5796	0.5900
priPhyloP	.6823	0.6800	0.6759	0.6800	0.5354	0.6800

Table 1. The performance metrics and statistics of the RF model and individual mutational prediction scores on cancer-specific canonical dataset.

METHOD/SCORE	ACCURACY	TPR	TNR	PPV	NPV	F- SCORE
LR	0.8661	0.8858	0.8408	0.8776	0.8511	0.8817
LR_score	0.8321	0.8230	0.8427	0.8611	0.8007	0.8416
RadialSVM_score	0.8384	0.8230	0.8566	0.8719	0.8033	0.8467
FATHMM_score	0.8273	0.8982	0.7435	0.8056	0.8606	0.8493
priPhCons	0.5421	0.8943	0.7344	0.5421	0.7956	0.6755
Uniprot_aapos	0.6906	0.6327	0.7592	0.7566	0.6360	0.6891
verPhyloP	0.5923	0.6637	0.5079	0.6148	0.5607	0.6383
GerpN	0.5423	0.6753	0.6232	0.5424	0.6122	0.6015
MutationAssessor_score	0.7002	0.6991	0.7016	0.7349	0.6634	0.7165

Table 2. The performance metrics and statistics of the LR model and individual mutational prediction scores on cancer-specific canonical dataset.

Our results supported arguments that ensemble-based functional predictors and conservation predictors may have a higher sensitivity than structural scores. We proposed that sequence-based approaches and a group of emerging top conservation metrics can capture driver mutations overlooked by structure-based predictors, whereas structural tools may be used to complement and validate machine learning predictions to aid in the interpretability of models.

2.6 Classification of Missense Mutations in Cbioportal Cancer Genes: A Comparative Analysis with Functionally Validated Mutations and Structural Mutational Hotspots

Machine learning models are typically evaluated using accuracy metrics on available validation datasets. However, the large cancer genomics datasets are highly diverse, reflecting the heterogeneous mutational patterns in cancer genomes and covering a range of phenotypic impacts rather than “black and white” driver/passenger separations. Applying standard evaluation metrics to these sets is often insufficient to obtain new insights into molecular determinants of cancer-causing mutations. Inspecting and explaining individual examples and predictions is a worthwhile complementary approach to assess trustability of predictions. The insights given by dissection of specific predictions and explanations are particularly helpful in identifying ways to improve quality and trustability of the machine learning models.

We employed the developed RF model for the prediction and analysis of cancer driver mutations using missense variants from Cbioportal database (Cerami, et al., 2012; Gao, et al., 2013). The studied dataset

included a total of 80,081 missense mutations from 310 cancer genes (Supplemental Tables S1,S2). A total of 56,634 unique missense mutations were finally examined and classified by the RF model for which all functional, conservation and ensemble-based scores were computed using the dbWGFP web server (Wu, et al., 2016). In these experiments, the performance of a simple RF model in classification of cancer driver mutations was assessed through a detailed comparison with three different sets of functionally annotated driver mutations : a) functional data of 1,049 experimentally validated somatic mutations in various cell lines,33 b) PanCancer multi-center mutation calling data on 579 driver mutations identified by consensus of multiple sequence-based and structure tools,32 and c) structural analysis of mutational hotspot clusters that identified potential driver mutations in 3,405 residues of protein structures from 503 genes.20

At the onset of this analysis, we would like to emphasize that comparisons and correspondence of our machine learning predictions with functional validation experiments and other studies can evaluate and confirm the number of true positive predictions but gives little information regarding false negatives. Nonetheless, a comparative analysis can be effectively used to aid in the interpretability of machine learning

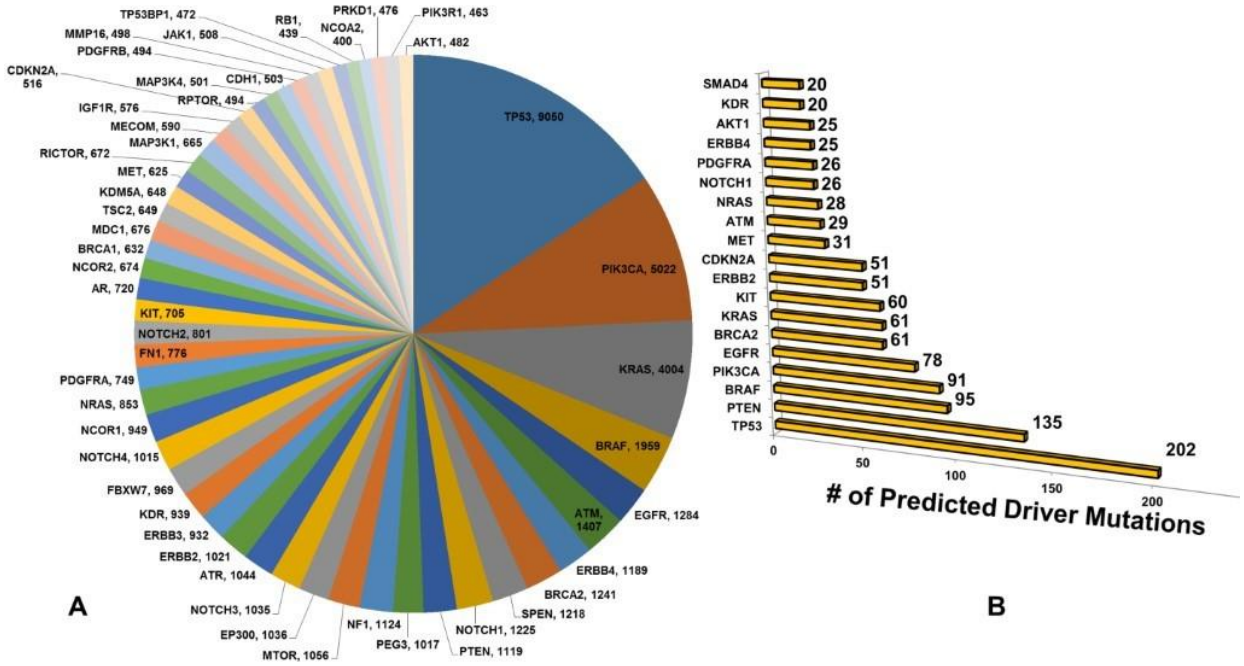


Figure 5. The gene-based distribution of examined cancer mutations and predicted driver mutations from Cbioportal cancer genomics dataset

predictions and identification a pool of novel potential driver mutations that may be strong candidate for follow-up experimental testing and validation.

The spectrum of cancer mutations analyzed by our models covers a highly representative group of well-known oncogenes and tumor suppressor genes, with the large number of missense mutations in TP53, PTEN, KRAS, BRAF, EFFR and other genes (Figure 6A). By applying the RF model to this dataset, we predicted the largest number of cancer driver mutations in TP53, PTEN, BRAF, PIK3CA, and EGFR (Figure 6B).

First, we matched our predictions against a set of 923 functionally validated driver mutations that were tested by high-throughput functional genomic platform sensitive to weak driver mutations (Ng, et al., 2018). Since our model is based on binary classification of missense mutations, to facilitate a meaningful comparison we considered activating, inactivating and inhibitory mutations from this experimental set as drivers, while neutral, non-inhibitory and undetermined variants were assigned as potential passengers. We predicted 380 cancer drivers based on comparison with consensus functional annotation, 345 drivers by mapping our results against MCF10A cancer cell line annotation, and 285 driver mutations in matching our predictions against Ba/F3 functional annotation (Figure 7A). Our predictions tended to slightly overestimate the number of cancer drivers (Figure 7A) while underestimating the number of passengers (Figure 7B). A direct overlap between the predicted driver mutations and experimentally validated mutations was significant (Figure 7A), providing support to the robustness of a simple model that can achieve a good accuracy in classification of missense mutations. These results are also consistent with the machine learning analysis of the canonical dataset, suggesting the higher sensitivity of the RF model and a tendency to capture a broader range of potential driver mutations. We suggest that even though a spectrum of predicted driver mutations may be larger than the experimentally tested group of drivers, the predicted cancer variants may not be merely an aberration and still be functionally relevant, potentially reflecting a different degree of driver effect (weak-to-strong) rather than a simple binary classification.

To facilitate functional interpretability of machine learning results, we compared the predicted driver mutations against validated activating mutations tested in MCF10A and Ba/F3 cancer cell and consensus functional annotation (Ng, et al., 2018). The RF model classified as drivers a total of 241 mutations from 263 experimentally validated activating mutations according to consensus functional annotation, predicted 231 driver mutations among 287 activating mutations in MCF10A cell line, and assigned 137 driver mutations from a total of 195 activating mutations tested in Ba/F3 cell line (Figure 7C). Hence, a significant fraction of experimentally validated activating mutations (~80%-90%) from different cell lines can be correctly classified as potential driver mutations. Although the machine learning model is largely determined by several ensemble-based and cancer-specific conservation features, it achieved a robust performance in classifying activating mutations in cancer genes that are often associated with structural effects by targeting localized functional regions and modulating functional conformational transitions. The functional dataset of 923 functionally validated missense driver mutations is dominated by activating mutations that constitute ~ 75% of total tested variants (Figure 7D). Not only our predictions classified a

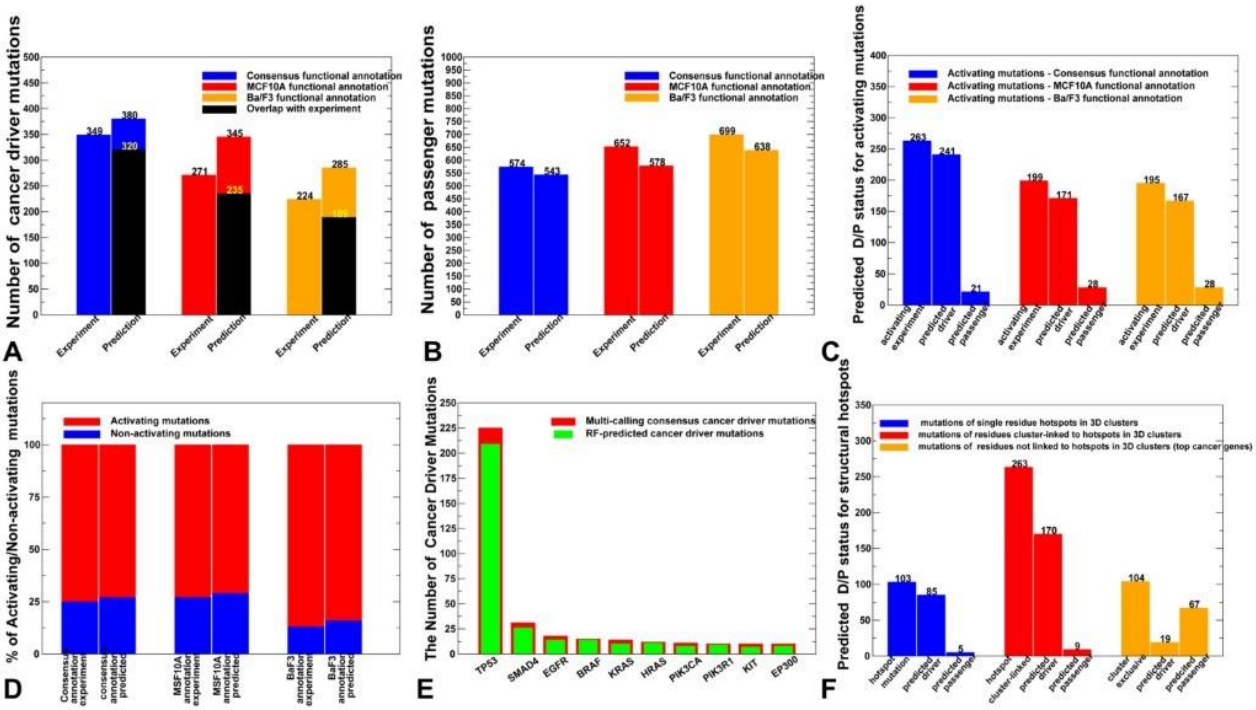


Figure 6.Machine learning predictions and functional comparisons of driver mutations in Cbioportal dataset.

significant fraction of validated cancer-causing mutations as potential drivers, but the model also preserved the same ratio of activating/inactivating mutations among predicted cancer driver mutations (Figure 7D).

We also compared our predictions with a recent comprehensive analysis of oncogenic driver mutations from PanCancer Atlas project (Bailey, et al., 2018). In this comparison we matched our predictions against a set of 579 mutations predicted by consensus of three categories of approaches: sequence-based tools distinguishing benign versus pathogenic mutations, sequence-based tools distinguishing driver versus passenger mutations, and structure-centric tools discovering statistically significant three-dimensional clusters of missense mutations. A significant fraction of these mutations was also experimentally tested and validated (Bailey, et al., 2018). By considering consensus 579 mutations that covered a large pool of most significant oncogenes and tumor suppressor genes, we found that our analysis classified ~85%-90% of these variations as cancer driver mutations (Figure 7E). Of particular interest is a strong correspondence between our predictions and consensus multi-calling results (Bailey, et al., 2018) for major cancer genes including TP53, SMAD4, KRAS, BRAF, EGFR, KIT, PIK3CA and PIK3R1 (Figure 7E).

To facilitate interpretability of machine learning results, we mapped our cancer driver predictions against the catalog of 3D clusters of mutational hotspots. According to previous studies, one of the hallmarks of cancer driver mutations is the emergence of statistically significant clusters of missense cancer mutations in protein structures (Tokheim, et al., 2016; Gao, et al., 2017; Niu, et al., 2016; Kamburov, et al., 2015; Engin, Kresiberg, & Carter, 2016). In this analysis, we specifically focused on a comparison with a set of single residue hotspots observed in 3D clusters (a total of 103 residues) and hotspot-linked sites that correspond to mutated residues clustered in protein structure with a known single residue hotspot (a total of 263 residues) (Gao, et al., 2017).

The RF classifier correctly identified a high percentage of single residue mutational hotspots and mutations coupled to these hotspots as potential drivers, while the distribution of mutations not coupled to a single residue hotspot revealed a clear shift towards passengers (Figure 7F). Our predictions retrieved 82% of single hotspot residues as cancer driver mutations, while only 64% of hotspot-linked residues were

classified as drivers (Figure 7F). According to these results, while functional impact in hotspot-linked cluster positions may be highly relevant, the likelihood of being a strong driver can be reduced. Importantly, we also observed that driver-to-passenger ratio among residues populating non-functional clusters lacking known mutational hotspots was shifted towards passengers (Figure 7F). Strikingly, nonetheless, an appreciable population of these cluster-exclusive residues was predicted as potential driver sites. Functional annotation and classification of these cluster-exclusive positions not linked to mutational hotspots is particularly intriguing and often overlooked even though these regions may harbor weak driver mutations that act cooperatively with mutational hotspots in cancer progression. We examined structural role of these potential driver positions and suggest specific viable candidates for further experimental validation.

2.7 Structure-Functional Analysis of Cancer Driver Mutations in Oncogenes and Tumor Suppressor Genes: Towards Interpretability of Machine Learning Predictions

To leverage our predictions beyond conventional comparisons with the experimental data and aim at extracting novel information about driver mutations, we conducted an extensive structure-based analysis of the predicted driver positions to characterize molecular signatures of driver sites in oncogenes and tumor suppressor genes. Based on this analysis, we propose for experimental testing a group of novel potential driver mutations that can act by altering structure, global dynamics and allosteric interaction networks in important cancer genes.

We attempted to first address these objectives by examining structural maps of the predicted driver mutations in the context of structural mutational hotspot clusters of well-known oncogenes (BRAF, EGFR, PIK3CA) and tumor suppressor genes (TP53, PTEN, SMAD4). While comparisons of our predictions with mutational hotspots supported the notion that structural clustering of missense mutations is a hallmark of oncogenes, we were intrigued by the observations suggesting a similar bias towards clustering of driver mutations in tumor suppressor genes. By performing structural mapping of the predicted driver mutations

in these oncogenes (Figure 8) and tumor suppressor genes (Figure 9), we generally observed a broader and more delocalized distribution of potential driver positions as compared to consensual mutational hotspots. Structural mapping of all predicted driver positions and mutational hotspot residues in the EGFR (Kovacs, Zorn, Huang, Barros, & Kuriyan, 2015; Shan, et al., 2012; Shan, Arkhipov, Kim, Pan, & Shaw, 2013) and BRAF activating dimer conformations (Rajakulendran, Sahmi, Lefrancois, Sicheri, & Therrien, 2009; Thevakumaran, et al., 2015) showed a considerable overlap, also highlighting preferential targeting of dynamic and exposed regions in the activation loop (Figure 8A,B). Mutational hotspot residues in these oncogenic kinases are located in the regulatory regions that are responsible for modulating functional conformational transitions between the inactive and active states (Kovacs, Zorn, Huang, Barros, & Kuriyan, 2015; Shan, et al., 2012; Shan, Arkhipov, Kim, Pan, & Shaw, 2013; Rajakulendran, Sahmi, Lefrancois, Sicheri, & Therrien, 2009; Thevakumaran, et al., 2015). Mutational hotspot residues in PIK3CA structures tend to be localized in protein interaction interfaces (Figure 8C). Cancer driver mutations E542K, E545K in this region can compromise the negative regulation of PIK3CA by preventing binding of phosphotyrosine peptides and inhibitory binding between nSH2 and PIK3CA proteins (Huang, et al., 2007; Thorpe, et al., 2017; Miller, et al., 2014; Mandelker, et al., 2009).

Of particular interest were several predicted drivers that reside near binding interfaces and spatially proximal to a group of known mutational driver hotspots E542, E545, Q546, E547, N564, K567, V344, N345, and C420. Among these potential candidates for further experimental validation were predicted S379T, N380S, and E418K mutations which target positions near the binding interface with PIK3R1 and are clustered together with sites of validated driver mutations N345K and C420R.32 Another group of predicted drivers included P539R, I391M, D549N and G451R/V mutations (Figure 8C). Although these variants have not been directly validated, our predictions are supported by several lines of experimental evidence. According to the experiments, I391M mutation increased cell proliferation and cell viability as compared to wild-type PIK3CA, in one of two different cell lines.33 D549N has not been biochemically characterized, but is predicted to confer a gain of function due to the increased transformation ability in two

different cell lines (Ng, et al., 2018). Another predicted mutation P539R can confer a gain of function, as indicated by constitutive phosphorylation of downstream proteins Akt and S6 (Gymnopoulos, Elsliger, & Vogt, 2007; Mankoo, Sukumar, & Karchin, 2009).

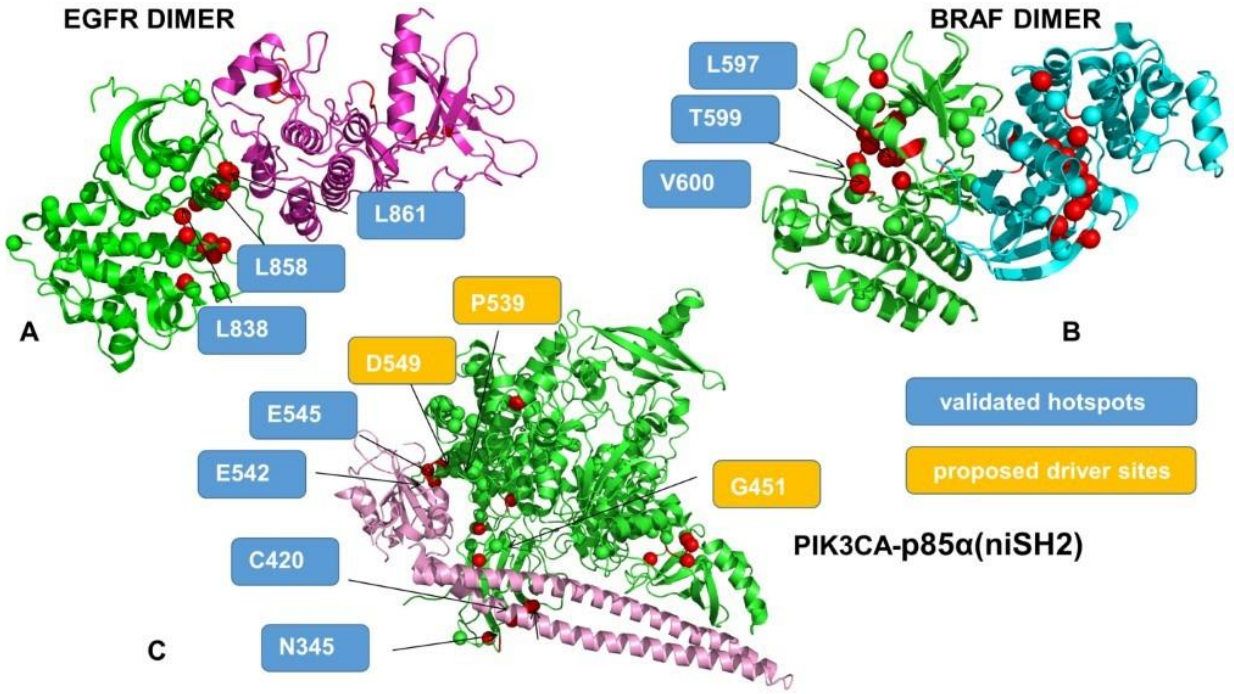


Figure 7. Structural mapping of the predicted driver mutations and validated mutational hotspot drivers in oncogenes

A considerable structural overlap between mutational hotspots and predicted driver positions confirmed preferential localization of these positions near the binding interfaces, particularly with PIK3R1 where functional residues from both partners can form clusters of both validated and non-tested mutations (Bailey, et al., 2018). In addition, we observed that the proposed driver positions are broadly distributed in the protein structures, targeting functional regions that are more dynamic and solvent-exposed regions, likely to facilitate activating conformational changes and enable structural plasticity required for access to diverse binding partners (Figure 8). These findings are illustrated by structural mapping of activating driver

mutations for EGFR, BRAF and PIK3CA genes (Supporting Information, Figure S1) that were experimentally validated and also emerged from consensus functional annotation of PanCancer multi-center mutation calling data (Bailey, et al., 2018). It is evident that activating driver mutations are consolidated in localized functional regions near binding interfaces that modulate activity of these oncogenes. We argue that the predicted driver positions may be functionally relevant and act as weak drivers that cooperate with spatially proximal major hotspots in exerting a cumulative effect on cancer progression.

We also mapped positions of known 3D mutational hotspot clusters (Gao, et al., 2017) and predicted driver positions in TP53 and PTEN tumor suppressor genes (Figure 8). The key hotspot residues that are most frequently mutated are located near the TP53-DNA binding interface (R248, R273) and correspond to positions that when altered can perturb the structure of the TP53-DNA binding surface (R175, G245, R249 and R282). Interestingly, the predicted driver mutations in TP53 occupied structurally diverse positions that were in the close proximity of mutational hotspots, resided near DNA-binding interface, and effectively linked spatially different mutational hotspots (Figure 9A). To enable a complete structural mapping of all predicted driver positions, we used the crystal structure of a multidomain TP53 oligomer bound to the CDKN1A(p21) p53-response element (Emamzadah, Tropia, & Halazonetis, 2011) (Figure 9A). In this case, we observed a much broader and delocalized distribution of potential driver positions where clusters seem to be formed in more flexible interacting regions and binding interfaces of a multidomain TP53 oligomer bound to the CDKN1A–p53-response element (Figure 9A). The difference in structural mapping of hotspot positions and all predicted driver residues is apparent, revealing a high density of hotspot residues in the core and DNA-binding interfaces of the tetramer, while the predicted driver positions are more broadly distributed and enriched in homo-oligomerization sites.

We specifically examined a group of residues that are not linked to known hotspot clusters but nonetheless were predicted as potential driver positions in TP53 (Figure 9A). Among predicted driver mutations in TP53 that targeted residues lacking functional annotation were A159P, A161V, A189T, Q331H, R337C,

R337H, F341S, G334W and R342P. Residues A159, A161 and A189 are located in the core of the DNA binding domain but have not been biochemically characterized. Other residues in this list can be structurally

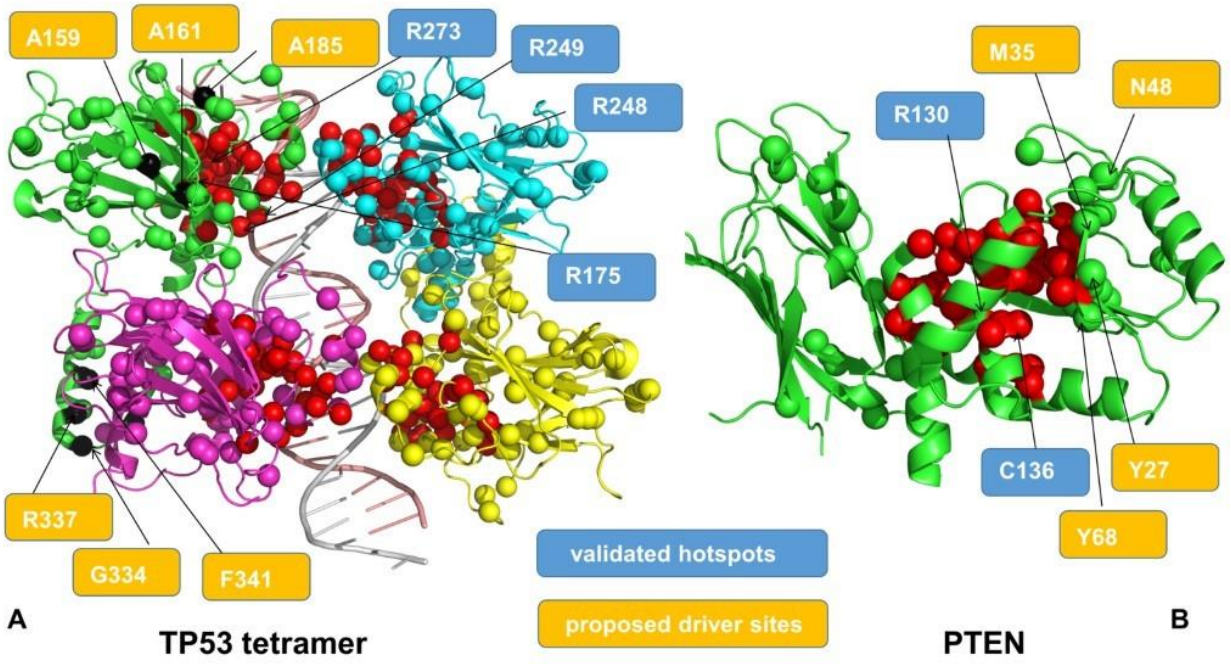


Figure 8. Structural mapping of the predicted driver mutations and validated mutational hotspot drivers in tumor suppressor genes

assessed only in the TP53 tetrameric form (Figure 9A). The tetramerization domain of TP53 harbors loss-of-functions mutations in these positions that compromise the ability to form the tetrameric structure required for TP53 function. In particular, R337 lies within the tetramerization domain of the TP53 protein (Joerger & Fersht, 2007; Kamada, Nomura, Anderson, & Sakaguchi, 2011) and R337H results in the decreased TP53 tetramerization and transactivation activity in cell culture (Imagawa, Terai, Yamada, Kamada, & Sakaguchi, 2008) and increased TP53 nuclear accumulation in patient samples (Seidinger, et al., 2015). G334 is also located at the oligomerization domain of the TP53 protein and G334W leads to a loss of TP53 transactivation activity in yeast (Kawaguchi, et al., 2005). Mutations F341S/V have not been biochemically annotated but impair formation of TP53 tetramers and induce the decreased TP53 transcriptional activity in yeast (Kawaguchi, et al., 2005).

This analysis revealed a broad spatial distribution of predicted driver positions and tendency of these residues to occupy more dynamic regions, including a group of sites involved in homo-oligomerization interfaces (Figure 9A). Although mutated oligomerization sites can be equally represented among oncogenes and tumor suppressors, only tumor suppressors are significantly enriched with functional mutations in these regions (Engin, Kresiberg, & Carter, 2016). Structural mapping of mutational hotspots in PTEN showed enrichment in the protein core, while the predicted driver positions similarly expanded to more dynamic regions that may be involved in protein binding interactions (Figure 9B). We were particularly intrigued by a group of predicted driver mutations that were not linked to any known mutational hotspots such as Y27C, M35R, N48I/K, P38L, I33S, and Y68C/D. Interestingly, even though these variants have not been biochemically validated as potential drivers, there is a substantial evidence pointing to functional importance of these mutations that may impair PTEN interactions with downstream partners (Rodriguez-Escudero, et al., 2011; Vega, et al., 2003) affecting PTEN-AKT3 signaling cascade and contributing to cancer development (Madhunapantula & Robertson, 2009).

To summarize, the predicted driver residues that coincide with validated mutational hotspots tend to occupy structurally stable positions and are often consolidated inside the protein core, especially for tumor suppressor genes, or localized at several dense binding interfaces. At the same time, some of the driver positions often occupy more dynamic regions that can be involved in diverse protein-protein binding interfaces clusters. By revealing a broader spectrum of potential cancer driver variations and their structural preferences, our results highlighted limitations of the binary driver/passenger classification, suggesting that functionally relevant cancer mutations may span a continuum spectrum (weak-to-strong) of driver effects.

2.8 Distinct Dynamic Signatures of Predicted Cancer Driver Mutations in Oncogenes and Tumor Suppressor Genes

Based on our predictions and structural analysis, we suggested that dynamics profiles of functional sites targeted by cancer driver mutations could be different in oncogenes and tumor suppressor genes. We

employed a coarse-grained model for large-scale characterization of residue rigidity/flexibility profiles in crystal structures of 3D-hotspot annotated oncogenes and tumor suppressor genes (Gao, et al., 2017). By using flexibility-rigidity index (FRI) method (Opron, Xia, & Wei, 2014; Nguyen, Xia, & Wei, 2016; Opron, Xia, Burton, & Wei, 2016) which is a robust matrix decomposition-free method that utilizes topological network connectivity in protein structures to derive a kernel generalization of the local density model, we analyzed the distribution of rigid and flexible regions in the protein structures harboring predicted driver mutations. We first computed the distribution of residue-based solvent-accessible surface area (SASA) (Figure 10A,B) which can be rapidly estimated for a large number of structures using analytical equations and their first and second derivatives as implemented in the web server GetArea (Fraczkiewicz & Braun, 1998). These distributions were obtained by using all crystal structures employed in the 3D-hotspot data set (Gao, et al., 2017) and we compared respective densities for all residues, mutational hotspot sites and predicted driver mutation positions. While the overall residue-based SASA distributions were similar for oncogenes and tumor suppressor genes with a dominant peak corresponding to more buried residues, an appreciable shift towards the increased solvent exposed sites was seen in the predicted drivers of oncogenes

(Figure 10A). In contrast, for tumor suppressor genes, all distributions revealed prevalence of mostly buried residues among mutational hotspots and predicted driver positions (Figure 10B).

Using the dynamic profiles for the crystal structures of oncogenes and tumor suppressor genes, we attempted to identify and characterizes molecular signatures of cancer driver mutations that can uniquely

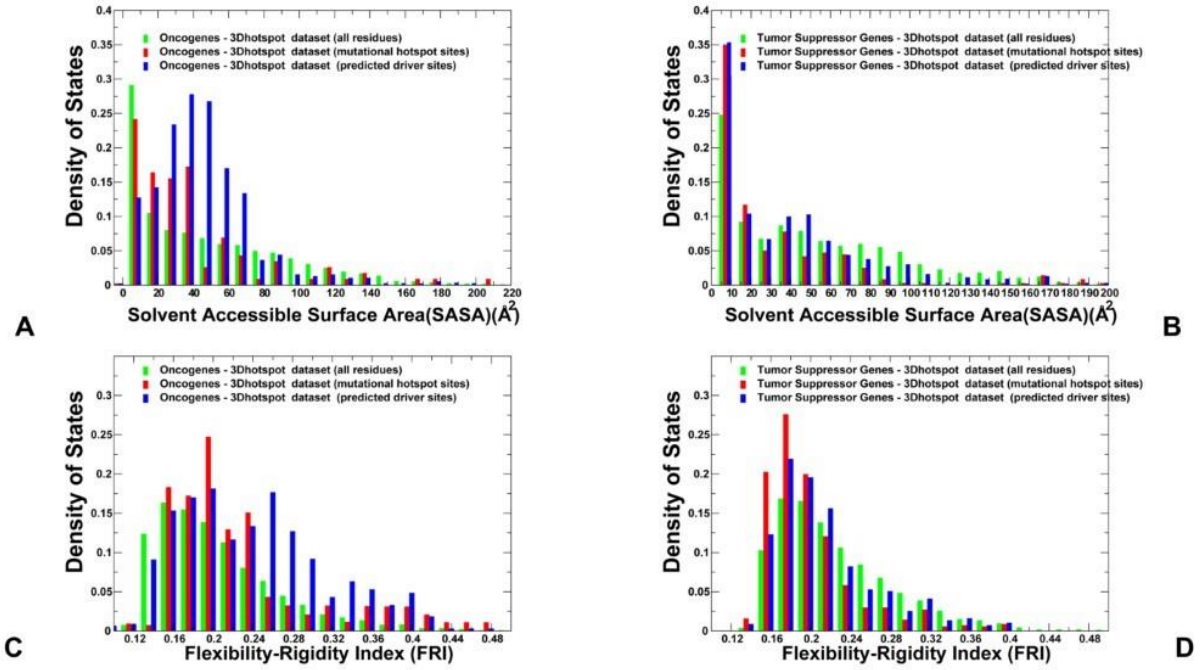


Figure 9. The distributions of residue-based solvent-accessible surface area (SASA) and flexibility-rigidity index (FRI) in the crystal structures of oncogenes and tumor suppressor genes.

define these classes of cancer proteins. For this analysis, we compared the residue-based distribution of conformational mobility against densities obtained for annotated mutational hotspot residues and the predicted driver positions (Figure 10C,D). The overall density is characterized by a well-defined peak of low FRI values (structurally stable residues) and a tail of higher FRI values corresponding to more flexible regions. The mobility density of structural mutational hotspots in oncogenes showed a shift towards larger FRI values (more flexible regions) and the predicted driver positions showed even a more pronounced redistribution towards larger FRI values, implying the increased average mobility of residues targeted by predicted driver mutations (Figure 10C). In a sharp contrast, for tumor suppressor genes, the mobility distributions for mutational hotspots and predicted driver sites showed the increased peak at low FRI values,

signaling that these functional sites are preferentially localized in structurally stable regions inside the core or at the stable binding interfaces (Figure 10D).

A more detailed analysis of mobility profiles in two prominent oncogenes BRAF and PIK3CA (Supporting Information, Figure S2) highlighted preferential concentration of driver positions in specific functional regions. Although most BRAF mutants display elevated kinase activity compared to the wild type, several cancer driver mutants in the conserved DFG motif of the A-loop (G466E, G466V, G596R and D594V) are inactivating alterations. We observed that the corresponding residues featured low FRI values and are structurally stable (Supporting Information, Figure S2). This suggests that kinase-dead mutations may occur in the regions of higher structural stability, and in the case of D594V mutation may lead to the increased thermodynamic stability of the inactive kinase form (Rajakulendran, Sahmi, Lefrancois, Sicheri, & Therrien, 2009; Thevakumaran, et al., 2015). This mutation can increase steric barrier for conformational transitions to the active conformation, thus rendering D594V as a kinase-inactivating mutation. In contrast, activating driver mutations (L597V/S/Q/R/L, A598V, T599I/A, V600E/K/R/M/R) target more dynamic residues of the functional regions in the kinase N-terminal lobe that are prone to conformational changes (Supporting Information, Figure S2), inducing functional transitions to the active form of BRAF kinase. A similar mechanism underlies dynamic preferences of various PIK3CA driver mutations (Supporting Information, Figure S2) that stimulate lipid kinase activity by mimicking and enhancing dynamic events and allosteric motions that occur in the wild-type enzyme (Burke, Perisic, Masson, Vadas, & Williams, 2012). These unifying dynamic signatures of driver mutations in oncogenes may be associated with the underlying mechanism of gain-of-function activating mutations that promote activating conformational transformations, alter the balance between the inactive and active states, and ultimately induce thermodynamic stabilization of a constitutively active state (Rajakulendran, Sahmi, Lefrancois, Sicheri, & Therrien, 2009; Thevakumaran, et al., 2015; Tse & Verkhivker, 2016; Stetz, Tse, & Verkhivker, 2017).

In some contrast, dynamic residue profiles derived from crystal structures of tumor suppressors TP53 and PTEN showed that loss-of-function driver hotspots in these genes occur at spatially distinct regions in and

broadly distributed, yet these cancer mutations preferentially target structurally stable residues that featured low FRI values (Supporting Information, Figure S2). Using this analysis to improve interpretability of machine learning results, we suggested that the expanded pool of predicted driver mutations in oncogenes may be functionally significant and describe a range of oncogenic potentials, from strong to weaker drivers, which may be linked with corresponding dynamic variations. In particular, a group of proposed PIK3CA drivers (S379T, N380S, and E418K) target sites of the increased conformational flexibility and are structurally adjacent to known drivers N345K and C420R which accelerate functional transitions from an inactive cytosolic conformation to an activated form (Burke, Perisic, Masson, Vadas, & Williams, 2012). Some other proposed driver mutations (P539R, D549N, and G451R/V) have similar dynamic signatures and are implicated in affecting constitutive phosphorylation of downstream proteins (Gymnopoulos, Elsliger, & Vogt, 2007; Mankoo, Sukumar, & Karchin, 2009)

2.9 Structure-Based Residue Interaction Networks and Centrality Analysis Highlight Mediating Allosteric Function of Driver Mutation Sites in Tumor Suppressor Genes

We also constructed and analyzed the organization and global properties of the residue interaction networks in the crystal structures of 3D-hotspot annotated oncogenes and tumor suppressor genes (Gao, et al., 2017). A global centrality measure, residue betweenness, was employed to characterize the distribution of highly connected residues that mediate stable interaction networks and allosteric communications in protein structures (del Sol, Fujihashi, Amoros, & Nussinov, 2006; del Sol, Fujihashi, Amoros, & Nussinov, 2006; Vijayabaskar & Visheshwara, 2010; Stetz & Verkhivker, 2015; Stetz & Verkhivker, 2016; Stetz & Verkhivker, 2017). In the network model, the peaks in the residue centrality profiles often corresponded to major mediating sites localized in structurally stable regions. By computing residue centrality profiles for the crystal structures of 3D-hotspot genes, we generated the network centrality distributions for oncogene and tumor suppressor structures. As expected, the overall centrality distributions in both classes of proteins are similar, where the characteristic long tail signals presence of a small number of high centrality residues associated with global mediating role in interaction networks (del Sol, Fujihashi, Amoros, & Nussinov,

2006; del Sol, Fujihashi, Amoros, & Nussinov, 2006; Vijayabaskar & Visheshwara, 2010; Stetz & Verkhivker, 2015; Stetz & Verkhivker, 2016; Stetz & Verkhivker, 2017). Strikingly, the distribution of mutational hotspots and predicted driver positions in oncogenes and tumor suppressor genes was markedly different (Figure 9). We found that validated hotspots of activating mutations in oncogenes are enriched in dynamic sites with a relatively moderate centrality (Figure 11A) and are often located at the intersection of high and low stability regions prone to conformational changes. In contrast, a pronounced shift towards the higher centrality was seen for known and predicted driver positions in tumor suppressor genes, indicating that these residues can mediate allosteric interactions in the protein structure (Figure 11B). Strikingly, our predictions can recapitulate this trend, as the distributions of predicted driver mutations in tumor suppressor genes follow closely the centrality signature of known mutational hotspots. According to our findings, most of the potential driver mutations in tumor suppressor structures target structurally stable high centrality residues that control allosteric interactions and signal transmission. This analysis suggested that a mechanism of inactivation in these genes may proceed through point mutations that reoccur at global mediating sites that server as primary coordinators of allosteric signaling. In a sharp contrast, the centrality distribution of sites targeted by the predicted driver mutations in oncogenes is shifted towards lower centrality values, indicating that gain-of-function activating mutations may be enriched in more dynamic

functional residues that serve as sensors of allosteric signals are often involved in the execution of allosteric conformational changes between the inactive and active states.

The analysis of residue centrality profiles in BRAF and PIK3CA structures (Supporting Information, Figure S3) showed that activating driver mutations (L597V/S/Q/R/L, A598V, T599I/A, V600E/K/R/M/R) target mobile residues of moderate centrality. On the other hand, inactivating mutations G466E, G466V, G596R and D594V corresponded to high centrality positions. According to these findings, inactivating mutants are centrally positioned and serve as mediating centers in the allosteric interaction network of the catalytic domain. As a result, mutations of these residues could severely impair allosteric interactions in the functional dimer and completely abrogate kinase activity. The centrality profiles in the TP53 and PTEN structures showed that majority of hotspots residues and predicted driver mutations targeted high centrality residues in the protein core and in homo-oligomerization sites (Supporting Information, Figure S3). Of particular interest are the centrality peaks in TP53 structure that corresponded to predicted drivers lacking experimental validation and proposed for further testing (Q331H, R337C, R337H, F341S, G334W and

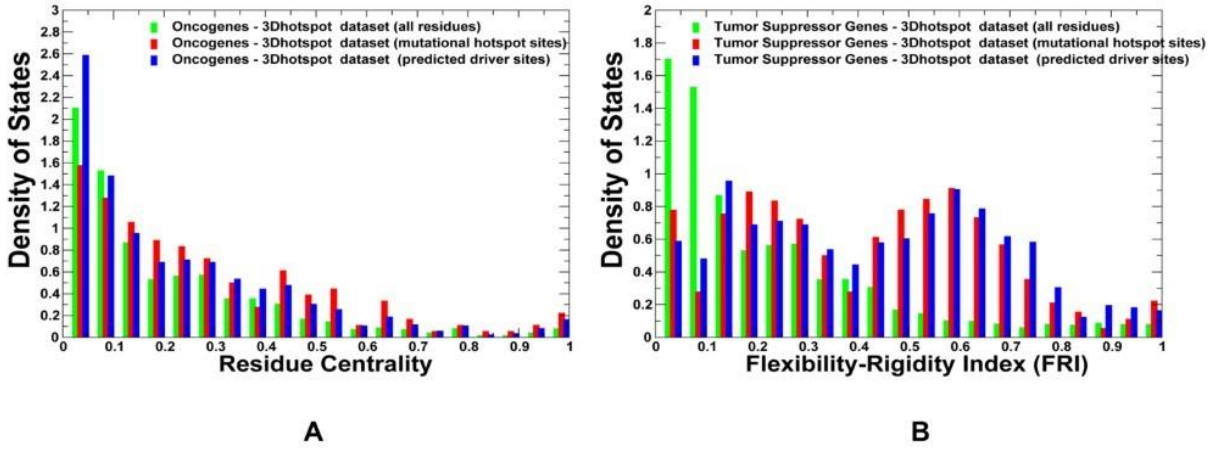


Figure 10. The distributions of residue-based centrality in the crystal structures of oncogenes and tumor suppressor genes.

R342P). The respective residues received relatively high centrality and mediate binding interfaces in the

tetramerization domain of TP53 structure. Hence, the network analysis revealed that high centrality centers of known mutational hotspots and proposed driver sites can act cooperatively to orchestrate an allosteric cross-talk to mediate DNA-binding and protein-protein interactions. We argue that spatial diversity of loss-of-function mutational positions may be linked with their global cooperativity as hubs of long-range signal transmission. Collectively, our results suggested that sites of the predicted driver mutations, which are not directly connected with known hotspots and target more dynamic regions, could be involved in allosteric interaction networks and act cooperatively with known mutational hotspots in exerting a collective functional effect.

2.10 Conclusion

In this study, we integrated machine learning with large scale structural analysis, protein dynamics profiling and modeling of residue interaction networks to determine distinct molecular signatures of cancer driver mutations in oncogenes and tumor suppressor genes. We developed two cancer-specific machine learning classifiers that were validated on canonical datasets and applied for prediction of driver mutations in Cbioportal cancer genomics dataset. By using detailed comparative analysis with various structure-functional experimental data and multi-center mutational calling results from Pan Cancer Atlas studies, we demonstrated robustness of our models. To facilitate interpretability of machine learning results, we compared our cancer driver predictions against the catalog of 3D clusters of mutational hotspots and characterized molecular signatures of functional regions harboring cancer driver mutations in oncogenes and tumor suppressor genes. The predicted driver residues that coincide with validated mutational hotspots tend to occupy structurally stable positions tumor suppressor genes and are often consolidated inside the protein core or localized at several dense binding interfaces. At the same time, putative driver positions in oncogenes tend occupy more dynamic sites in localized functional regions involved in activating transitions. By using several case studies for important cancer genes, we demonstrated that sites of the predicted driver mutations, which are not directly connected with known hotspots and target more dynamic regions, could be involved in allosteric interaction networks and act cooperatively with major driver sites

in exerting a collective functional effect. By carefully inspecting predictions of machine learning models on specific individual examples, we obtain useful insights into mechanisms underlying effects of cancer mutations and identify directions to improve quality, interpretability and reliability of machine learning model approaches.

Chapter 3: Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations

3.1 Introduction

Deep sequencing studies have enabled a detailed characterization of cancer genomes and unveiled important genespecific signatures of somatic mutations (Davies et al., 2002; Bardelli et al., 2003; Futreal et al., 2004; Samuels et al., 2004; Stephens et al., 2004, 2005; Wang et al., 2004; Sjoblom et al., 2006; Greenman et al., 2007; Wood et al., 2007; Vogelstein et al., 2013; Watson et al., 2013). The steadily growing amount of data generated in cancer genomic studies and next-generation sequencing (NGS) have been the impetus behind formation of international cancer genomic projects and development of large bioinformatics data resources such as Cancer Genome Atlas (TCGA), Genomics Data Commons Portal (https://portal.gdc. cancer.gov/) (Weinstein et al., 2013; Jensen et al., 2017), COSMIC database (http://cancer.sanger.ac.uk) (Forbes et al., 2015), and the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010; Zhang et al., 2011; Klonowska et al., 2016; Hinkson et al., 2017). The Cancer Gene Census of the Catalog of Somatic Mutations in Cancer (COSMIC) database has grown from 291 well-characterized cancer genes (Futreal et al., 2004) to more than 500 entries (Forbes et al., 2015) where some cancer genes can be commonly mutated across cancer types, while other genes are predominantly cancer-specific. The cBio Cancer Genomics Portal (https://www.cbioportal.org/) is an open-access resource for exploration of large cancer genomics data sets (Cerami et al., 2012; Gao et al., 2013). These datasets have allowed for comprehensive genome-wide analyses of genetic alterations in multiple tumor types (Poulos

and Wong, 2018). A relatively small fraction of somatic variants known as driver mutations have considerable functional effects and can be acquired over time as a result of a range of mutational processes, rather than inherited (Haber and Settleman, 2007; Lawrence et al., 2013; Vogelstein et al., 2013). A comprehensive analysis of cancer driver genes and mutations has provided classification of 751,876 unique missense mutations, producing a dataset of 3,442 functionally validated driver mutations (Bailey et al., 2018). Another significant dataset of 1,049 experimentally tested and functionally validated driver mutations (Ng et al., 2018) has expanded our knowledge of cancer-causing variants in oncogenes and tumor suppressor genes. TCGA organized the Multi-Center Mutation Calling in Multiple Cancers (MC3) network project which generated a comprehensive and consistent collection of somatic mutation calls for the 10,437 tumor samples dataset (Ellrott et al., 2018). Computational approaches that assess the impact of somatic mutations are often characterized by different basic assumptions, types of input information, models, and prediction targets such as driver gene or driver mutation (Gonzalez-Perez et al., 2013; Cheng et al., 2016). A number of somatic variant callers based on various statistical and machine learning approaches are now available for somatic mutation detection, including MuTect2 (Cibulskis et al., 2013), MuSE (Fan et al., 2016), VarDict (Lai et al., 2016), VarScan2 (Koboldt et al., 2012), Strelka2 (Kim et al., 2018), SomaticSniper (Larson et al., 2012), and SNooPer (Spinella et al., 2016). A deep convolutional neural network (CNN) approach termed DeepVariant can identify genetic variation in NGS data by discerning statistical relationships around putative variant sites (Poplin et al., 2018). To facilitate systematic and standardized somatic variant refinement from cancer sequencing data, random forest (RF) models and deep learning (DL) approach were utilized, showing that these machine learning techniques could achieve high and similar classification performance across all variant refinement classes (Ainscough et al., 2018). A machine learning approach called Cerebro increased the accuracy of calling validated somatic mutations in tumor samples and outperformed several other somatic mutation detection methods (Wood et al., 2018). Many computational methods have been proposed for prediction of cancer driver genes. Some of these approaches use cohort-based analysis to detect driver genes, including ActiveDriver (Reimand and Bader, 2013), MutSigCV (Lawrence et al., 2013), MuSiC (Dees et al., 2012), OncodriveCLUST (Tamborero et

al., 2013), OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), and OncodriveFML (Mularoni et al., 2016). The success of hybrid methods for scoring coding variants has indicated that integration of different tools may enhance predictive accuracy for both coding and non-coding variants (Li et al., 2015). A deep learning-based method (deepDriver) predicts driver genes by CNN trained with mutation-based feature matrix constructed using similarity networks (Luo et al., 2019). Since many methods are often found to predict distinct or partially overlapping subsets of cancer driver genes, a consensus-based strategy was recently proposed, showing considerable promise and outperforming the individual approaches (Bertrand et al., 2018). A unified machine learning-based evaluation framework for analysis of driver gene predictions compared the performance of these methods, showing that the driver genes predicted by individual tools can vary widely (Tokheim C. et al., 2016; Tokheim C. J. et al., 2016). Computational methods designed to identify driver mutations have become increasingly important to facilitate an automated assessment of functional and clinical impacts (Gnad et al., 2013; Ding et al., 2014; Martelotto et al., 2014; Raphael et al., 2014; Cheng et al., 2016). Functional computational prediction methods include Sorted Intolerant From Tolerant (SIFT) (Sim et al., 2012), PolyPhen-2 (Adzhubei et al., 2010), Mutation Assessor (Reva et al., 2011), MutationTaster (Schwarz et al., 2010), CONsensus DELeteriousness score of missense mutations (Condel) (Gonzalez-Perez and Lopez-Bigas, 2011), Protein Variation Effect Analyzer (PROVEAN) (Choi et al., 2012), and Functional Analysis Through Hidden Markov Models (FATHMM) (Shihab et al., 2013). Cancer-specific High throughput Annotation of Somatic Mutations (CHASM) (Carter et al., 2009; Douville et al., 2013; Masica et al., 2017), Cancer Driver Annotation (CanDrA) (Mao et al., 2013), and FATHMM (Shihab et al., 2013). Many new approaches have recently addressed a problem of locating driver mutations within the non-coding genome regions (Piraino and Furney, 2016). The identification of cancer mutation hotspots in protein structures has been a fruitful approach for identifying driver mutations (Dixit et al., 2009; Dixit and Verkhivker, 2011; Gao et al., 2013; Gauthier et al., 2016; Niu et al., 2016; Tokheim C. et al., 2016; Tokheim C. J. et al., 2016). To consolidate functional annotation for SNVs discovered in exome sequencing studies, a database of human non-synonymous SNVs (dbNSFP) was developed (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016). This resource allows for computation of a total of 48

functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches and 15 conservation features (Wu et al., 2016). In our recent investigation, two cancer-specific machine learning classifiers were proposed that utilized 48 functional scores from dbWGFP server in classification of cancer driver mutations (Agajanian et al., 2018). In this work, we explore and integrate RF and DL/CNN machine learning approaches for prediction and classification of cancer driver mutations. We first explore the ability of CNN models to identify and classify cancer driver mutations directly from raw nucleotide sequence information without relying on specific functional scores. The performance of these classifiers was compared to RF and gradient boosted tree (GBT) methods to provide a comparative analysis of various classification models. These raw sequence-derived scores are advantageous because they can be obtained for any mutation with a known chromosome and position, whereas the functional scoring features can be limited to subsets of genomic mutations. By developing a successful classification scheme that could leverage information from raw DNA sequences, the universe of classifiable mutations can be greatly expanded leading to more general and robust machine learning tools. The results of this study reveal that CNN models can learn high importance features from genomic information that are complementary to the ensemble-based predictor scores traditionally employed in machine learning classification of cancer mutations. We show that integration of the DL-derived predictor score with only several ensemble-based features can recapitulate the results obtained with a large number of functional features and improve performance in capturing driver mutations across a spectrum of machine learning classifiers. Machine learning predictions are leveraged in biophysical simulations and network analysis of protein kinase oncogenes to obtain more detailed functional information about molecular signatures of activating driver mutations, aiding in the interpretability of cancer mutation classifiers.

3.2 Mutational Datasets and Feature Selection

In our earlier study (Agajanian et al., 2018) we used RF classifier to predict cancer driver mutations using a combination of two golden datasets (Mao et al., 2013; Martelotto et al., 2014). Here, we expanded this dataset by adding the predicted cancer driver mutations and passengers from the analysis of missense

mutations in Cbioportal database (Agajanian et al., 2018). By leveraging the earlier analysis, we created a dataset consisting of functionally validated 6,389 cancer driver mutations and 12,941 passenger mutations. The driver/passenger classifications for 2,570 of these mutations were present in the two aforementioned golden datasets, and our RF classifier made predictions on the remaining 16,760 missense mutations from the Cbioportal database. Given the performance level of our model (Agajanian et al., 2018), we conjectured that a combination of the two golden datasets and the missense mutations in the Cbioportal database would yield an informative dataset for the current study. The initially selected features for RF predictions were obtained from dbWGFP web server (Wu et al., 2016) of functional predictions for human whole-genome single nucleotide variants. A total of 32 sequence-based, evolutionary and functional features identified in our previous study (Agajanian et al., 2018) were initially used for machine learning experiments with the new dataset of cancer mutations. In cancer driver mutation predictions, traditional input data contain distinct features that cannot be directly applied to CNN models due to their lack of spatial meaning. Using the chromosome and the position on that chromosome that corresponded to the mutated nucleotide, we could retrieve the surrounding nucleotides of the mutation of interest to perform classification with only this raw string of nucleotides. To represent the original nucleotide and its mutated version, we placed two nucleotide sequences on top of each other, one containing the original string, and the other contained the mutated version. This would only result in a one nucleotide difference between the two, allowing to effectively utilizing the sliding window format of the CNN models. The schematic workflow diagram of the CNN approach employed in this study is presented in Figure 12.

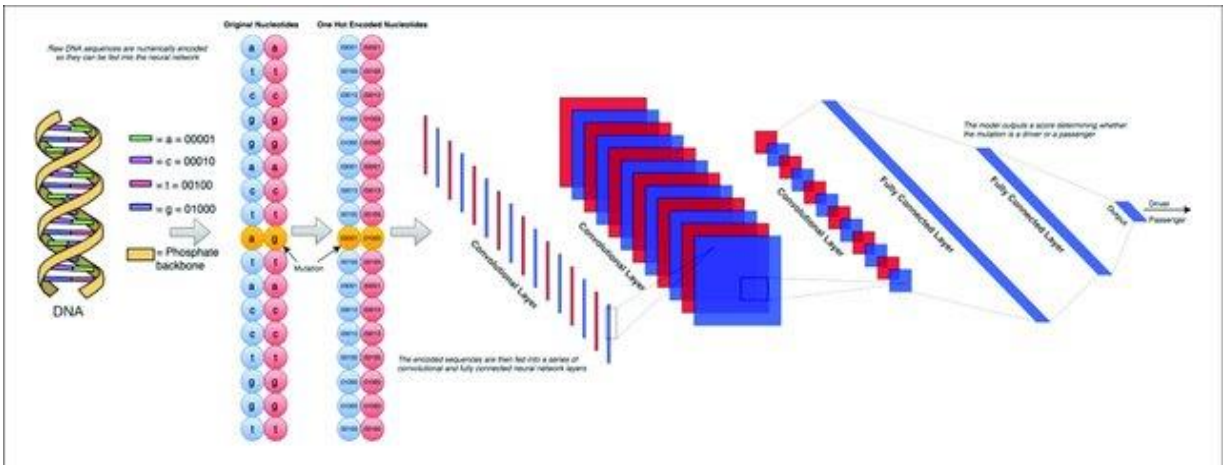


Figure 11. The schematic workflow diagram of the CNN approach employed in this study

To create this dataset, we parsed information from University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) (Tyner et al., 2017) which takes a chromosome (CHR) and a position (POS) on that chromosome as arguments and returns back all nucleotides within the sequence. Using the dataset consisting of 6,389 driver mutations and 12,941 passengers, we created 5 different datasets of various window sizes around each given CHR/POS pair. The explored window sizes (10, 50, 100, 500, and 5,000) produced nucleotide strings of length 21, 101, 201, 1,001, and 10,001, respectively. To represent the type of mutation (A->C, A->G, etc.) we stacked two of the same nucleotide sequences on top of each other, having one contain the original nucleotide at the position passed in initially, and the other containing the mutated version (Figure 13A). This operation resulted in a total input matrix size of (2, 21), (2, 101), (2, 201), (2, 1001), and (2, 10001), respectively. Three different preprocessing techniques were then applied to the dataset to allow it to be passed into the CNN model in the numerical form: label encoding (Figure 13B), one-hot encoding (Figure 13C; Goh et al., 2017), and embedding (Figure 13D). Label encoding involves assigning each nucleotide its own unique ID (A->0, C->1, etc.) This imposes an ordering on the nucleotide sequences that may have implications for the neural network learning (Figure 13B). This technique was implemented using the Scikit-learn LabelEncoder package for the Python programming language. We also tried one-hot encoding the dataset by assigning each nucleotide its own bit encoded string (A -> [0,0,0,0,1], C-> [0,0,0,1,0]) (Figure 13C). This tends to be a favorable preprocessing function

for weight-based classifiers because no artificial ordering is imposed on the samples. This technique tends to be the default representation choice for categorical variables due to how it is interpreted. Because each nucleotide gets its own index in a 5 bit string, a 1 in any particular index means that nucleotide is present in that location. For example, since A->[0,0,0,0,1], this can essentially be read as "There are 0 'n,' 0 'g,' 0 't,' 0 'c,' and 1 'a' nucleotides present at this location." Since the one-hot encoding preprocessing technique lengthens the string, the resulting dimensionalities were (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005), respectively. The final preprocessing technique employed for the DNA sequences involved learned embeddings created with the word2vec algorithm (Mikolov et al., 2013). This technique analyzes the sequential context of the nucleotides assigning them a numeric representation in vector space. Using this representation, the nucleotide segments with similar meaning in the word2vec model would yield similar vectors in an N-dimensional representation. This technique was implemented using the Word2Vec model from the genism library for the Python programming language. Since the vocabulary in this application is fairly small, consisting of only 5 bit components, we chose to convert the nucleotide to 2 dimensional vectors which is sufficient to effectively encode this set. This resulted in the input sizes (2, 42), (2, 202), (2, 402), (2, 2002), and (2, 20002), respectively (Figures 12, 13). The implementation and execution of these three preprocessing techniques provides adequate and efficient nucleotide representations for the CNN classifier.

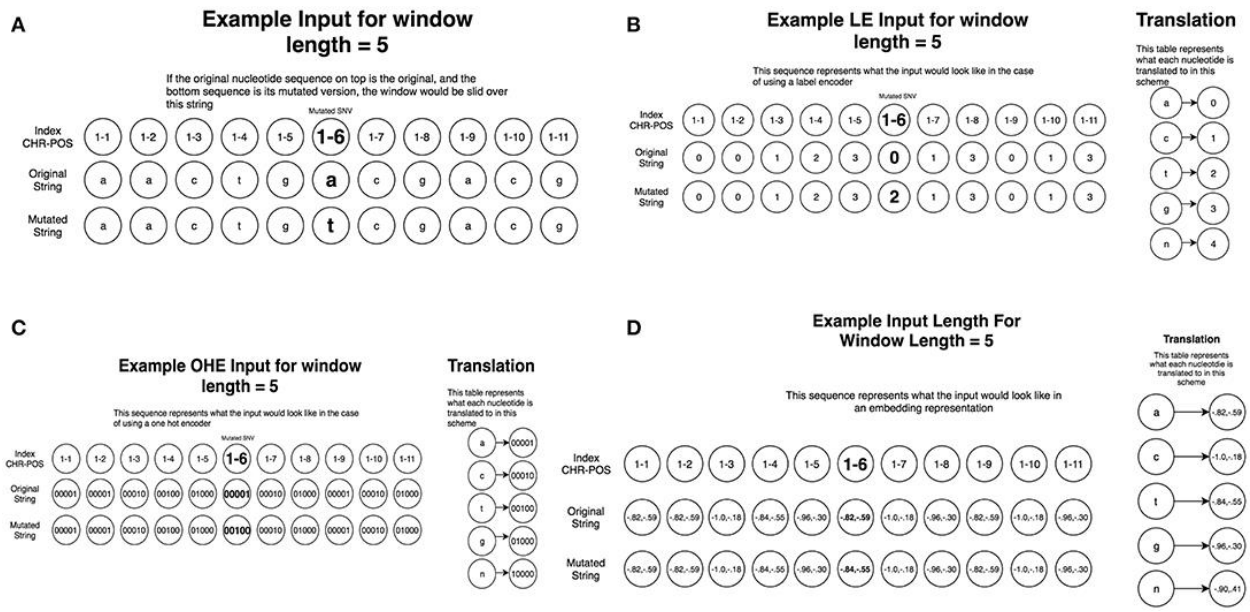


Figure 12. Preprocessing of the nucleotide information for CNN machine learning of cancer driver mutations.

3.3 Machine Learning Models

We used and compared performance of tree based classifiers and DL/CNN machine learning models. For the tree based methods, we used previously established protocol for obtaining hyperparameters (Agajanian et al., 2018). The model training and tuning was done using Scikit-learn free software machine learning library for the Python programming language (Pedregosa et al., 2011; Biau, 2012). The Keras framework was used for training, validation and testing of CNN models (Erickson et al., 2017). We initially held out 20% of the data in a stratified manner as a testing set so that it had the same distribution of passengers/drivers as the total dataset. We then used the remaining 80% of the dataset as the training set to learn and tune its hyper-parameters. To choose between the hyperparameters attempted, we test our model out on unseen data so that we have an unbiased estimate of its performance. To do this, we performed 3-fold cross validation, splitting the training set up into three equal sized portions. The model trains on two of them and makes predictions on the third. This is repeated three times so that each of the three portions has been predicted on. A workflow diagram of the CNN approach (Figure 12) was carefully engineered to determine the optimal architecture. For this, we performed a grid search over a total of 72 different neural network architectures. These 72 architectures consisted of between 1 and 3 convolutional layers and 1–3

fully connected layers following. The number of nodes in each of these layers was also varied between 2 and 256 in powers of 2. The simplest architecture covered in this search contains 1 convolutional layer with 2 filters feeding into 1 fully connected layer with 2 nodes, and the most complex would have 3 convolutional layers feeding into 3 fully connected layers, all containing 256 nodes. The ReLU activation function was used, which returns max (0, X). All 72 different architectures (Table 3) were tested using this cross-validation algorithm and the architecture that had the highest F1 score across all 3-folds was chosen. Our neural networks were trained for 100 epochs, which means that they will pass through the entire dataset 100 times to complete their training. In between each epoch, the model recorded its predictions on the validation fold, and the epoch with the best performance on the validation set was recorded. Dropout was applied in between layers, so that inputs into a layer are randomly set to 0 with a certain probability. This prevents the neural network from overfitting, forcing it to learn without random features present. The best architecture was used for predictions on the test set.

Architecture	*# Layers*	*# Nodes per Layer*
0	2	32,3
1	3	16,8,2
2	3	16,16,2
3	3	32,16,2
4	3	32,8,2
5	3	64,32,2
6	3	64,16,2
7	4	64,64,16,2
8	4	128,64,16,2
9	4	128,64,32,2
10	5	128,64,32,16,2

Table 3. The parameters of displayed CNN architectures in classification of cancer driver mutations.

3.4 Biomolecular Simulations of Cancer Mutation Effects: Rigidity Decomposition and Protein Stability Analysis

We used FIRST (Floppy Inclusion and Rigid Substructure Topography) approach (Jacobs et al., 2001; Rader et al., 2002; Chubynsky and Thorpe, 2007) and the Python-based Constraint Network Analysis (CNA) interface (Hespenheide et al., 2002; Kruger et al., 2013; Pfleger et al., 2013a,b) to analyze partition

of rigid and flexible regions in a set of protein kinases with the predicted cancer driver mutations. The employed parameters are consistent with our previous studies of protein kinases (Stetz et al., 2017). Protein stability computations that evaluated the effect of cancer driver mutations on the functional forms of the ErbB kinases were performed using CUPSAT (Cologne University Protein Stability Analysis Tool) (Parthiban et al., 2006, 2007). This approach was successfully adopted for the energetic analysis of cancer mutation hotspots (Dixit et al., 2009; Dixit and Verkhivker, 2011). We also employed the Foldx method (Guerois et al., 2002; Schymkowitz et al., 2005; Tokuriki et al., 2007; Van Durme et al., 2011) that allows for robust assessment of mutational effects on protein stability. These calculations were done with the user interface for the FoldX force field calculations (Schymkowitz et al., 2005) implemented as a plugin for the YASARA molecular graphics suite (Van Durme et al., 2011).

3.5 Deep Learning Classification of Cancer Driver Mutations from Nucleotide Information

We began with an attempt to recapitulate our predictions by using various DL/CNN architectures informed by raw nucleotide sequence data evaluated the ability to make predictions based solely on raw genomic information. The inclusion of the three different preprocessing techniques allowed us to select the most informative representation of the nucleotides. The one hot encoded sequences yielded the model with the best performance, and for clarity of presentation we report only the dimensions and performance of the one hot encoded model. This preprocessing model resulted in input matrices of size (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005) corresponding to the different window sizes (10, 50, 100, 500, 1,000) surrounding the original nucleotide. It is worth noting that the embedding algorithm also learned meaningful representations of the nucleotides. The missing place indicator, “n,” was predictably separated from the original nucleotides, which were arranged in 2 neat clusters (Figure 13D). Cluster 1 consisted of the adenine and tyrosine nucleotides, and cluster 2 consisted of the guanine and cytosine nucleotides. These two clusters are easily identified due to the fact that their constituent components are very close to each other while simultaneously being far away from the other cluster. We employed 72 different DL architectures (Table

3) and the results for the window size of 10 are presented since they revealed more variance (Figure 14). The figures below display the 10 best performing models out of the 72 attempted. The training accuracy continued to increase for the duration of training (Figure 14A), while on the validation testing set of cancer mutations, the best DL/CNN architecture achieved an average validation accuracy of 86.68% with an F1 score of 0.61 (Figure 14B). Interestingly, we found that the DL model seemed to learn early on, overfitting with each successive epoch (Figure 14B). In fact, the model achieved its highest validation accuracy on the first epoch, and proceeds to decline as learning proceeds in subsequent epochs. Furthermore, the AUC score of the model as well as the F1 score consistently stayed the same throughout all of the process. This is further contextualized by the tree based method's performance on the same dataset. The GBT classifier exhibited an F1 score of 0.57 with an average validation accuracy of 66.59%, and the RF classifier exhibited an F1 score of 0.58 and an average validation accuracy of 69.86%. We analyzed predictions by the DL/CNN model by assigning the predicted values for the entire dataset as a separate new feature termed DL score. Although we probed a variety of different architectures and several nucleotide-encoding protocols, a direct brute-force application of DL/CNN models to predict driver mutations only as a function of surrounding nucleotides appeared to be challenging. As a result, we suggested that a diverse set of more informative features may be required to recapitulate the level of robust performance achieved in our earlier work with sequence-based conservation and functional features (Agajanian et al., 2018).

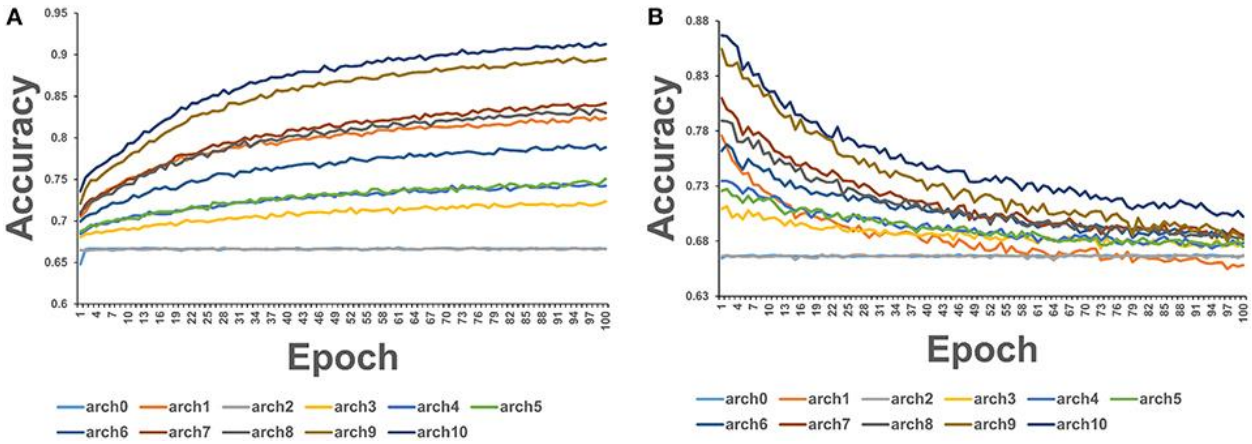


Figure 13. The average accuracy of CNN model using exclusively nucleotide information.

We first used the RF classifier on the cancer mutation dataset with functional and conservation features obtained from dbWGFP server and adopted in our previous study (Agajanian et al., 2018). A database of human non-synonymous SNVs (dbNSFP) was developed as a one-stop resource for analysis of disease-causing mutations (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016) storing 8.58 billion possible human whole-genome SNVs, with capabilities to compute a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches, 15 conservation features from 4 different tools including ensemble-based predictors RadialSVM, LR, and MSRV scores. The initially selected features were obtained from dbWGFP web server of functional predictions for human whole-genome single nucleotide variants that provided 32 functional prediction scores and 15 evolutionary features (Agajanian et al., 2018). Functional prediction scores refer to scores that predict the likelihood of a given SNV to cause a deleterious functional change in the protein, and evolutionary scores refer to scores providing different conservation measures of a given nucleotide site across multiple species. Some of the score features (SIFT, PolyPhen, LRT, Mutation Assessor, MutationTaster, FATHMM, RadialSVM, LR, MSRV, and SinBaD) can be applied only to SNVs in the protein coding regions, while other scores (Gerp++, SiPhy, PhyloP, Grantham, CADD, and GWAVA) can evaluate SNVs spreading over the whole genome. The ensemble-based scores RadialSVM and LR are integrated features that used machine learning approaches to combine information from 10 individual component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, Gerp++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) (Agajanian et al., 2018). In this baseline experiment we evaluated feature performance of 32 input features on the expanded dataset (Figure 15A). Similar to our previous investigation (Agajanian et al., 2018), we found that the ensemble-based scores LR and RadialSVM considerably overshadowed the contributions of other features (Figure 15). By adding DL score to the original 32 features, we applied the RF model for predicting cancer driver mutations with this expanded set of features. The first question was to analyze feature importance of the RF model with the DL score included and determine whether the nucleotide-based scoring feature can contribute to the prediction performance in a meaningful and appreciable way (Figure 15). In the second round of RF classification experiments, we added DL score to the original list of

32 features (Figure 15B). Strikingly, the DL score ranked third following the ensemble-based LR and RadialSVM scores (Figure 15B). Moreover, it was evident that these three feature scores completely dominated feature importance distribution, with the DL score contributing almost as much as the ensemble-based RadialSVM feature (Figure 15B). Quite remarkably, the DL-based score derived by CNN exclusively from primary nucleotide information can deliver significant information content and enrich predictions.

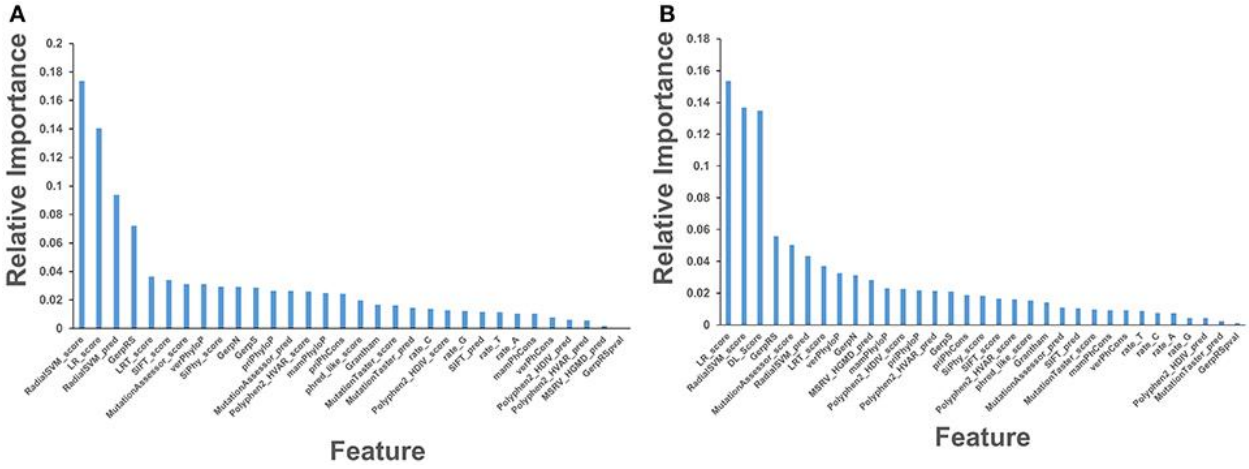


Figure 14. Feature importance of the RF machine learning model on the cancer mutation dataset.

Using Spearman's rank correlation coefficient, we computed the pairwise correlations between different prediction scores (Figure 16). In this analysis, we found that the two dominant feature scores RadialSVM and LR are only moderately correlated with DL score, with the correlation coefficient of 0.486 and 0.423, respectively. Interestingly, RadialSVM and LR scores are more significantly correlated, suggesting that these ensemble-based features could be complementary with the nucleotide-based DL score. Accordingly, we argued that a combination of these dominant and yet complementary scores may allow for feature reduction and more robust performance of the RF classification models.

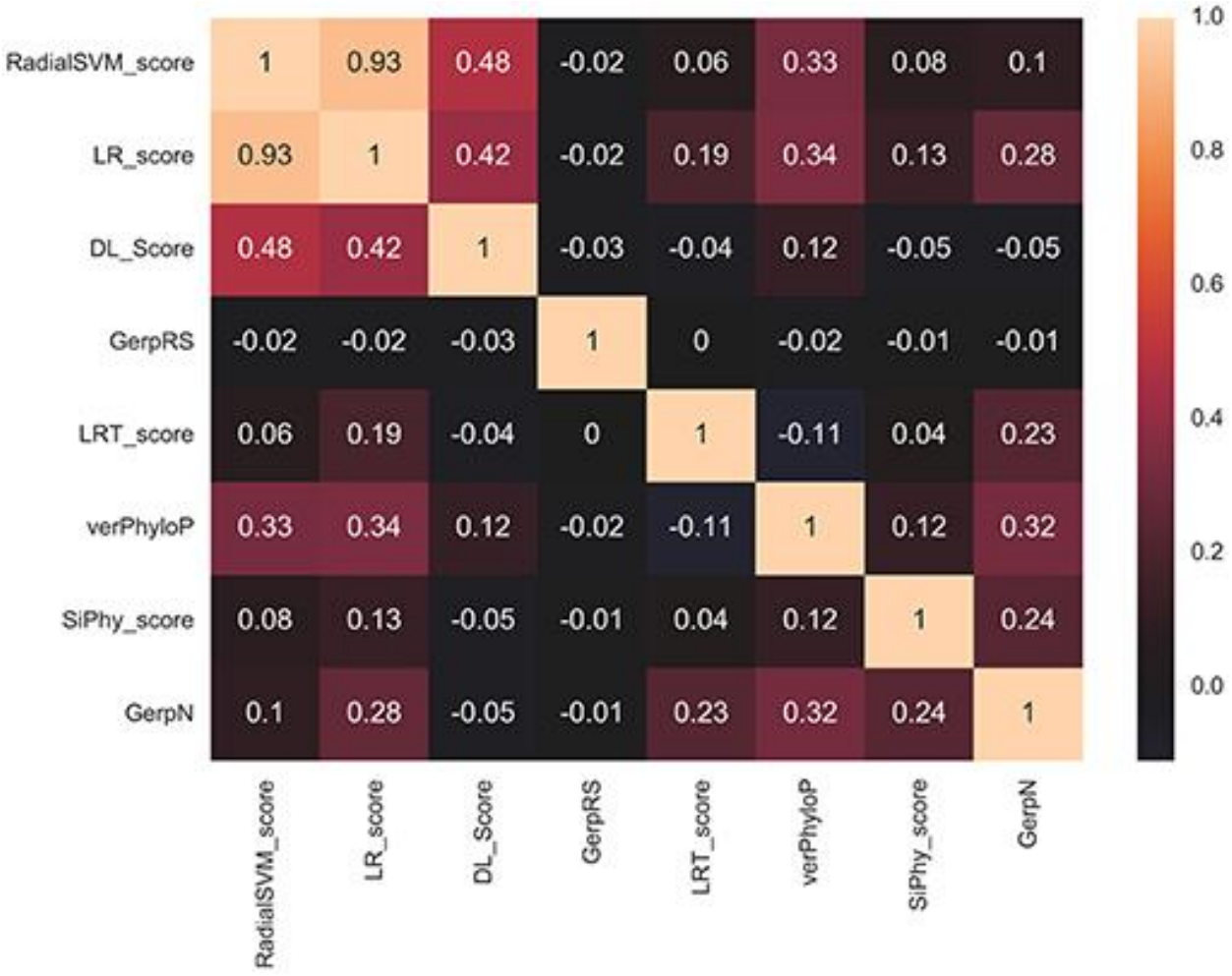


Figure 15. The pairwise Spearman's rank correlation heat map between different prediction scores.

3.6 Integration of CNN Predictions with Ensemble-Based Features in Classification Models of Cancer Driver Mutations

Based on these findings, we evaluated feature selection again aiming to recreate the same accuracy with only 8 features: RadialSVM score, LR score, DL score, GerpRS, LRT score, verPhyloP, SiPhy score, GerpN (Figure 17A). The RF model with only 8 features produced a similar ranking in which the ensemble-based scores and DL score contributed the most (Figure 17A). Other contributing features included evolutionary conservation scores derived from multiple sequence alignments and reflecting functional specificity, such as GerpRS (Davydov et al., 2010), SiPhy (Garber et al., 2009), and PhyloP (Garber et al., 2009) also showed appreciable information score values (Figure 17A). We then tested the performance of

the RF model and feature importance by performing machine learning of cancer driver mutations using only 3 top features (Figure 17B).

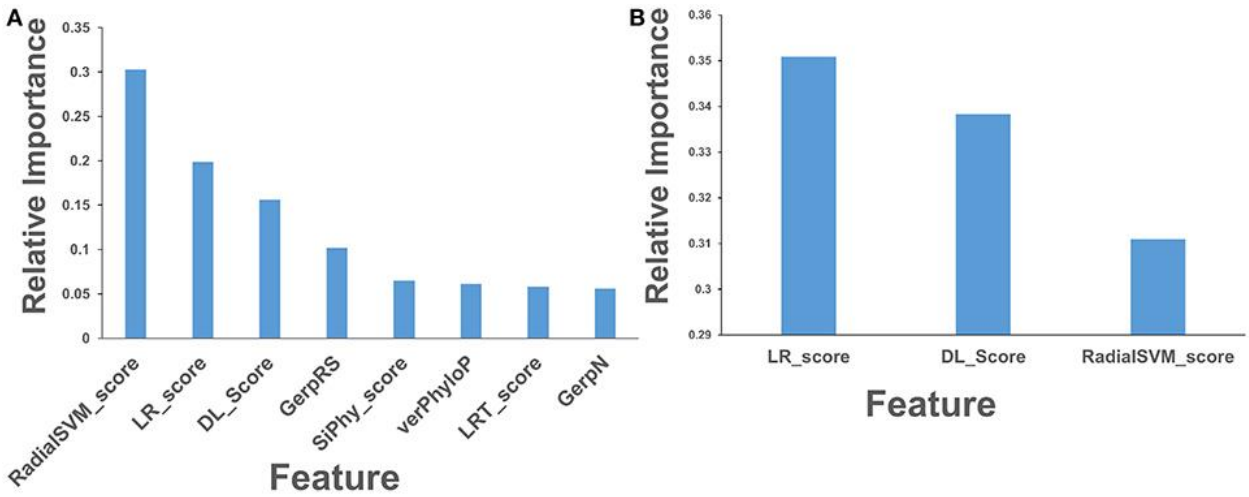


Figure 16. Feature importance of the RF model on the cancer mutation dataset with the reduced number of features.

The predictive performance of the RF models with different set of features was examined using area under the curve (AUC) plots (Figure 18). First, we examined difference in the AUC curves for RF-based classification with 32 functional features and with additional DL score (Figure 18A). The results showed a very similar high-level prediction performance with AUC = 0.95–0.96. It is worth noting that due to high AUC value for RF classification with 32 informative functional features, the addition of DL could not significantly enhance it. However, we showed that this nucleotide-derived predictor score provides an additional information content and is complementary to the ensemble-based RadialSVM score and LR score. In this context, it was instructive to observe that addition of DL score may marginally improve separation between TPR and FPR at higher values of these parameters (Figure 18A).

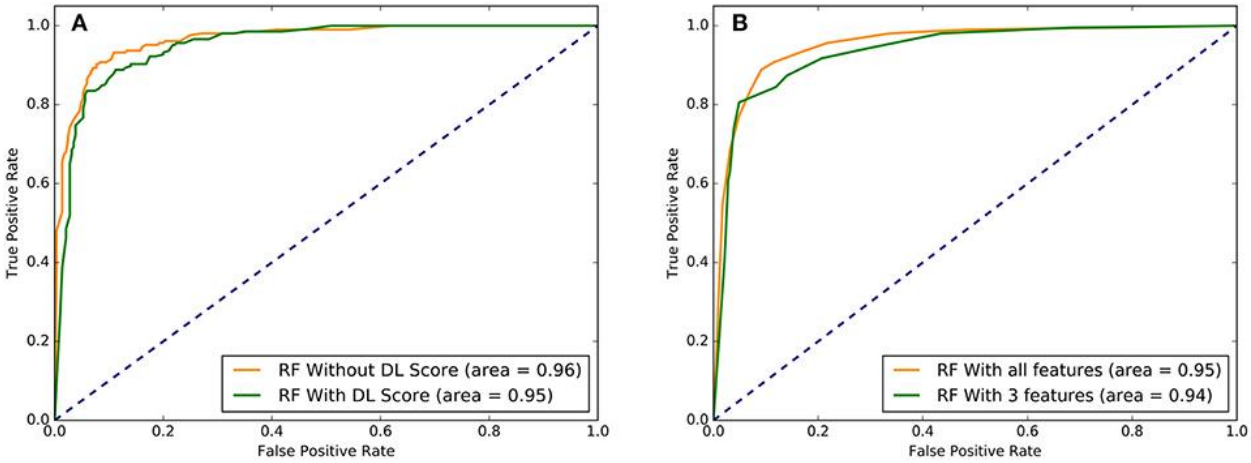


Figure 17. The ROC plots of sensitivity (TPR) as a function of specificity

Strikingly, RF learning model that relied on only 3 top features (RadialSVM score, LR score, and DL score) yielded AUC = 0.94, thereby showing that these features may be sufficient to achieve robust classification of cancer driver mutations on a fairly large dataset of somatic mutations employed in this study. Combined with the findings that DL score only weakly correlated with the ensemble-based scores, we concluded that unexpectedly few highly informative parameters can achieve high level of performance (Figure 18). We then tested several machine learning models including RF, GBTs and support vector machine (SVM) on the dataset with the top 8 features to benchmark performance against the original RF model with 32 features (Agajanian et al., 2018). The performance of classification models was carefully assessed (Table 4). All methods achieved a high classification accuracy of ~90%. The sensitivity values were higher for the SVM and RF models, but all methods yielded similar high-performance classification on the dataset with only limited number of major features that included DL score (Table 4).

	Boosted Trees	*SVM*	*Random Forest*
Accuracy	0.896	0.890	0.896
F1 Score	0.900	0.890	0.900
Precision	0.900	0.890	0.900
Recall	0.900	0.890	0.900
True positive rate	0.850	0.848	0.857
False positive rate	0.112	0.797	0.123
True negative rate	0.115	0.016	0.107
False negative rate	0.913	0.748	0.907

Table 4. The relative performance metrics and statistics of various machine learning models in classification of cancer driver mutations with the top 8 features.

To summarize, our results supported the notion that machine learning-derived ensemble functional predictors may play a central role in classification of cancer driver mutations. The central finding of these machine learning experiments was that combination of ensemble-based features and DL score derived by CNN model from nucleotide information are complementary and when combined can yield classification accuracy comparable and often exceeding the one obtained with a full set of features. The important lesson from this analysis is that integrated high-level features derived by machine learning approaches from primary nucleotide and protein sequence information may be sufficient to predict an important functional phenotype. Although structure-derived features and other functional scores contribute to feature importance ranking and tightly linked with the mutational phenotype, the success of machine learning tools in deciphering predictive features from primary sequence information is encouraging and should be further explored in other applications.

3.7 Leveraging Machine Learning Predictions in Structure-Functional Analysis of Molecular Signatures of Driver Mutations in Oncogenic Protein Kinases

Machine learning driver/passenger classifications typically consider activating, inactivating and inhibitory (or resistant) mutations as drivers, often leaving aside a more detailed characterization and assignment of driver positions. Direct predictions of these specific classes may not be adequately suited for machine learning tools due to smaller datasets. To expand our predictions and aim at extracting a more granular functional information about driver mutations, we conducted rigidity decomposition simulations and analyzed conformational flexibility of the predicted driver positions in protein kinase genes. The objective of this analysis was to facilitate functional validation and interpretation of machine learning results through coarse-grained biophysical simulations as an effective post-processing tool of machine learning classification. In fact, the proposed simulation analysis of mobility at the driver positions allows to expand classification of driver mutations further and characterize activating drivers. Previous studies have suggested that conformational mobility of many oncogenic kinases may be linked with preferential localization of activating cancer mutations in flexible functional regions (Paladino et al., 2015; Kiel et al.,

2016; Stetz et al., 2017). We examined flexibility of specific functional regions targeted by driver mutations in oncogenic protein kinases and probed functional propensity of these drivers to promote transitions to constitutively active states. The primary focus of this analysis is on the family of the ErbB protein tyrosine kinases (Lemmon and Schlessinger, 2010; Roskoski, 2014). A number of human cancers are associated with mutations causing the increased expression of the ErbB kinases. A large number of activating and drug resistance EGFR mutations have been extensively studied at the molecular and functional levels (Paez et al., 2004; Kobayashi et al., 2005; Zhou et al., 2009; Eck and Yun, 2010). Oncogenic kinase mutants are known to act by destabilizing the inactive dormant kinase form while promoting conformational transitions and stabilization of a constitutively active kinase state—a salient functional characteristic linked with the initiation or progression of cancer (Carey et al., 2006; Wang et al., 2011). We used the crystal structures of the EGFR, ErbB2, ErbB3, and ErbB4 kinases that constitute this family to perform rigidity decomposition and then align the positions of the predicted cancer driver mutations with the structural mobility maps (Figure 19). We examined how the predicted driver mutations for ErbB protein kinases are distributed on the rigidity/flexibility map of the catalytic core and whether the dynamic preferences of mutational sites can be linked with their primary function as activating drivers. To explore these questions, we examined the predicted cancer driver mutations for the ErbB kinase family. Structural mapping of these cancer mutations onto the crystallographic ErbB conformations showed that activating driver mutations are preferentially localized in the flexible regions and target positions where they can readily promote conformational changes to the active form without severely compromising thermodynamic stability (Figure 19).

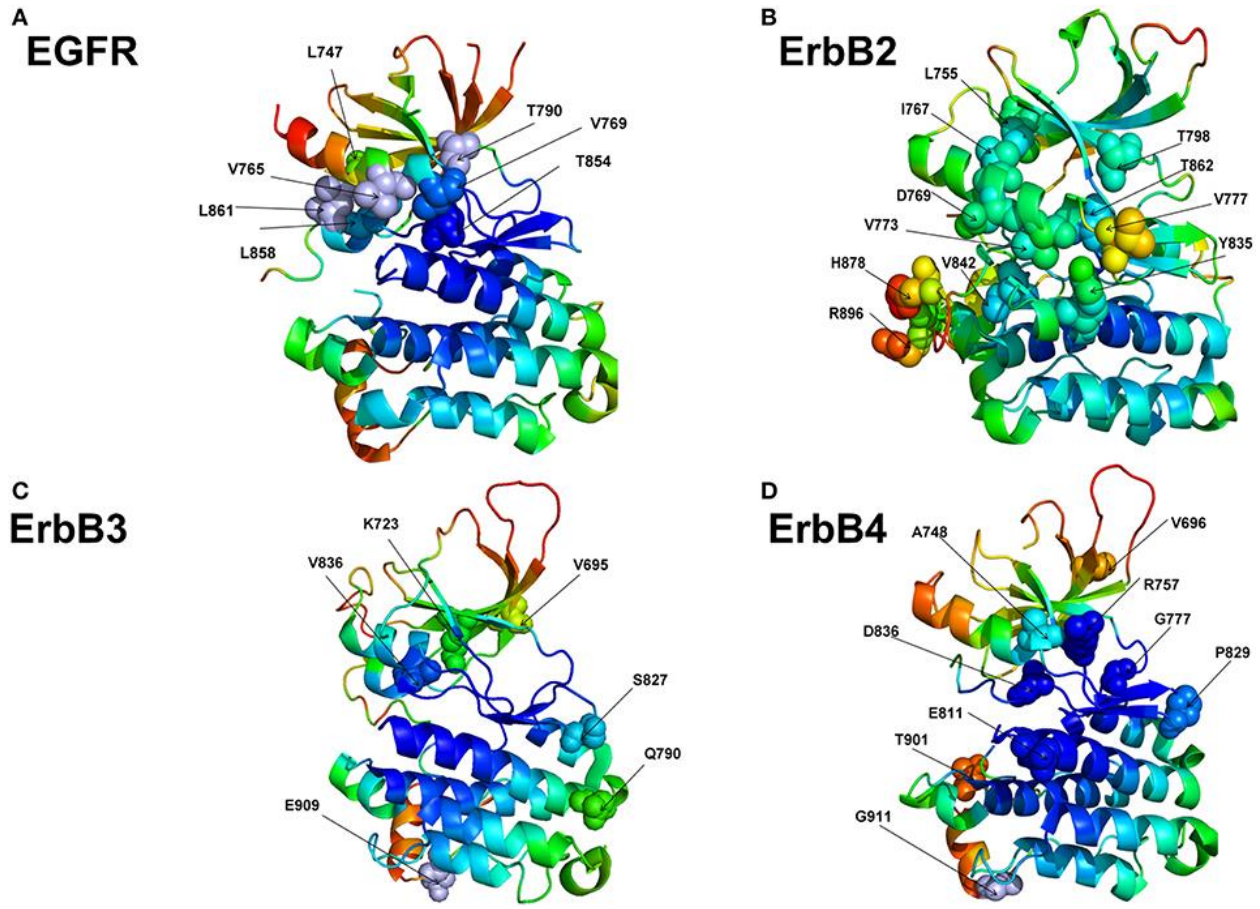


Figure 18. Structural maps of rigidity decomposition and mobility signatures of cancer mutation drivers in the ErbB protein kinases.

To quantify these arguments further, we also characterized the free energy differences between wild-type and cancerdriver mutations for the ErbB proteins in both inactive and active kinase forms (Figure 20). Since both CUPSAT and FoldX approaches yielded similar results, we illustrated our findings by presenting FoldX-derived protein stability changes (Figure 20). The results of this simulation-driven functional classification of predicted driver mutations were compared with the biochemical and mutagenesis data. The analysis of driver mutations in EGFR confirmed that L858 and L861 positions target flexible regions as can be manifested by classical activating driver mutations L858R and L861Q (Littlefield and Jura, 2013; Red Brewer et al., 2013). The energetics of these activating drivers is consistent with a common mechanism of the constitutive activation of kinases by driver mutations (Figure 20A). This mechanism reflects a combined effect of activating mutations producing a more significant destabilization of the inactive state as compared

to the active state, triggering shift of the thermodynamic equilibrium toward the active conformation. We found that some EGFR mutations such as T854A are mapped onto more stable regions of the kinase (Figure 19A) and showed similar destabilization in the inactive and active forms. Accordingly, this predicted cancer driver mutation is likely not activating but rather may be attributed to inhibitory or resistant mutations. Indeed, the recent experimental studies showed that T854A mutation is the acquired mutation causing resistance to known drugs (Bean et al., 2008). Another EGFR mutation V769M/L showed an intermediate level of mobility (Figure 19A) and greater stabilization of the active state. These results are in line with recent functional experiments showing that EGFRV769M mutation is indeed activating that may explain the role of this driver mutation in the development of multiple lung cancers in a pool of lung cancer patients (Deng et al., 2018).

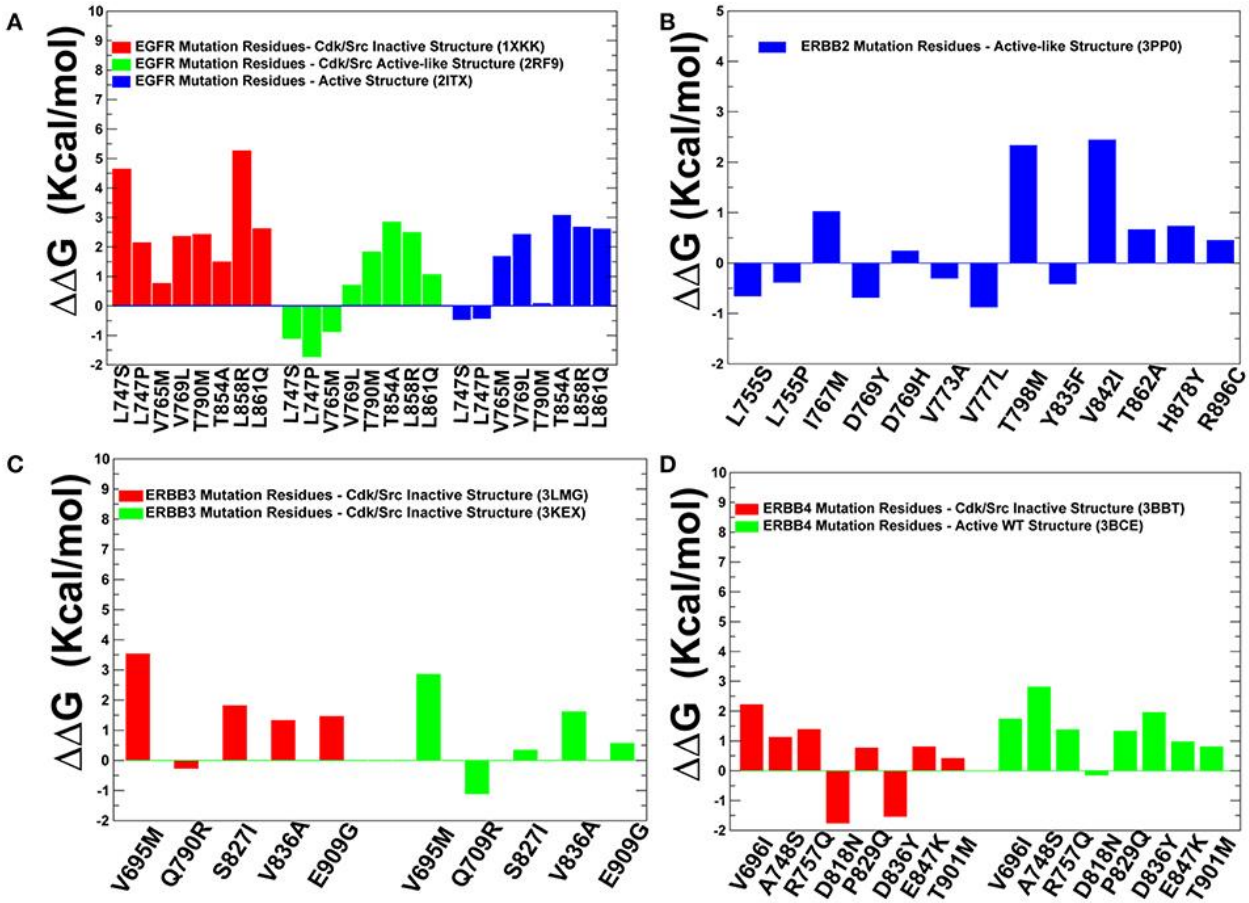


Figure 19. Protein stability analysis of the predicted cancer driver mutations. Protein stability differences calculated between the wild-type and mutants for predicted cancer driver mutations in the ErbB kinases using FOLDx approach.

The positions of almost all predicted driver mutations in ErbB2 kinase target highly flexible regions and can be assigned in our model to activating driver mutations (Figures 19B, 20B). Our previous biophysical simulations and network analysis of activation mechanisms in the ErbB proteins similarly indicated that almost all oncogenic ErbB2 variants are localized in the mobile αC-β4 loop and highly dynamic in their inactive states promoting transition to the active form and causing an uncontrollable activity (James and Verkhivker, 2014). These findings are consistent with the experimental studies (Fan et al., 2008; Aertgeerts et al., 2011). While the majority of somatic mutations in the EGFR and ErbB2 kinases increase the kinase activity, a number of the classified ErbB4 cancer mutants have been shown to inhibit or reduce the kinase activity (Tvorogov et al., 2009). In particular, some cancer-associated mutations of ErbB4 can promote loss of ErbB4 kinase activity as these alterations weaken the important functional interactions in the catalytic core and may interfere with the protein stability. According to experimental data, some cancer mutations have only minor or no effect on kinase activity (V696I, E785K, A748S, P757Q, P829Q, and T901M), while K726R abolishes kinase activity and D818N and D836Q are known as kinase-dead mutations (Tvorogov et al., 2009). We found that predicted cancer driver mutations are mapped onto more stable regions in ErbB4, owing to the greater rigidity of this catalytic domain (Figures 19D, 20D). Accordingly, the respective driver mutations cannot function as activating but rather may cause significant distortions of the kinase structure, causing abolishment of kinase activity which is the functional signature of most cancer drivers in ErbB4 kinase. The performed simulation-driven post-processing of machine learning predictions facilitated in silico functional characterization of cancer mutations and allowed to properly assign activating or inhibiting phenotypic effects to a pool of pathogenic kinase variants. To provide more quantitative insights, we used the predicted cancer mutations in the ErbB kinases and conducted protein structure network analysis to identify whether positions of deleterious mutations would overlap with the global mediating nodes in the interaction networks. The betweenness of a residue node is defined as the number of shortest paths that can go through that node, thus estimating the contribution of the node to the global communication flow in the system. High betweenness nodes can influence the spread of information through the network by facilitating, hindering, or altering the communication between others. According to

our hypothesis, cancer mutations may preferentially target the essential mediating residues with a high centrality that play an important role in activity and signaling of protein kinase genes. The centrality analysis revealed important differences in the distribution of mediating centers in the ErbB kinase structures (Figure 21). We particularly observed that the betweenness of the active form of EGFR (Figure 21A) and ErbB4 (Figure 21D) was on average higher than for the inactive states. Importantly, the location of the properly classified EGFR mutations with the highest oncogenic potential (L858R, T790M, L838V, V742A, V851A, I853T) corresponds to some of the high centrality peaks of the profile (Figure 21A). In addition, these residues showed appreciable differences in the betweenness values between the inactive to the active states, as the residue centrality in these positions typically increased in the functional active form (Figures 21A,D). These findings suggested that a number of key activating mutations in the ErbB kinases target mediating sites of global allosteric communication in the protein structures. We believe that by adding this significant additional component to our study, we have been able to further quantify and explain the protein rigidity/flexibility analysis of predicted cancer mutations in the kinase genes. In our view, by complementing machine learning predictions with the structural and network-based analyses we can obtain useful insights into mechanisms underlying effects of cancer mutations and also identify limitations of classification models and ways to improve interpretability and reliability of machine learning model approaches.

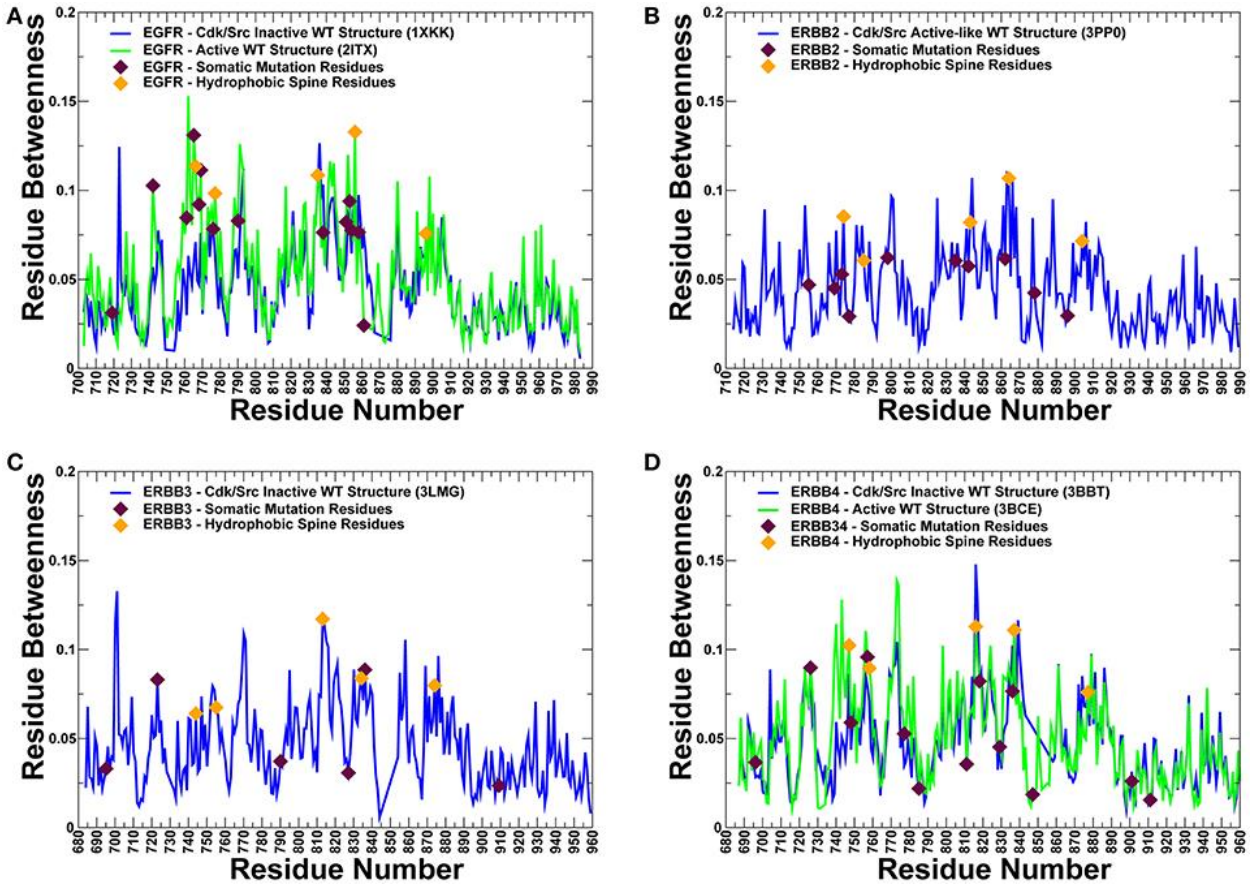


Figure 20. The residue-based betweenness profiles of the ErbB kinase structures.

3.8 Discussion

As large-scale biological data are available from high-throughput assays, and methods for learning the thousands of network parameters have matured, we can now assess feasibility and practicality of using specialized neural network architectures as classification tools for recognizing cancer-causing variants and associated cancer types. Given rapid proliferation and increasing popularity of deep learning tools to address various biological problems, there are several fundamental questions arising in the context of classification of cancer driver mutations. Will deep learning make all other models obsolete? Can deep learning models achieve robust classification and recognition of cancer driver mutations based solely on nucleotide information? What is the role of many functional and structural predictors derived from biophysical perspective in this context? In this work, we have explored and integrated different machine learning approaches for prediction and classification of cancer driver mutations. We first explored the

ability of CNN models to identify and classify cancer driver mutations directly from raw nucleotide sequence information without relying on specific functional scores. The results of this study have demonstrated that while CNN models can learn high level features from genomic information that has sufficiently high importance, accurate classification of cancer mutation driver phenotype using exclusively nucleotide data continues to be challenging. This problem is admittedly more complex than the experimental design suggests, due to the complex nature of protein interactions in the human body. This experimental setup considered only the primary sequence form of the nucleotides, which could only ever partially explain the onset of cancer. The secondary, tertiary, and quaternary form of these same strings would certainly contain more information, due to the folding processes that occur in these steps. Additionally, this technique ignores all of the possible interactions that can be had with other structures in the body, which further dilutes the informational value present in the dataset. As such it's unreasonable to assume that our solely primary sequence based dataset would be able to explain all of the variance present in a complex problem like determining a single mutation's level of effect on the onset of cancer. The experimental inclusion of the different window sizes was also an attempt to allow increasing numbers of surrounding nucleotides to have an influence on our chosen mutation's effect. An obvious assumption here is that more nucleotides would in fact bring in more information. This, however, proved not to hold up as the only dataset that provided any significant variance in performance was the window size = 10 dataset. This suggests that more nucleotides only confuse the model and disallow it from learning informative patterns. This problem could possibly be combatted in future research by testing out larger architectures. The benefits of integrating CNN-derived predictors obtained from nucleotide information with protein sequence features, evolutionary and functional scores were then carefully examined. By exploring various encoding techniques and an array of different CNN architectures, we have found that neural networks can quickly learn an important functional signal, but can rarely steadily improve the initial performance spike with the number of additional epochs. The juxtaposition of monotonically increasing training accuracy with monotonically decreasing validation accuracy is a telltale sign of overfitting. This suggests that there is only a small amount of useful information that can be learned very early on, and subsequent epochs only

cause the model to learn noisy patterns that are only exhibited in the training set. It is difficult to determine exactly what was learned by the model due to the black box nature of neural networks, however due to the short path to optimality it is safe to say that any learned concepts cannot be overly complex. We have pursued a synergistic strategy in which the prediction score generated by CNN models was integrated with physics-based functional, structural and evolutionary conservation features. The important lesson of this analysis was the revelation that CNN-derived features may be complementary to the ensemble-based predictors often employed for classification of cancer mutations. These other scores are not calculated from raw sequence based techniques, which supports this DL score as a novel inclusion into a portfolio of scores due to its unique derivation. By combining deep learning-generated score with only two main ensemble-based functional features, we were able to achieve a high performance level for cancer driver mutations. The robustness of this approach was verified by several traditional machine learning classifiers, including RF, SVM, and GBTs. We have found that integration of CNN-derived predictor score with only several ensemble-based features can recapitulate the results obtained with a large number of functional features and improve performance in capturing driver mutations across a spectrum of machine learning classifiers. Our findings have also demonstrated that synergy of nucleotide-based deep learning scores and integrated metrics derived from protein sequence conservation scores can allow for robust classification of cancer driver mutations with a reduced number of highly informative features. This is an interesting and highly informative result, as the law of parsimony holds for machine learning models so simpler models with comparable performance are typically preferred over their more complex counterparts. Part of this model complexity includes the number of features that a model relies on. As such a reduction in features is a universally positive outcome. In addition to the improved quality of the model, it also expands the universe of predictable nucleotides that are available to us since we depend only on the presence of two ensemble-based scores. The DL score can be derived for any mutation with known coordinates, so this is not a limiting factor. In this respect our initial goal of expanding the nucleotides we can make predictions for was partially achieved. This increase in the generalization of these models facilitates the logical conclusion of driver classification efforts, accurately classifying all known nucleotides. While machine learning approaches can

often produce robust and accurate predictors, the ultimate goal of research is fundamental understanding of the underlying phenomena which requires a mechanistic model of the world. In this context, machine learning predictions are leveraged in biomolecular simulations to enable analysis of cancer mutation mechanisms and obtain a more specific information about an important subset of cancer mutations, activating drivers. The results of our investigation suggested that through integration of machine learning classification and biomolecular simulations of cancer mutations we can often validate the predictions and facilitate a more detailed functional analysis of activating driver mutations. These findings can provide insight and new angle to the problem of interpretability of "black box" machine learning results. By carefully inspecting predictions of machine learning models in the context of dynamic and energetic signatures of mutational sites for oncogenic protein kinases, this study offered instructive strategy for simulation-based post-processing of machine learning predictions and detailed functional specification of cancer driver mutations. The proposed synergistic integration of machine learning and biomolecular simulations into a single computational platform allows to rapidly process large datasets and make robust predictions on functionally significant cancer drivers. The results of this study may also inform and guide design of targeted and personalized therapeutic agents combating a spectrum of mutational changes occurring in cancer.

Chapter 4: Autonomous Molecular Design of Protein Inhibitors

4.1 Review of Molecular Design Techniques and Tools

Drug discovery applications seek to design small molecules that have specific targets. This typically requires extensive trial and error when creating and assessing molecules. Given the levels of success achieved in previous machine learning aided molecular design applications (Maziarka, et al., 2020; Cao & Kipf, 2018; Kadurin, et al., 2017; Yu, Zhang, Wang, & Yu, 2017), we decided to see if the same idea could be applied to generate and alter molecules to display targeted properties. Our goal is to create a framework to make protein kinase inhibiting small molecules as a result of direct generation, or alteration of known

molecules. Protein kinase inhibitors are small molecules that block the actions of enzymes called protein kinases. Protein kinases are involved in many cellular functions including metabolism, cell cycle regulation, survival, and differentiation (Kannaiyan & Mahadevan, 2018). Dysregulation of these protein kinases has been implicated in various carcinogenic processes (Kannaiyan & Mahadevan, 2018). This has led to excitement from the biochemical research community towards the creation of protein kinase inhibitors that can be used as anticancer therapeutic agents.

Generative deep learning models have exceled as tools to aid in navigating the large space of known molecules and in the creation of new molecules. These models are fed various representations of molecules as inputs and learn to perform a variety of things, such as the optimization of these molecules towards a targeted property. This task requires a large amount of data to perform successfully, which in turn requires non-trivial computational resources. An additional functionality that generative models provide is making alterations to inputs to yield transformed outputs. As discussed in chapter 1, this idea has been applied in the chemistry domain with the Mol-CycleGAN (Maziarka, et al., 2020) and MolGAN (Cao & Kipf, 2018). SeqGAN (Yu, Zhang, Wang, & Yu, 2017), created molecules one token at a time, which machine learning models had trouble doing before (Yu, Zhang, Wang, & Yu, 2017). JT-VAE (Jin, Barzilay, & Jaakkola, 2019) and Chemical VAE (Gomez-Bombarelli, et al., 2017) are two successful VAE approaches that have high potential for assisting a molecular generation task. These variational autoencoders are trained using the Simplified Molecular Input Line Entry System (SMILES) format and outperform traditional string encoding techniques while providing a continuous representation. Notably, druGAN combined GAN and VAE by training an adversarial autoencoder to efficiently sample molecules from the latent space (Kadurin, Nikolenko, Khrabrov, Aliper, & Zhavoronkov, 2017).

MolGAN was designed to create new molecules that optimize a portfolio of different properties that include drug likeliness (qed) (Bickerton, Paolini, Besnard, Muresan, & Hopkins, 2012), synthesizability (SAS) (Ertl & Schuffenhauer, 2009), and water-octanol partition coefficient (logP) (Wildman & Crippen, 1999). This requires the model to learn a probabilistic pattern about molecular structure. The authors of this paper

achieved state of the art performance in the optimization of all these selected properties, while maintaining almost 100% validity of generated molecules (Cao & Kipf, 2018). However, the MolGAN framework struggled with creating unique molecules, suffering from mode collapse due to the fact that it could only sample nine atoms to create molecules (Cao & Kipf, 2018). Regardless of the mode collapse issues, MolGAN exhibited impressive performance learning a complicated task and proving that machine learning models are capable of executing molecular design.

Maziarka et al., proved that machine learning models trained in a reinforcement learning paradigm could successfully make structural alterations to small molecules and maintain the validity of these altered molecules with Mol-CycleGAN (Maziarka, et al., 2020). Furthermore, they successfully explored the generative adversarial network's ability to optimize selected properties of these molecules while satisfying a set of constraints designed to aid the validity of the molecules (Maziarka, et al., 2020). This includes forcing these altered molecules to maintain a level of similarity with the original set, effectively setting the "size" of the alteration made to the molecule. This constraint assisted the model in creating valid molecules while still being able to perform the optimization tasks (Maziarka, et al., 2020). The authors obtained the tanimoto similarity of the Morgan fingerprints for these molecules and found that the higher the imposed similarity constraint, the lower the average improvement of the optimized property (Maziarka, et al., 2020). Most importantly, the authors proved that machine learning models could gain enough of an understanding of the complex molecular structure to make targeted improvements to molecules.

Capitalizing on the results achieved by the previous two models, we believe the same idea could be applied to generate and alter molecules to display targeted properties. Development of a model that can produce specific molecules with desired characteristics is highly dependent on three factors. The first is a dataset of molecules encoded in an informative representation that is usable by machine learning models. The second are scoring functions that assess a molecule's quality and assist the model in differentiating desired molecules from undesirable ones. The third, a model or framework that can generate molecules. Choosing among the options for these three factors is not trivial, as each has its own set of strengths and weaknesses.

4.2 How Do Computers Read Molecules?

There is a wealth of active research dedicated towards the determination of an optimal representation for molecules with respect to machine learning models. Each comes with its own set of pros and cons, and there isn't a clear best choice. 3D molecules, in the form of voxels (Kuzminykh, et al., 2018), struggle with the invariance of their representations. Different rotations, translations, or permutations of the atomic indexing can yield different 3D grids that all represent the same molecule. 2D representations cannot encode as much information as their 3D counterparts, but don't suffer from as many of the same invariance issues. They typically are represented by an 80x80 grayscale image, but RGB channel style representations have been attempted. 1D string representations, otherwise known as SMILES (Weininger, 1988), are the most widely used due to their simplicity and wide availability. 1D string representations suffer from similar invariance issues as 3D representations, and restrictive set of constraints that cause models to have problems with the generation of valid SMILES strings. SMILES strings can represent seven important characteristics of a molecule:

1. Atoms
    a. Represented by the standard abbreviation of an element
2. Bonds
    a. Represented with these symbols: 0. `- = # $ : / \`
3. Rings
    a. Numbers are used to show breaks in the ring
4. Aromaticity
    a. In Kekulé form with alternating single and double bonds, e.g. `C1=CC=CC=C1`,
    b. Using the aromatic bond symbol `:`, e.g. `C:1:C:C:C:C:C1`, or
    c. Most commonly, by writing the constituent B, C, N, O, P and S atoms in lower-case forms `b`, `c`, `n`, `o`, `p` and `s`, respectively.
5. Branching
    a. parenthesis

6. Stereochemistry
    a. / and \
7. Isotopes
    a. Bracketed number representing integer isotopic mass

Due to the wide availability and consistent use of SMILES representation in numerous molecular modeling approaches, they are a natural choice to explore. Given that some of the techniques we are using

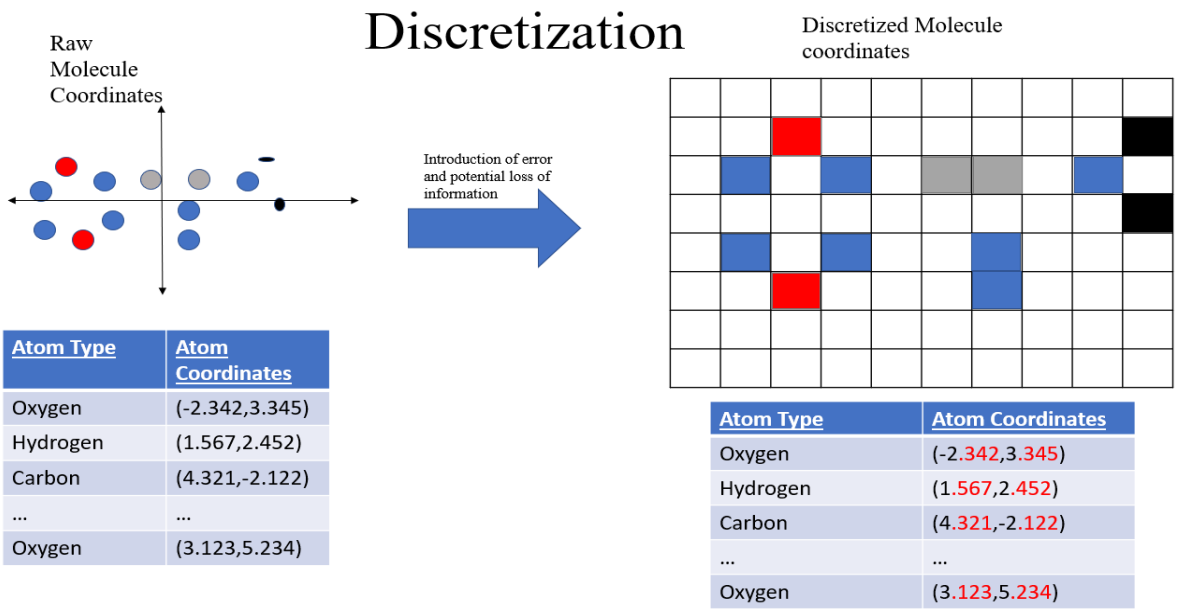


Figure 21. Discretization of Continuous Atomic Coordinates

were developed with image translation in mind, we also included 2D and 3D molecular representations in our portfolio of experiments, which were obtained in the MOL2 format. MOL2 files contain a list of all the atoms and their coordinates, along with bond and other chemical information about a molecule. Given that machine learning models cannot operate directly on textual information and need consistently sized inputs, simple lists of coordinates will not suffice. The molecules need to be discretized so the models can use a uniform representation. To do this, a resolution needs to be chosen. Numerous resolutions were tested, including 256x256x11, 32x32x32x11, 32x32x11, 64x64x64x11 and 64x64x11 matrices. The final dimension of length 11 represents the different atoms that can be contained in any given cell. This introduces a few difficulties into the process. First, this causes our data to become extremely sparse, with

higher resolution only magnifying the sparsity. Machine learning models are known to have a harder time learning with sparse data. Second, the size of the data also increases massively as matrix cells need to be maintained for empty space. Third, the act of discretization introduces rounding error into the positioning of atoms. Finally, as previously mentioned, there are many ML aided representations of molecules. In an attempt to gain the representationally flexible advantages of 2D encoding, while bypassing the sparsity and size imposed, we explored the ecosystem of variational autoencoders including JT-VAE and Chemical VAE.

Chemical VAE tailored the variational autoencoder to the biochemical realm by simultaneously training the model to predict the qed, SAS, and logP (Gomez-Bombarelli, et al., 2017). These combined tasks yield a model that can accurately encode a SMILES string into a 196-dimensional vector, and then decode that vector back into the same SMILES string (Gomez-Bombarelli, et al., 2017). By optimizing the recreation objective function, the model is forced to learn information about the nature of the molecules causing the latent space to preserve important characteristics of their inputs. This makes these types of models an attractive choice for our experiments due to their strong performance and the public availability of pretrained architectures that were exposed to large datasets of molecules.

JT-VAE is a similar variational autoencoder approach that attempted to improve on the approach used in Chemical VAE by operating directly on molecular graphs rather than simply on SMILES representations (Jin, Barzilay, & Jaakkola, 2019). The authors hypothesized that this would allow the model to capture molecular similarity, a property SMILES representation is not as readily designed to capture. Similarly, they hypothesized that operating directly on the molecular graphs would allow for improved expression of the molecular properties such as SAS, qed, and logP. JT-VAE also exhibited 100% validity of decoded molecules and a 76.7% accuracy in reconstruction tasks while lowering the dimensionality to 56 (Jin, Barzilay, & Jaakkola, 2019).

4.3 Survey of Publicly Available Biochemical Databases In Search of a Molecular Design Training Set

Typically, models learn how to create or alter molecules by training with as much data as possible and only successfully accomplish the task by iterating hundreds of thousands of times over large datasets. Our experiment is no exception. Numerous large databases are available that contain molecules in a variety of representations including SMILES, 2D, and 3D. From these databases we will need a large dataset of known SRC protein kinase inhibitors to analyze and emulate. In addition, we will need a baseline set of random similar small molecules to understand what differentiates the SRC set from any other molecule and to act as a control group. There is no shortage of publicly available biochemical databases, enumerating up to 166 billion molecules (Ruddigkeit, van Deursen, Blum, & Reymond, 2012) in some cases. These databases serve as catalogues to search through in order to choose molecules with desired properties. A massive amount of effort is being dedicated to building and maintaining these databases, as models can only ever be as good as the data they are fed. Examples of these databases included GDB, PubChem, ZINC, CAS, ChEMBL, and DrugBank. GDB-17 was created by capitalizing on work previously performed on the GDB-13 database. Due to inefficiency of implementation for the enumeration of GDB-13 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012), it was only able to represent molecules with 13 or less atoms. The authors of the paper operated on the molecular graphs directly and enhanced the memory efficiency of their graph analysis to overcome the limitations of GDB-13 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012). This allowed for the enumeration of much larger molecules, yielding the massive 166 billion molecules currently deposited in the database. GDB-17's 166 billion molecule size is much larger than alternative options such as PubChem-17 or CAS-17 (Ruddigkeit, van Deursen, Blum, & Reymond, 2012). The authors compared the size and composition of their molecular databank with public archives from PubChem, ChEMBL, and DrugBank. GDB-17 was the leading dataset for compliance of reference datasets while having the highest percentage of molecules with at least one small ring out of all datasets (Ruddigkeit, van Deursen, Blum, & Reymond, 2012). Another attractive component of GDB-17 is that it has a more uniform distribution of topologies than the other datasets. Notably, DrugBank-17 exhibited the most

uniform distribution for the different categories (heteroaromatic, aromatic, heterocyclic, carbocylic, acyclic) This allows us to obtain a very large set of random small molecules to serve as the foundation for our molecular generation experiments that is sufficiently representative of the underlying distribution. We obtained a sample of this database corresponding to 163,953 random small molecules from a variety of domains, with the following atoms (C,N,O,S,F,Cl,Br,I,At,Ts). Additionally, these molecules are available in a variety of formats. We gathered the smiles strings and 3D coordinates of all molecules. Next, we obtained our target set of SRC protein kinase inhibitors.

To complement our random baseline molecules found in GDB-17, we also turned to the ZINC database. ZINC specializes in commercially available drug compounds enhanced with properties about these molecules such as molecular weight, calculated logP, and number of rotatable bonds (Irwin & Shoichet, 2005). The creators of the ZINC database processed the molecules by desalting them with OpenEye's convert.py tool and filtering out bad molecules. The logP coefficient was estimated by a fragment-based implementation the authors adopted from molinspiration, a chemoinformatics software suite (Irwin & Shoichet, 2005). Additionally, ZINC maintains an active list of known protein kinase inhibitors, which allow us to create labeled a training set for any downstream machine learning processes used in our experiments. We obtained 52,348 total kinase inhibiting small molecules across 385 gene targets with an average of 200 conformations per gene. All molecules obtained from zinc have molecular weight less than 700, calculated logP between -4 and 6, less than 6 rotatable bonds, and only contain a set of 10 atoms (C, N, O, F, S, P, Cl, Br, I).

4.4 Determining the Quality of Generated Molecules: How Do We Know if a Molecule is good?

The main goal of our experiments is to create a framework for the generation of novel SRC protein kinase inhibitors. The difficulty with this is that there is no one metric to track that perfectly encapsulates this goal. As such we have chosen a portfolio of performance metrics commonly used in drug discovery research that together should allow us to analyze our molecules appropriately. Our first set of metrics, the drug property

measures qed, SAS, and logP, pertain to the molecules' chances of not only being successfully synthesized, but their ability to be absorbed and used by the body and do not translate to the kinase domain specifically. Notably, we ensure that the logP adheres to Lipinski's rule of 5 (Lipinski, 2004) for maximum chance of success in synthesis. We will also perform structural analysis on the molecules, using RDKit to obtain the number of rings. These ensure that our output molecules are "druglike" in nature and have potential to show therapeutic effects. The second kind of metric we track, % validity, monitors the underlying strategy's ability to generate valid molecules. One important thing to note about percent validity is that it is highly dependent on the quality of our variational autoencoder. All approaches tested in this experiment have a final step where they decode through this network, so they will only ever be as good as it is. The latent space is not evenly distributed with validity (Gomez-Bombarelli, et al., 2017), and some areas lead to better decode success than others. That being said, we are testing these strategies' ability to identify high potential areas in the latent space to yield valid SRC kinase inhibitors. It is important that these optimization techniques not only search in good areas but sample points in these areas that can actually be decoded back into valid SMILES strings. The third metrics, average similarity and kinase inhibition likelihood, pertain more explicitly to our main goal and will be the primary metrics to optimize. We will also compute the **Δ** values between the metrics and those scored by SRC protein kinase inhibitors to compare them more directly as the SRC set is our target. However, some of these metrics are harder to track than others, and we need to create our own models to estimate them. While there are currently large amounts of effort being dedicated to increasing our understanding of protein kinase inhibitors, we still don't know everything about them. This makes estimating the likelihood that a selected molecule is a kinase inhibitor difficult. Inconsistent representations, limited data, and poor understanding of structural motifs contribute to the difficulty of the problem which is a major obstacle in creating a potent scoring function. However, upon examination of the Chemical VAE latent space strong clusters emerged which show promise in creating this function.

One of the focal discoveries of our experiments was the fact that the Chemical VAE latent space encodes critical information about kinase inhibition properties. This is characterized by the organization of the hyperplane and the emergence of highly skewed clusters that demonstrate potential for accurate classification. During our data exploration phase, we sampled a set of 100,000 random small molecules from GDB-17, 1883 ABL1 Kinase inhibiting small molecules, and 3477 SRC kinase inhibiting small molecules both from ZINC. The 196-dimensional vectors representing these molecules were fed through principle component analysis (PCA) so that we could visualize them in two dimensions. We assigned a different color for each data point so that they could be differentiated on the graph and saw that the kinase inhibiting molecules were heavily concentrated in one area. This means that the encoded representations of these molecules conserve information critical to the nature of our target molecules.

It is also important to note that even though the cluster of kinase inhibitors isn't pure and there are green molecules from GDB-17 present within them, this does not necessarily mean that these green molecules aren't kinase inhibitors. It is valid to assume that some of these might exhibit kinase inhibition potential because there are many molecules that might simply be undiscovered kinase inhibitors. Following this observation, we decided to create a kinase inhibition likelihood scorer ourselves to fuel the molecular design. We obtained the latent variables for each of the molecules in our datasets and assigned a label of "1" for all SRC kinase inhibitors and "0" for all other molecules. After fitting a random forest to the data, our assumptions about the quality of the clusters were validated as shown by the strong performance of the model, achieving an AUC score of 0.92, a precision of 0.89, a recall of 0.96, and an F1-score of 0.92 (shown in Figure 23). These scores mean that we can rely on the model to accurately predict whether a latent variable is an SRC protein kinase inhibitor or not.

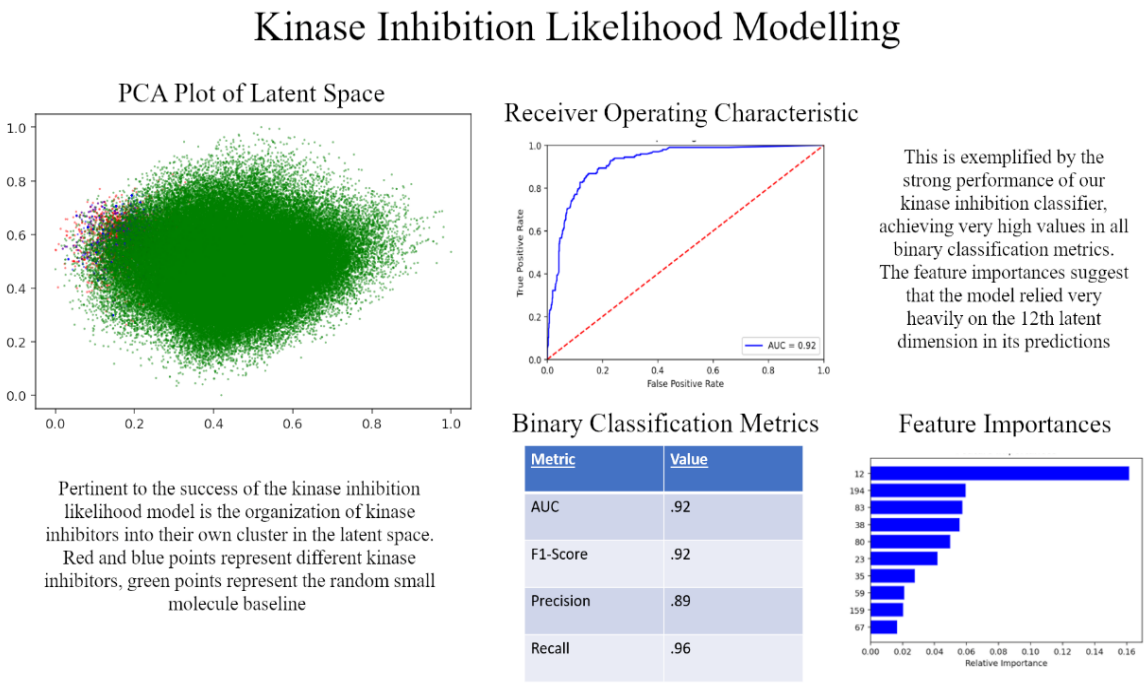


Figure 22. Classification Performance of Kinase Inhibition Likelihood Model

After successful creation of the kinase inhibition classifier, we have all the necessary ingredients to create the scoring function for our generation techniques. The model will output values close to 1 for SRC kinase inhibitors and values close to 0 for all other molecules. So, all we need to do is generate a latent variable and we can assign a score to it that we want to maximize. However, this scoring function can be enhanced by imposing additional constraints on it. In addition to kinase inhibition likelihood maximization, we also combined the SAS, qed, and logP values into the scoring function so we could increase the drug likeliness and chance for successful synthesis. We obtain these measures as a direct output of the Chemical VAE neural network. Average similarity is also computed using the python library RDKit. To do so, we obtain the Morgan fingerprints of the molecule, and compute the tanimoto similarity between it and all of either A. the SRC inhibitor target set or B. the random small molecule baseline. We take the average of all similarities to return the average similarity of the molecule. We also impose validity constraints on the created molecules by filtering out any latent locations that have a 0% decode validity. To do so, we repeat the following process. First, sample the latent representation of a given molecule and predict qed, SAS, logP, kinase inhibition likelihood, and average similarity for that data point. Next the probabilistic Chemical

VAE neural network attempts to decode the latent variable into a valid smiles string 100 times, each time feeding it into RDKit to assess validity. If the network is able to yield at least 1 valid molecule out of the 100 decode attempts, we move the molecule into the next validity stage, and if not, we return 0 for the reward function. The next validity gate checks the length of the smiles string to ensure that no molecule makes it through if it has a length less than 10. This forces our model to output real molecules and not cheat by simply creating 1 or 2 atom molecules. Early on in our experiments, we observed the model exhibiting bias towards prediction of 1 atom smiles strings, and this stage stops that from being allowed. Finally, we return the value of our reward function at that sampled latent variable. Higher values of this function will translate to better molecules.

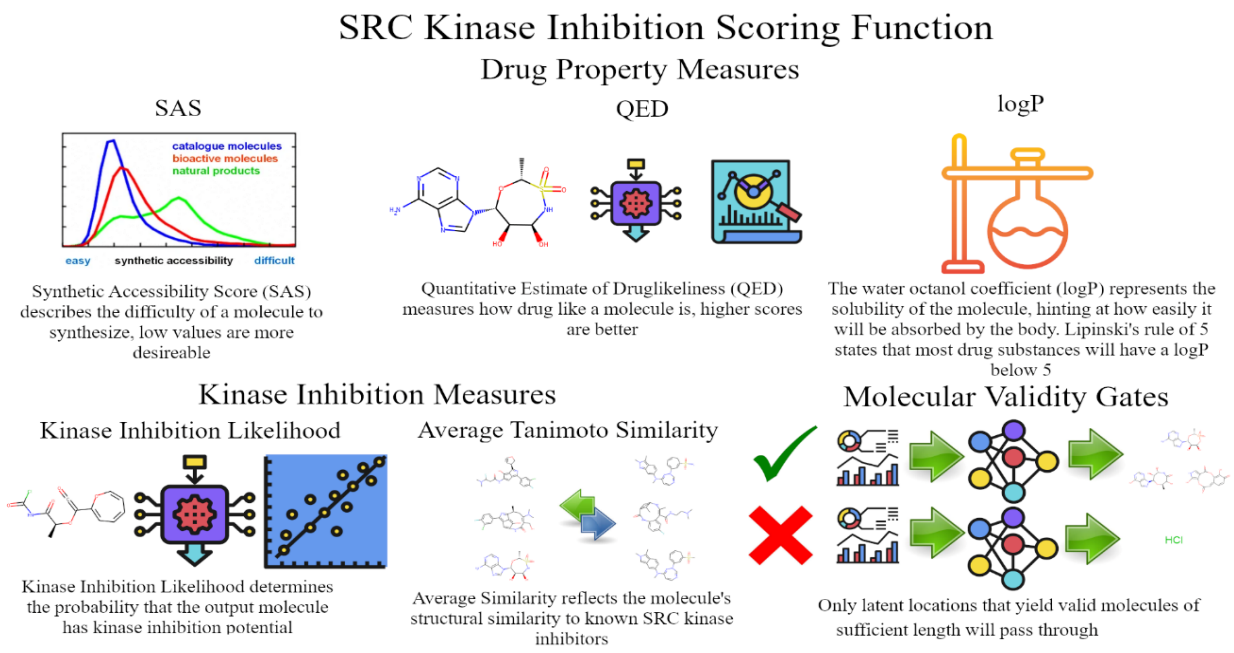


Figure 23. Presentation of Components of SRC Kinase Inhibition Scoring Function

Chapter 5: Presentation and Comparison of Strategies for Molecular Design

In this chapter, we will describe the strategies we created for molecular design and then review the results gathered from implementing the strategies described above. First, we will present each strategy for molecular design and present the aggregate results achieved in each strategy. Next, we will compare the

molecules across all strategies to compare their performance in aggregate. Finally, we analyze the top molecules from each strategy to observe if there were any high potential outliers present in the output sets.

5.1 Discussion and Creation of Strategies for Targeted Molecular Design

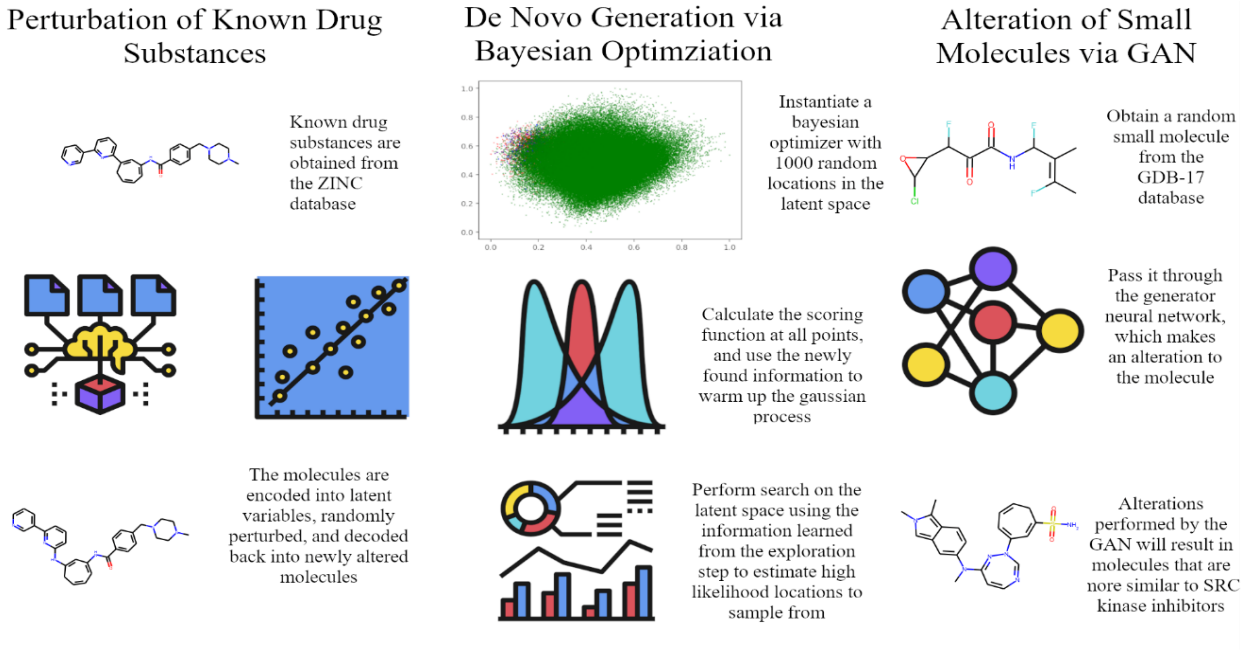


Figure 24 Presentation of three strategies for targeted molecular design

Following the acquisition of datasets for inputs and design of a framework to assess the quality of outputs, a variety of tools become available to use to accomplish the targeted molecular design goal. They vary in their degrees of complexity, but each comes with its own set of strengths. VAE are used very heavily in all approaches, and all of the below methods rely on them for successful mapping back into SMILES strings. We start with the simplest strategy, perturbation of known drug substances, move to de novo generation via Bayesian Optimization, and end with the most complex, targeted alteration via GANs.

5.1.1 Perturbation of Known Drug Substances

In addition to providing benefits, such as compression, autoencoders provide another key benefit: the creation of a continuous space that can be searched (Hinton & Salakhutdinov, 2006). In the SMILES domain, there are discrete changes from one string to another and no clear graphical representation that can

be observed or searched. This leads to difficulty in experimenting with or optimizing candidate molecules. Autoencoders solve this problem by mapping these string representations into a numeric domain where much of the structural chemical properties are maintained. This allows us to use the large ecosystem of tools available to operate on numbers and then simply decode back into the SMILES domain following experimentation. Essentially, we can create a random vector of numbers the same length as the bottleneck layer in order to convert it into a SMILES string (agnostic of validity).

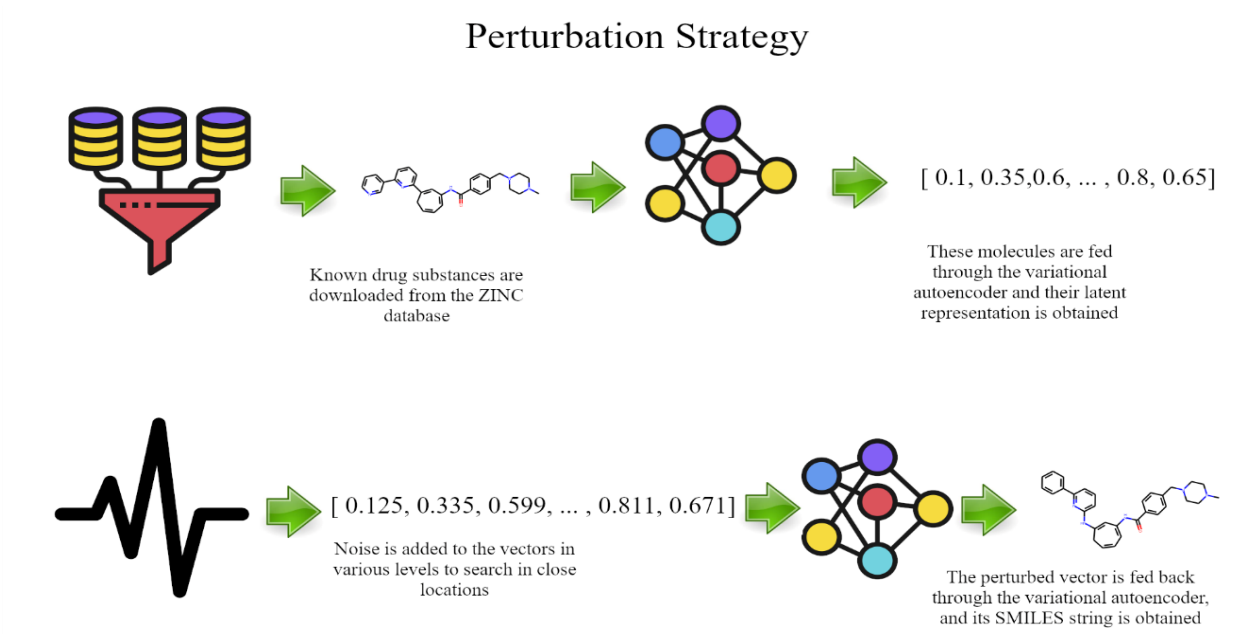


Figure 25 Perturbation strategy for molecular design

Not only is the latent space continuous but, if trained correctly, movements in the space should also correlate to movements in some integral characteristic of the inputs. In other words, molecules in the latent space should cluster themselves by some informative property. The only issue is that due to the black box nature of neural network training, it is unclear what that property might be. The assumption is that important structural properties of the molecules, such as rings and atomic composition, are encoded into one of the many latent dimensions. This allows us to feasibly hone in on the structural motifs of SRC protein kinase inhibitors in order to generate new ones. Furthermore, this tool provides us with a complete and continuous representation of the known molecular landscape, where we can find any molecule we want with the right tools.

However, searching becomes easier when you know where to look. If the assumptions hold that the variational autoencoders capture most of the information from molecules they are fed, then the best place to search for new SRC kinase inhibitors is close to other SRC kinase inhibitors. We made an attempt to create new SRC kinase inhibitors by first observing the location of known SRC kinase inhibiting drug molecules in the latent space through dimensionality reduction and utilization of principal component analysis. We then chose 24 heavily studied SRC kinase inhibitors to treat as anchor points in the latent space. After observing these locations, we searched around them by sampling from similar locations in increasing radii around the known anchor points. This yielded novel molecules with high potential to be SRC protein kinase inhibitors. We performed this experiment with increasing noise levels of 5,10,15,20,25, and 30.

Drug Compounds	Smiles
Sorafenib	CNC(=O)c1cc(Oc2ccc(NC(=O)Nc3ccc(Cl)c(C(F)(F)F)c3)cc2)ccn1
Erlotinib	C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1
Lestaurtinib	C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6ccccc6n2c5c31)CNC4=O
Amp	Nc1ncnc2c1ncn2[C@@H]1O[C@H](COP(=O)(O)O)[C@@H](O)[C@H]1O
Meletin	O=c1c(O)c(-c2ccc(O)c(O)c2)oc2cc(O)cc(O)c12
Llagate	O=c1oc2c(O)c(O)cc3c(=O)oc4c(O)c(O)cc1c4c23
Niclosamide	O=C(Nc1ccc([N+](=O)[O-])cc1Cl)c1cc(Cl)ccc1O
Neratinib	CCOc1cc2ncc(C#N)c(Nc3ccc(OCc4ccccn4)c(Cl)c3)c2cc1NC(=O)/C=C/CN(C)C
Sunitnib	CCN(CC)CCNC(=O)c1c(C)[nH]c(/C=C2\C(=O)Nc3ccc(F)cc32)c1C
Afatinib	CN(C)C/C=C/C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2cc1O[C@H]1CCOC1
Sprycel	Cc1nc(Nc2ncc(C(=O)Nc3c(C)cccc3Cl)s2)cc(N2CCN(CCO)CC2)n1
Nilotinib	Cc1cn(-c2cc(NC(=O)c3ccc(C)c(Nc4nccc(-c5cccnc5)n4)c3)cc(C(F)(F)F)c2)cn1
Pazopanib	Cc1ccc(Nc2nccc(N(C)c3ccc4c(C)n(C)nc4c3)n2)cc1S(N)(=O)=O
Iressa	COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1
Imatinib	Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc(-c2cccnc2)n1
Bosulif	COc1cc(Nc2c(C#N)cnc3cc(OCCCN4CCN(C)CC4)c(OC)cc23)c(Cl)cc1Cl
ZINC23358248	COc1cc(Nc2c(C#N)cnc3cc(-c4coc(CN5CCN(C)CC5)c4)c(OC)cc23)c(Cl)cc1Cl
Xalkori	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl
Fluoxetine Glucuronide	CC(C)n1nc(-c2cc3cc(O)ccc3[nH]2)c2c(N)ncnc21
Ponatinib	Cc1ccc(C(=O)Nc2ccc(CN3CCN(C)CC3)c(C(F)(F)F)c2)cc1C#Cc1cnc2cccnn12
Caprelsa	COc1cc2/c(=N\c3ccc(Br)cc3F)nc[nH]c2cc1OCC1CCN(C)CC1
Ceritinib	Cc1cc(Nc2ncc(Cl)c(Nc3cccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1
Midostaurin	CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1(C)n1c3ccccc3c3c4c(c5c6ccccc6n2c5c31)C(=O)NC4
Nintendanib	COC(=O)c1ccc2c(c1)NC(=O)/C2=C(\Nc1ccc(N(C)C(=O)CN2CCN(C)CC2)cc1)c1ccccc1

Table 4. SRC Kinase Inhibiting Substances and their SMILES representation

The molecules created via perturbation of known drug targets are, on average, more similar to an SRC kinase inhibiting molecule than they are to an average random small molecule from the GDB-17 database. Similarly, the molecules created by the perturbation experiments achieved higher kinase inhibition scores than random small molecules from the GDB database. As a reference point, the known SRC kinase inhibiting molecules which are expected to have the highest score possible have an average kinase inhibition score of 0.462, compared to that of the best perturbation strategy: 0.350, and the random small molecule baseline: 0.02. These techniques also showed high validity yields, with an aggregated 42.3% validity. It should be noted that there was high variance in the validity depending on the drug substance being perturbed. Some drug substances, such as Sprycel, had very high validity yields whereas others, such as Lestaurtinib, had close to zero validity.

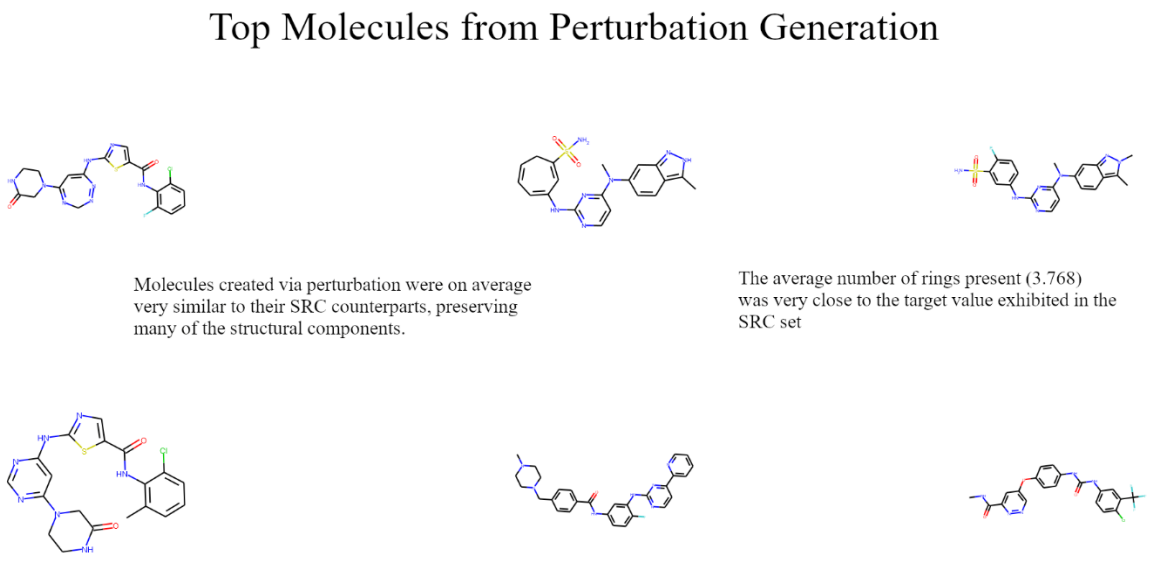


Figure 26. Top Molecules from Perturbation Generation

Noise Level 5

The molecules created using a noise level of 5 yielded an average qed of 0.600, average SAS of 3.092, and average logP of 3.804. They had an average kinase inhibition score of 0.350 and were 12.5% more similar to SRC molecules than GDB molecules, with an average similarity of 0.413 and 0.288, respectively.

Noise Level 10

The molecules created using a noise level of 10 yielded an average qed of 0.609, an average SAS of 2.919, and an average logP of 3.799. They had an average kinase inhibition score of 0.322 and were 12.4% more similar to SRC molecules than GDB molecules, with an average similarity of 0.408 and 0.284, respectively.

Noise Level 15

The molecules created using a noise level of 15 yielded an average qed of 0.610, an average SAS of 3.045, and an average logP of 3.633. They had an average kinase inhibition score of 0.358 and were 12.4% more similar to SRC molecules than GDB molecules, with an average similarity of 0.410 and 0.286, respectively

Noise Level 20

The molecules created using a noise level of 20 yielded an average qed of 0.599, an average SAS of 3.046, and an average logP of 3.885. They had an average kinase inhibition score of 0.346 and were 12.5% more similar to SRC molecules than GDB molecules, with an average similarity of 0.411 and 0.286, respectively

Noise Level 25

The molecules created using a noise level of 25 yielded an average qed of 0.613, an average SAS of 2.968, and an average logP of 3.685. They had an average kinase inhibition score of 0.325 and were 12.5% more similar to SRC molecules than GDB molecules, with an average similarity of 0.406 and 0.281, respectively

Noise Level 30

The molecules created using a noise level of 30 yielded an average qed of 0.614, an average SAS of 3.028, and an average logP of 3.742. They had an average kinase inhibition score of 0.347 and were 12.3% more similar to SRC molecules than GDB molecules, with an average similarity of 0.408 and 0.285 respectively

The noise level setting did not seem to affect the quality of the molecules very much. They all achieved very similar values for all of the metrics we tracked. When the noise level was set to 10, the molecules maximized the kinase inhibition likelihood function slightly. However, there was not enough differentiation to comment on one strategy being ultimately better. This was also shown to be dependent on the perturbation subjects used. The molecules were not evenly distributed with validity, so the quality of the output molecules is more dependent on quality of sampled drug anchor points than it is on noise level.

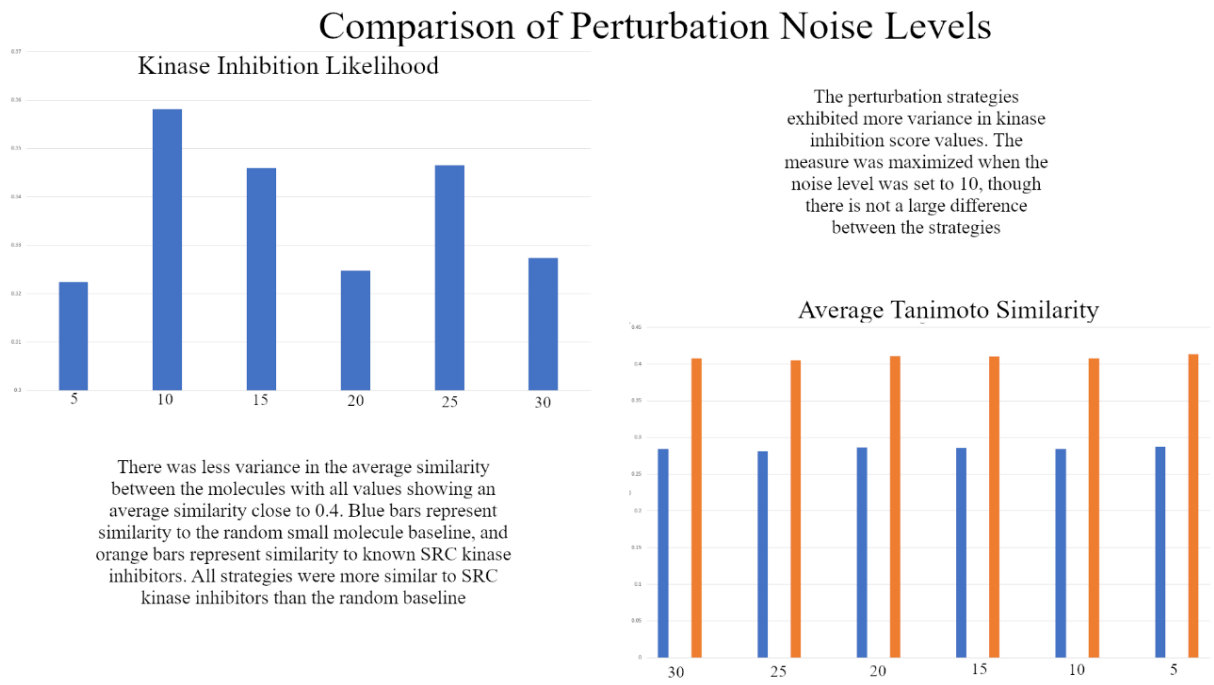


Figure 27. Comparison of Performance in Perturbation Noise Levels

5.1.2 De Novo Generation of SRC Kinase Inhibitors via Bayesian Optimization

An efficient technique is required for searching something as vast as the latent space. Every location searched and scored comes with heavy computational requirements. There are 196 dimensions to search through and no clear understanding of what these dimensions might encode. Bayesian Optimization is a technique for efficient sampling and testing of parameters with respect to a scoring function (Frazier, 2018). Heavily used in the hyperparameter tuning step of machine learning experiments (Snoek, Larochelle, & Adams, 2012), Bayesian Optimization is designed to learn information about how different parameters affect an observed function with minimal sampling. Typically, there are very large numbers of parameters that need to be tuned to reach optimal training performance. However, evaluating performance at each of these sampled parameter sets can be extremely expensive. While guaranteed to yield the best possible set of parameters, brute force approaches can be impossible to run due to computational intractability (Snoek, Larochelle, & Adams, 2012). Bayesian Optimization approaches attempt to solve that problem by training a gaussian process to approximate a function that represents expected scores with respect to hyperparameters. However, the applications of Bayesian Optimization are not just limited to

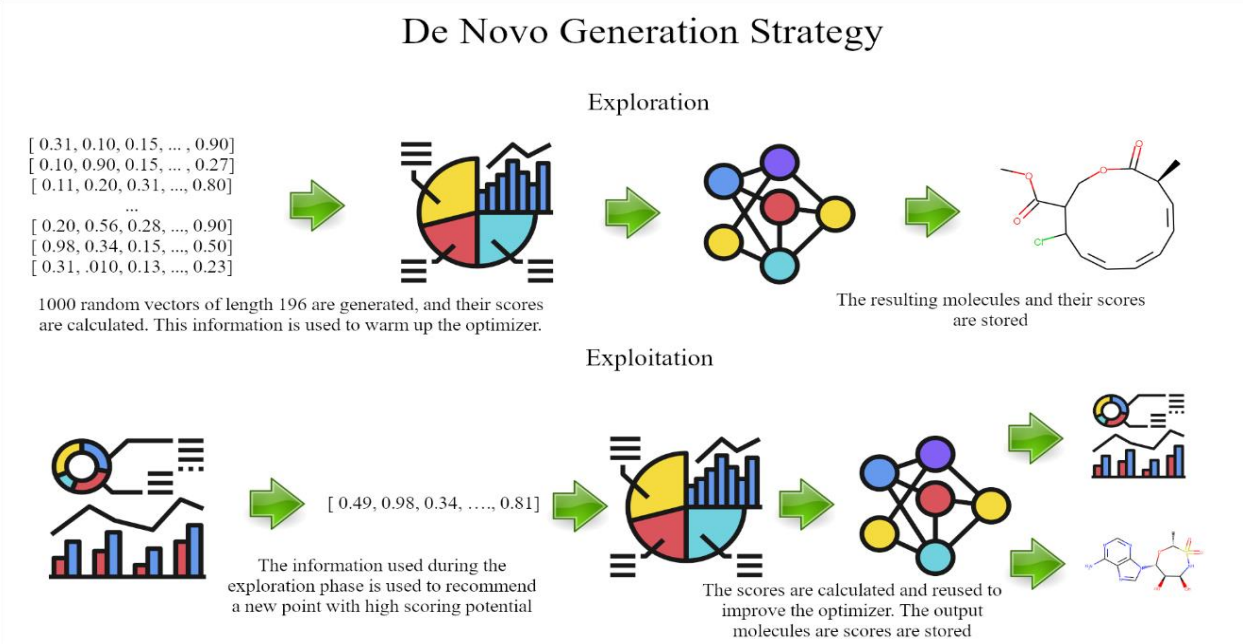


Figure 28. De novo generation of SRC kinase inhibitors via Bayesian Optimization

hyperparameter search problems. Bayesian Optimizers have been shown to perform well in molecular generation applications as well (Jin, Barzilay, & Jaakkola, 2019). Given that each of the dimensions of the latent space of a variational autoencoder can be thought of as a "parameter", it follows that de novo molecular generation can be treated as a parameter tuning exercise. The Bayesian Optimizer can estimate values for each latent dimension of an encoded molecule that maximize a scoring function. As such, this creates the ability to perform optimized generation of molecules to some property. Jin et al,. showed that they could optimize properties such as logP and SAS using Bayesian Optimization in the JT-VAE latent space (Jin, Barzilay, & Jaakkola, 2019). Given that the Bayesian Optimizer obtained state of the art performance in this task, we decided to see if it could maintain its strong performance in our protein kinase inhibitor generation experiments. However, rather than simply maximizing widely estimated values such as logP or SAS, we need to maximize a more nebulous function: kinase inhibition likelihood. Bayesian Optimization involves an exploration phase to warm up the knowledge about the scoring space and an exploitation phase to capitalize on the information learned during the exploration phase

During the exploration phase, the Bayesian Optimizer randomly generates a set of hyperparameters and then evaluates the scoring function at this particular location. Due to the fact that during this stage the optimizer does not know much about how these parameters affect the scoring function, it is simply performing a random search so that it can obtain an initial understanding of the parameters that explain variance in the scoring outcome. We are telling the Bayesian Optimizer "sample 1000 random points with no regard for exploitation, record their reward functions, and warm up your prior distribution over the latent space". This is necessary so that the breadth of the parameter space can be searched without getting stuck in local optima due to scoring bias. Many pockets of high potential scoring values are found during this step to force the model to explore more areas of the search space rather than becoming too focused on one area. There is typically a relationship between the number of parameters to be optimized and the amount of exploration steps that need to be performed for sufficient training. After all the exploration steps have been completed, the optimizer begins to exploit this information.

After the optimizer gains an initial understanding of the search space, it begins to sample at targeted locations where it expects high scoring function values. To do so, it calculates the set of parameters with the highest expected improvement, and then feeds this set of values to the scoring function. The value of the scoring function, with respect to these parameters, is used to update the optimizer's understanding of the scoring function so a potentially new set of parameters can be generated in a different location. As more points are sampled, the optimizer begins to converge to it's found optimum, yielding a set of parameters that maximize or minimize the desired scoring function

We also tried to generate new SRC kinase inhibitors without the bias of known drug molecules using Bayesian Optimization to search the latent space. To do so, we instantiated a Bayesian Optimizer to generate the latent coordinates of new molecules and update its gaussian process, maximizing our scoring function. We set the gaussian process to perform 1000 steps of random exploration of the latent space to identify regions that have the potential to maximize our score. Then it sampled 100 molecules from the areas with the highest estimated potential to have high scoring function values, capitalizing on the information learned in the exploration step.

The de novo designed molecules achieved an average kinase inhibition score of 0.327, a vast improvement over the baseline. Furthermore, these molecules are on average higher in similarity to known SRC kinase inhibiting small molecules (.316) than they are to the average small molecule in the GDB-17 database (.247). This means that the Bayesian Optimizer was able to capture some important patters in the structural information encoded into the SRC set.

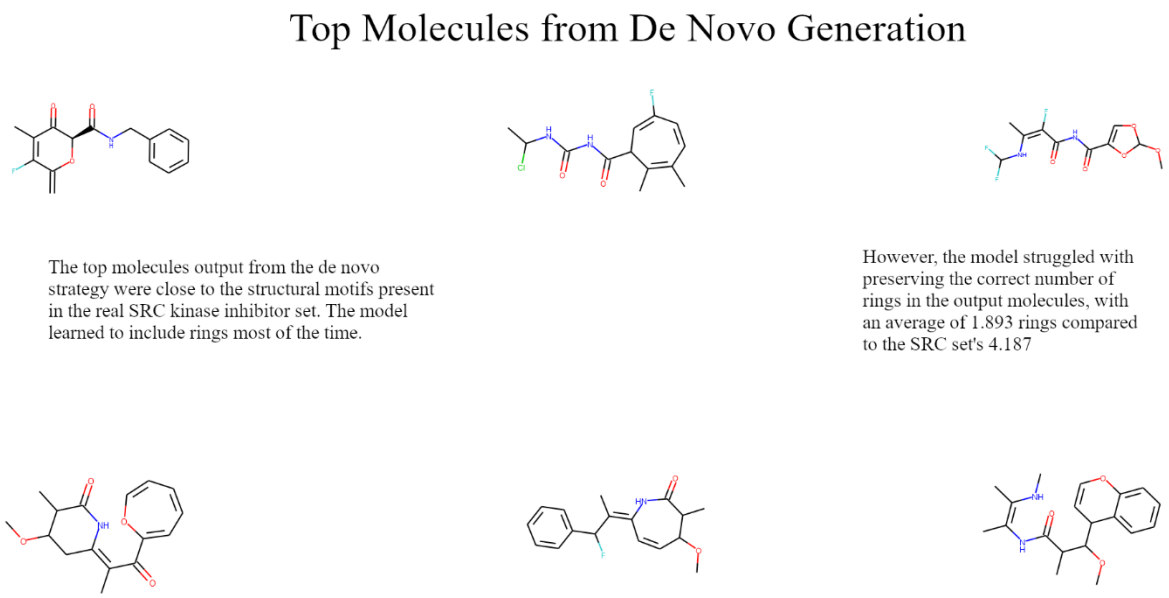


Figure 29. Top Molecules from De Novo Generation

The optimizer also maintained drug likeliness property levels near the original levels of the SRC kinase inhibitor set. These molecules had an average qed of 0.803 which is less than a 25% deviation from the SRC baseline. An average SAS score of 3.068 and average logP of 2.039 represented a 13 and 50% deviation, respectively. The De Novo strategy had a 23.2% validity yield. The top molecules approached the structure exhibited in the SRC set, though they typically fell short in the number of rings present.

5.1.3 GAN Alteration of Small Molecules

Our final strategy for generating SRC kinase inhibitors was to train a Cycle GAN to make kinase inhibiting alterations to small molecules. In this strategy, rather than perturbing known kinase inhibitors with random noise like in our first strategy, the alterations were more strategic as a result of the adversarial training algorithm as shown in Figure 31. The set X consisted of random small molecules and the set Y consisted of SRC protein kinase inhibitors. The GAN attempted to learn transformations that imbue kinase inhibiting potential onto the molecule.

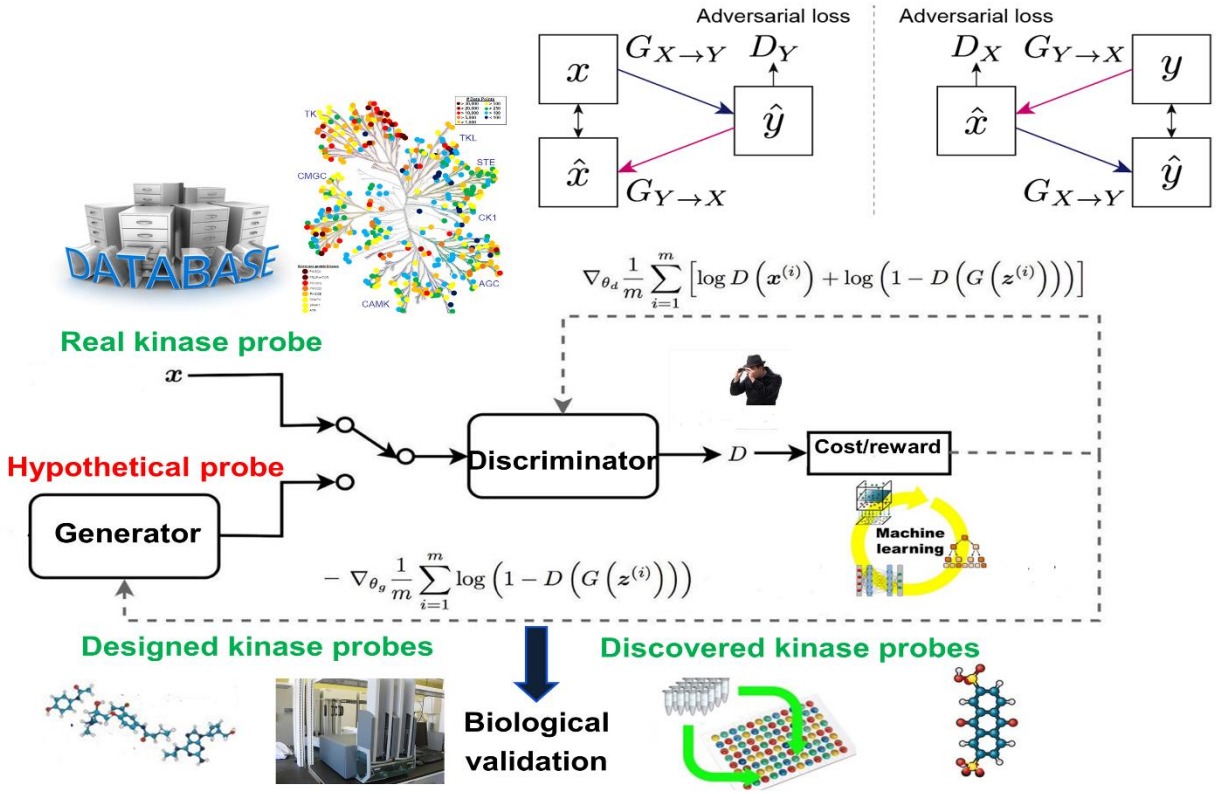


Figure 30. Kinase inhibition alteration via GAN

Machine	*Memory*	*GPU*	*GPU Memory*	*Strategies*
Personal Desktop	16 GB	GTX 1080	16 GB	Perturbation De Novo
Keck Cluster	100 GB	None	None	Perturbation De Novo
Schmid Cluster	800+ GB	8x Tesla V100	8x 32 GB	SMILES GAN 2D GAN 3D GAN

Table 5. Computing Environments Used During Experiments, the Hardware Available, and the Strategies Run on Them

GAN was the only strategy that was able to utilize SMILES, 2D, and 3D representations of the molecules. However, the model was only able to create valid outputs when fed autoencoded SMILES strings. The 2D and 3D versions of the model struggled with the sparsity observed in the inputs. Direct outputs of the model were unusable and processed versions were marginally better, therefore, we were unable to extract any valid molecules that could be used to generate statistics. Another problem that occurred was that the model seemed to struggle with the 11 channels that it had to handle. The fact that every matrix cell could contain 11 possible values corresponding to different atoms added a layer of complexity to traditional image classification. Typically, traditional image classification has three channels for red, green, and blue. This is exemplified by the density plots shown in Figure 32. The model struggled with maintaining the one-hot encoded nature of the atomic dimension of the inputs. These matrices are one-hot encoded in nature because only one atom can ever be occupying a space at one time. However, the model did not recognize this pattern because every output contained some non-zero values in all cells. The density plots above were obtained by calculating the most likely atom present in each channel. This did not yield a smooth structure as different atoms were present in cells directly next to each other. The models also shifted all feature-like

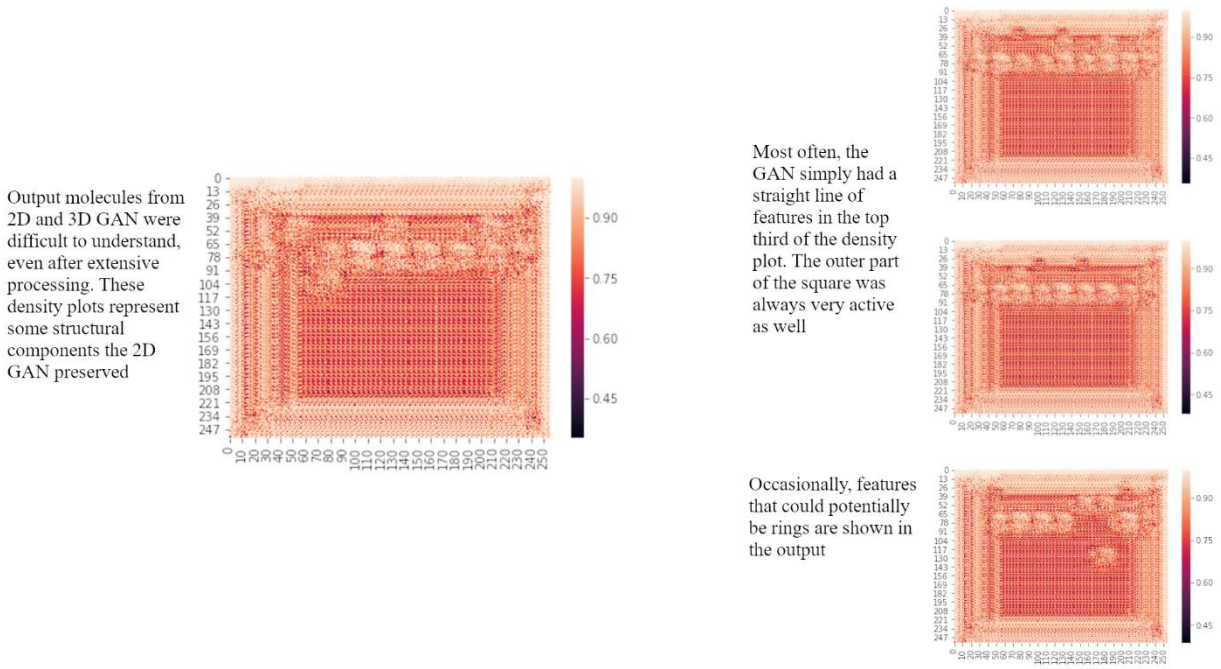


Figure 31. Processed Output of 2D GAN

components to the top third of the plot. This suggests that these models need increased training with more data or that the task (especially in the 3D case) was too difficult for the model in the given representation. While all GAN training required significant computational resources, the 3D version of the model required the most, necessitating massive hardware upgrades in order to complete (as shown in Table 2). Initially, we performed our training on a personal desktop which eventually ran out of memory. As a result, we attempted to upgrade our environment and train on Chapman's Keck Cluster, which has 84 GB more memory, however, similar results were obtained. Finally, we were required to transition to the Schmid Cluster, which has 700 GB more memory than the Keck Cluster and has access to GPUs. In this attempt, we were successful in completing the training. Ultimately, we were required to set up our environment on all three machines and develop the capabilities to move large datasets from machine to machine.

Hyperparameter	***Values Tested***
Hidden Layers	1,2,3
Batch Size	8,16,32,64
Activation Function	ReLU, Leaky ReLU
Epochs	20,40,60,80,100
Learning Rate	.1,.01,.001,.0001

Table 6. Hyperparameter Combinations Tested

Out of all SMILES GAN hyperparameter combinations attempted (shown in Table 3), the highest validity yield reached was 2.56%, excluding one and two atom molecules. This involved an architecture that includes: three hidden layers, batch size of 32, Leaky ReLU activation function and learning rate of 0.001 trained for 100 epochs. Including one and two atom molecules, the model was able to achieve 9.82% validity yield. These 2.56% valid molecules were made up mostly of complicated molecules that would be difficult to synthesize, characterized by the property values associated with the GAN output set and the top scoring molecules displayed in Figure 33. The GAN set of molecules had an average qed of 0.550, average

SAS of 4.103, and an average logP of 5.452, the only set out of the strategies tested that broke Lipinski's rule of 5. The GAN strategy slightly improved over the kinase inhibition baseline achieved by GDB-17 with a value of 0.04. The model still struggled to capture the structural similarity achieved by other strategies represented by its average similarity to SRC of 0.124 and its average similarity to GDB of 0.104. The results imply that the GAN molecules are just as similar to random molecules as they are to SRC protein kinase inhibitors.

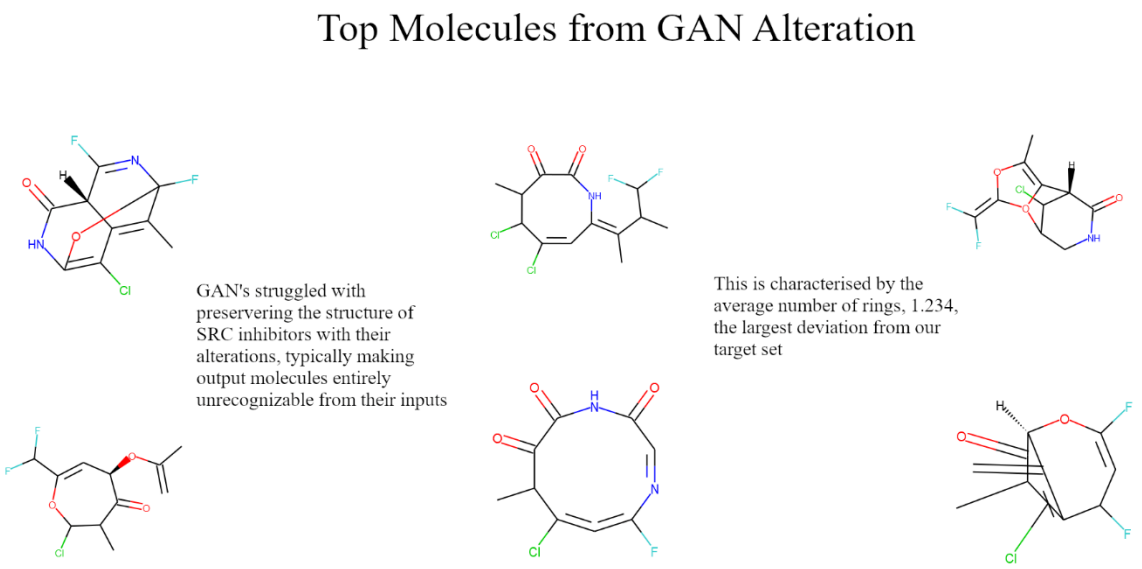


Figure 32. Top Molecules from GAN Alteration

5.2 Comparison of Performance for All Strategies

5.2.1 Comparison of the Molecules in Aggregate

While all models were able to yield real molecules, the perturbation methods outperformed the rest in most categories. Across all metrics, their values were always most in line with those seen in real SRC kinase inhibitors. They achieved the highest kinase inhibition scores, were the most structurally similar, and tended to have drug property values closest to those of kinase inhibitors. The De Novo strategy followed, with GAN performing the worst out of the group.

Metric	***GDB-17***	***SRC***	***P: 5***	***P: 10***	***P: 15***	***P: 20***	***P: 25***	***P: 30***	***De Novo***	***GAN***
Avg qed	0.791	0.591	0.600	0.609	0.610	0.599	0.613	0.614	0.803	0.550
Avg SAS	3.906	2.707	3.092	2.919	3.045	3.046	2.968	3.028	3.068	4.103
Avg LogP	2.058	4.137	3.804	3.799	3.633	3.885	3.685	3.742	2.039	5.452
Avg Kinase Inhibition Score	0.020	0.462	0.350	0.322	0.358	0.346	0.325	0.347	0.327	0.040
Avg Similarity to GDB	N/A	0.261	0.288	0.284	0.286	0.286	0.281	0.285	0.247	0.104
Avg Similarity to SRC	0.261	N/A	0.413	0.408	0.410	0.411	0.406	0.408	0.316	0.124
Δ qed	0.200	0	0.009	0.018	0.020	0.008	0.022	0.023	0.212	0.041
Δ SAS	1.200	0	0.385	0.212	0.338	0.340	0.261	0.322	0.362	1.396
Δ logP	-2.079	0	-0.333	-0.338	-0.504	-0.252	-0.452	-0.395	-2.098	1.315
Δ Kinase Inhibition Score	-0.442	0	-0.112	-0.140	-0.104	-0.116	-0.137	-0.116	-0.135	-0.422

Table 7. Aggregated Results of Design Strategies

5.2.2 Comparison of the Best Molecules from Each Strategy

The patterns shown in aggregate held when a micro level view was taken on the generated molecules. We selected the molecule that maximized the kinase inhibition likelihood score across the three overarching generation strategies and compared their scores and structural composition. The most optimal molecule created from a kinase inhibition likelihood perspective was when the noise level for our perturbation methods was set to 5, with a value of 0.569. The De novo method's best molecule had a kinase inhibition likelihood of 0.431, and GAN's was 0.105. The perturbation method's molecule was well above the average kinase inhibition score exhibited by known SRC kinase inhibitors. Notably, the de novo molecule was within 10% of this threshold. Structurally, these molecules displayed similar characteristics to SRC kinase inhibitors. A repeated pattern seen in the kinase inhibiting molecules in our datasets were 4 evenly spaced out rings, as evidenced by the fact that the set of SRC kinase inhibiting small molecules has an average of 4.187 rings. The perturbation strategies maintained this characteristic since they tended to make smaller alterations to the known drug substances, as evidenced by the four rings present in its highest scoring molecule. The De Novo strategy struggled with this concept, with its best molecule only containing two rings. Finally, the GAN exhibited the same number of rings, however there was nothing between these

rings. The perturbation molecule had a qed of 0.736, an SAS of 3.254, and a logP of 2.635. The De Novo molecule had a qed of 0.815, an SAS of 2.863, and a logP of 3.219. The GAN molecule had aqed of 0.410, an SAS of 3.131 and a logP of 3.235

The best molecule from an average tanimoto similarity perspective was still achieved by perturbation with noise level of 5, with value of 0.545. Once again, this was followed by the De Novo at 0.423, and the GAN at 0.284. Notably, the GAN's best molecule was the same for both kinase inhibition likelihood and average similarity. All strategies maintained the same number of rings in this scenario. The drug property values were similar, with the perturbation molecule achieving a 0.701, 3.203, and 2.701 for qed, SAS, and logP respectively. The De Novo had 0.801, 2.619, and 3.267 for the values, and the GAN had 0.450, 3.139, and 3.199.

Overall, the perturbation methods performed the best due to the simplicity of their task. Only required to make small alterations, these methods optimized all of our metrics in most cases. Notably, the de novo

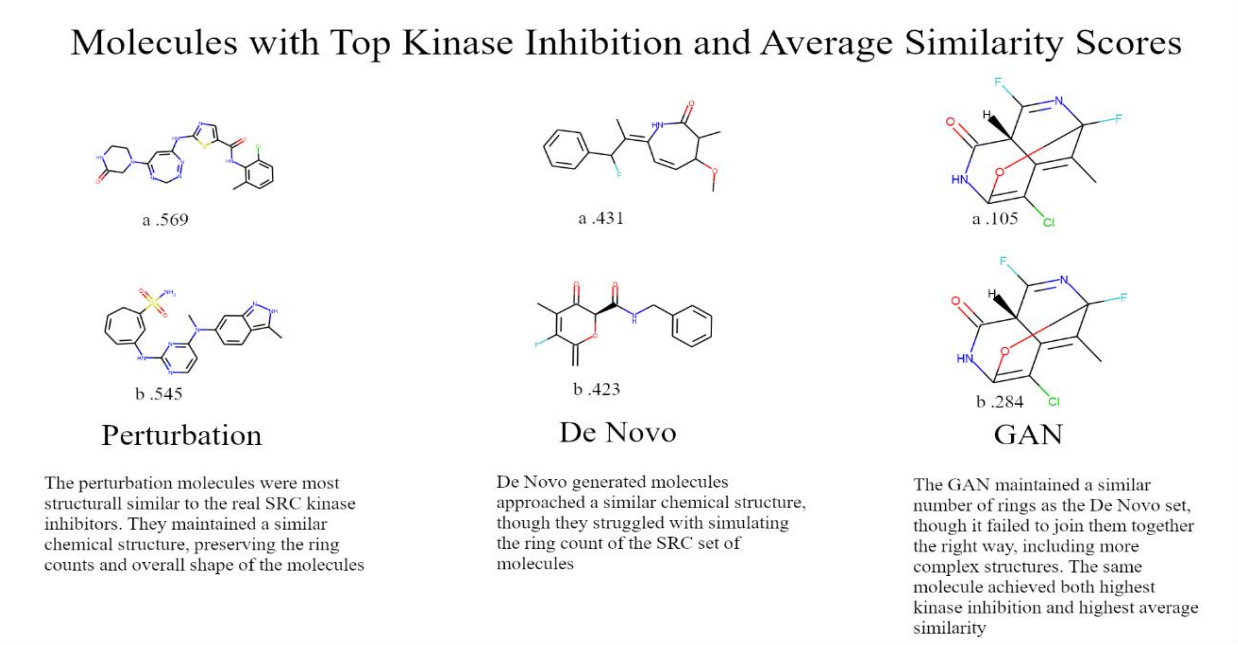


Figure 33. Molecules with the Highest Scores from all Strategies. (a) Kinase inhibition likelihood score and (b) Average Tanimoto Similarity

method was able to approach this performance even though it had the difficult task of complete generation. The molecules from both sets exhibited drug property measures well within the thresholds seen in real drug

substances. GANs had poorer performance in both of these tasks due to the size of the alterations made by the model. The resulting molecules were highly complex and looked much different from their inputs. All methods had acceptable validity yields from an in-silico design perspective.

5.3 Discussions of Results and Implications for Future Directions

The different in silico molecular generation techniques tested in our research demonstrated clear capability to create novel, valid SMILES strings that are structurally similar to known SRC protein kinase inhibitors. The average similarity of the generated molecules was always higher to SRC protein kinase inhibitors than it was to the random small molecule baseline. Simultaneously the average kinase inhibition likelihood levels approached that of the known set of SRC protein kinase inhibitors found on the ZINC database. Predictably, the molecules created via perturbing known drug substances were the highest performing on all of our tests. Given that this strategy merely altered molecules rather than complete generation, it is expected that this strategy would yield higher performing molecules. While the de novo generated molecules still beat the random small molecule baseline, the Bayesian Optimizer struggled to fully capture the nuance associated with the latent space. While all dimensions of the latent space contain some information about the molecules, the feature importance graphs show that the most variance was the most predictive in kinase inhibition potential. This suggests that some of the latent dimensions contain more structural information pertinent to kinase inhibition than others. We did not optimize for the correct number of latent variables to represent kinase inhibition potential. It remains to be investigated how many latent dimensions optimal to create kinase inhibitors. Too many, and the Bayesian Optimizer will have a hard time finding a pattern in the higher dimensional data. Too few and it will be insufficiently complex to capture the nuanced structural information encoded in SMILES strings. In our next experiment, we will test whether we can enhance the

perturbation method by fixing the values at the dimensions with highest feature importance and filling in random noise for the rest. This would allow us to increase the variance in perturbation from known drug anchor points but maintain much of the most important components that comprise an SRC protein kinase inhibitor.

Another interesting finding was that the noise level of perturbation seemed to not affect the quality of molecules. There was no clear trend in the similarity nor kinase inhibition likelihood with respect to perturbation noise level. This smooth reward plane suggests a more uniform distribution of kinase inhibition quality within the cluster locations. We observed that the higher quality molecules were obtained from the anchor point molecules with the higher success rates for recreation. There was a correlation of 0.71 between % valid molecules with 1000 decode attempts and average similarity, and a correlation of 0.65 between % valid molecules with 1000 decode attempts and average kinase inhibition score.

The De novo method performed well and picked up on most of the structural motifs present in SRC protein kinase inhibitors. However, they struggled to accurately represent the correct number of rings found in the

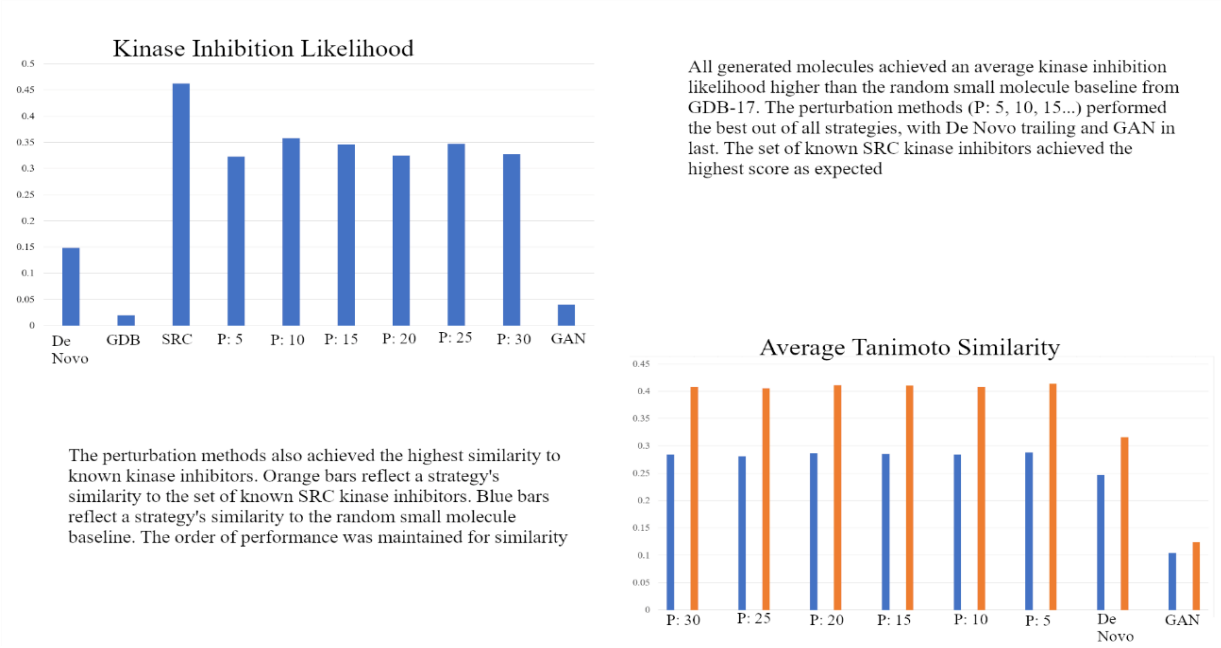


Figure 34. Presentation of Results Across Design Strategies

real set. The SRC kinase set had an average of 4.187 rings per molecule whereas the de novo set had an average of 1.893. One way that these could be enhanced to correct for this shortcoming is to add a ring count feature to the scoring function. If generated molecules have a number of rings too high or low they will be penalized. This would force the optimizer to search in areas of the latent space where molecules have close to four rings.

The GAN strategy had trouble with the size and breadth of its alterations to the molecule. Given how different the random small molecules proved to be from the known SRC kinase inhibitors, it seems that the GAN learned to make large alterations which completely changed the structure of the molecule. The resulting molecules also showed little potential for synthesis due to their highly complex nature, as shown by the drug property values present in their output set. Comparatively, perturbation techniques excelled due to the fact that they typically only made small alterations to the molecules. This suggests that the GAN could be retrained and constrained to only make small alterations, and this might help its performance. One way to approach this would be to change the training paradigm for the GAN from X = random small molecule baseline, Y = SRC set, to X = SRC set and Y = SRC set. Then the GAN would only have to learn to make smaller alterations that maintain kinase inhibition potential rather than making large alterations that imbue this potential. This GAN enhancement would require us to enhance our set of SRC kinase inhibitors. However, the perturbation and de novo strategies can be used to create this new dataset. The quality and structural similarity maintained by these methods would allow us to generate many more novel molecules that could be used in a training set. So, if we used generated molecules from both of these strategies to bolster our SRC protein kinase inhibitor dataset, the GAN might show improvement.

Furthermore, we recognized that the alterations made to the molecule by GAN would only imbue kinase inhibiting potential by making outputs resemble SRC kinase inhibitors. The loss functions used in the GAN training only captured the structural components of the molecule with no regard for actual kinase inhibiting

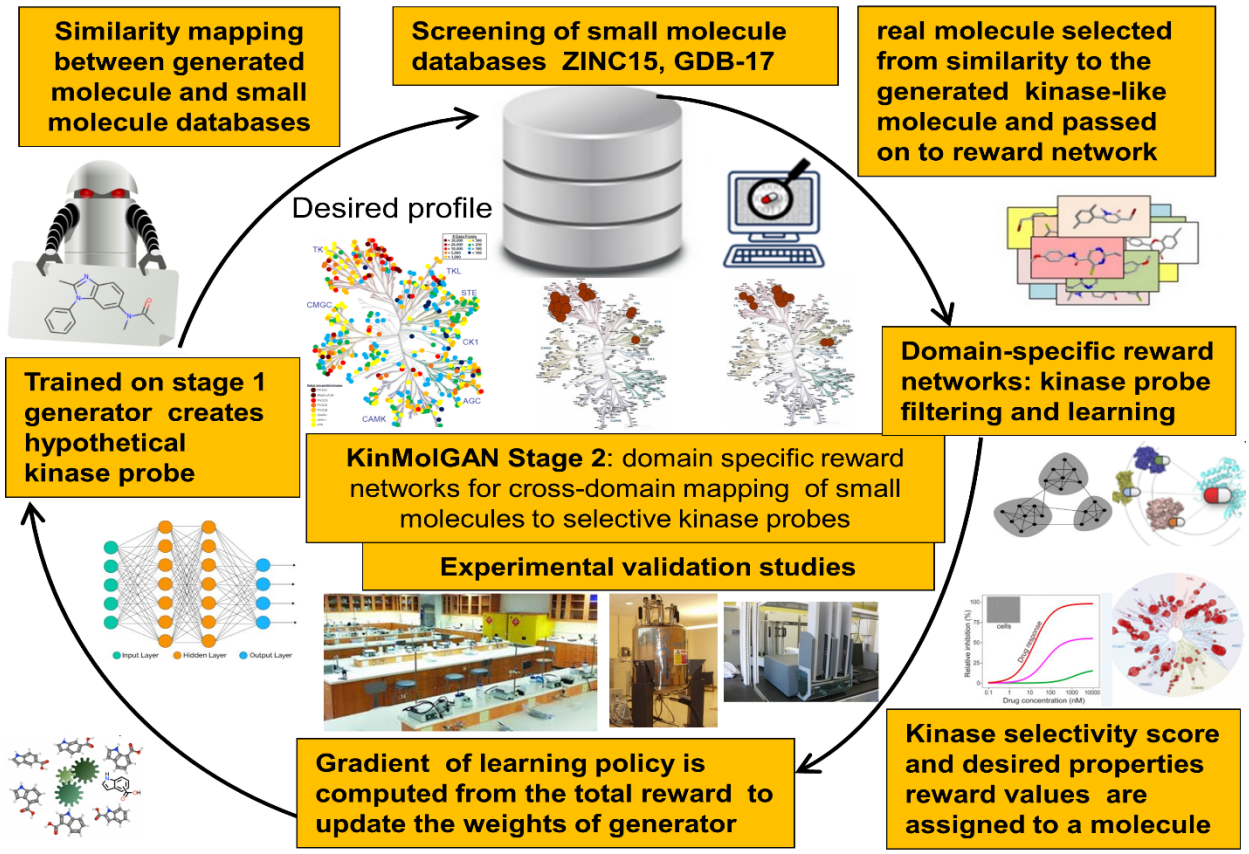


Figure 35. Stage 2 of GAN Alteration

scores, like the likelihood measure we designed or specific selectivity scores that can be calculated. We hypothesized that an additional stage 2 of training would also benefit the model as shown in Figure 36. In this stage the GAN would retrain with an additional loss functions derived from domain specific reward networks designed to model kinase inhibition while screening the small molecules through databases to identify close alternatives. We believe that this would allow the architecture to identify known molecules that might be kinase inhibitors in addition to directly generating them.

All these findings point to the conclusion that there is an observable pattern that is present within protein kinase inhibitors. Due to the increasingly large number of identified molecules present within online databases, variational autoencoders are able to capture much of the complex structural information contained within SMILES strings, and machine learning models are able to accurately pick up on some of the critical features of molecules. The growing body of computational and experimental studies has shown

that integration of data-driven biophysical and ML approaches can bring about new drug discovery paradigms, opening up unexplored venues for further scientific innovation and unique biological insights. The integration of computational and NMR approaches into a novel research platform that explores experiment-informed physical simulations, Markov state modeling, information-theoretical formalism of dynamic allosteric networks under the unified umbrella of machine learning will key to dissect molecular rules of allosteric regulation. The innovative cross-disciplinary approaches that expand the knowledge, resources and tools for studies of allosteric regulation can promote a broader usage of new technologies to understand and exploit allosteric phenomenon through the lens of chemical biology, material science, synthetic biology and bioengineering. By developing an open science infrastructure for machine learning studies of allosteric regulation and validating computational approaches using integrative studies of allosteric mechanisms, the scientific community can expand the toolkit of approaches and chemical probes for dissecting and interrogation allosteric mechanisms in many therapeutically important proteins. The development of community-accessible tools that uniquely leverage the existing experimental and simulation knowledgebase to enable interrogation of the allosteric functions can provide much needed impetus to further experimental technologies and enable steady progress.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky,, V. E., Gerasimova, A., Bork, P., & et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods 7*, 248–249. doi: 10.1038/nmeth0410-248.

Aertgeerts, K., Skene, R., Yano, J., Sang, B. C., Zou, H., Snell, G., & et al. (2011). Structural analysis of the mechanism of inhibition and allosteric activation of the kinase domain of HER2 protein. *J. Biol. Chem. 286*, 18756–18765. doi: 10.1074/jbc.M110.206193.

Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., & Verkhivker, G. M. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J. Chem. Inf. Model. 58*, 2131–2150. doi: 10.1021/acs.jcim.8b00414.

Agajanian, S., Oluyemi, O., & Verkhivker, G. M. (2019). Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modling of Cancer Driver Mutations. *Front Mol Biosci*, 6, 44.

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., & et al. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet. 50*, 1735–1743. doi: 10.1038/s41588-018-0257-y.

Angermueller, C., Parnamaa, T., Parts, L., & Stegle, O. (2016). Deep Learning for COmputational Biology. *Mol. Syst. Biol.*, 12, 878.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., & et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell 173*, 371–385.e318. doi: 10.1016/j.cell.2018.02.060.

Ballester, P. J., & Mitchell, J. B. (2010). a machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9), 1169-1175.

Bardelli, A., Parsons, D. W., Silliman, N., Ptak, J., Szabo, S., Saha, S., & et al. (2003). Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* , 300:949. doi: 10.1126/science.1082596.

Barto, A. G. (1994). Reinforcement learning control. *Curr Opin Neurobiol*, 4(6), 888-893.

Baskin, I. (2020). The power of deep learning to ligand-based novel drug discovery. *Expert Opin Drug Discov*, 1-10.

Bean, J., Riely, G. J., Balak, M., Marks, J. L., Ladanyi, M., Miller, V. A., & et al. (2008). Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. *Clin. Cancer Res 14*, 7519–7525. doi: 10.1158/1078-0432.CCR-08-0151.

Bertrand, D., Drissler, S., Chia, B. K., Koh, J. Y., Li, C., Suphavilai, C., & et al. (2018). Consensus driver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Res, 78*, 290–301. doi: 10.1158/0008-5472.CAN-17-1345.

Biau, G. (2012). Analysis of a random forest model. *J. Mach. Learn. Res., 13*, 1063–1095. Available online at: http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nat. Chem., 4*, 90– 98 DOI: 10.1038/nchem.1243.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Curr Opin Neurobiol*, 22(6), 956-962.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends Cogn Sci*, 23(5), 408-422.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1): 5-32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. A. (1984). *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Burke, J. E., Perisic, O., Masson, G. R., Vadas, O., & Williams, R. L. (2012). Oncogenic Mutations Mimic and Enance Dynamic Events in the Natural Activation of Phosphoinositide 3-Kianse p110alpha (PIK3CA). *Proc. Natl. Acad. Sci. U.S.A.*, 109, 15259-15264.

Cao, N. D., & Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. *arXiv [Preprint] [stat.ML]*, arXiv: 1805.11973.

Carey, K. D., Garton, A. J., Romero, M. S., Kahler, J., Thomson, S., Ross, S., & et al. (2006). Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Res.*, *66*, 8163–8171. doi: 10.1158/0008-5472.CAN-06-0453.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., & et al. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, *69*, 6660–6667. doi: 10.1158/0008-5472.CAN-09-1133.

Castro-Giner, F., Ratcliffe, P., & Thomlinson, I. (2015). The Mini-Driver Model of Polygenic Cancer Evolution. *Nat. Rev. Cancer*, 15, 680-685.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., & et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, *2*, 401–404. doi: 10.1158/2159-8290.CD-12-0095.

Chakrabarty, B., & Parkekh, N. (2016). NAPS: network analysis of protein structures. *Nucleic Acids Res*, *44*, W375–W382. doi: 10.1093/nar/gkw383.

Chary, M. A., Manini, A. F., Boyer, E. W., & Burns, M. (2020). The Role and Promise of Artifical Intelligence in Medical Toxicology. *J Med Toxicol*.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov Today*, 23(6), 1241-1250.

Chen, W., Tan, A. R., & Ferguson, A. L. (2018). Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *J Chem Phys*, 149(7), 072312.

Cheng, F., Zhao, J., & Zhao, Z. (2017). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinformatics*, *17*, 642–656. doi: 10.1093/bib/bbv068.

Chiavazzo, E., Covino, R., Coifman, R. R., Gear, C. W., Georgiou, A. A., & Hummer, G. (2017). Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc Natl Acad Sci U S A*, 114(28), E5495-e5503.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, 7:e46688. doi: 10.1371/journal.pone.0046688.

Chubynsky, M. V., & Thorpe, M. F. (2007). Algorithms for three-dimensional rigidity analysis and a first-order percolation transition. *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.*, 76:041135. doi: 10.1103/PhysRevE.76.041135.

Chun, S., & Fay, J. C. (2009). Identification of Deleterious Mutations within Three Human Genomes. *Genome Res.*, 19, 1553-1561.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., & et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31, 213–219. doi: 10.1038/nbt.2514.

Cortina, G. A., & Kasson, P. M. (2018). Predicting Allostery and microbial drug resistance with molecular simulations. *Curr Opin Struc Biol*, 52, 80-86.

Csáji, B. C. (2001). Approximation with Artificial Neural Networks. *Faculty of Sciences; Eötvös Loránd University, Hungary*.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., & et al. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417, 949–954. doi: 10.1038/nature00766.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., & et al. (2002). Mutations of the BRAF Gene in Human Cancer. *Nature*, 417, 949-954.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, 6:e1001025. doi: 10.1371/journal.pcbi.1001025.

De, S., & Ganesan, S. (2017). Looking Beyond Drivers and Passengers in Cancer Genome Sequencing Data. *Ann. Oncol.*, 28, 938-945.

Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., & Cavalli, A. (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics machine learning. *Nat Commun*, 6, 6155.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., & et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, 22, 1589–1598. doi: 10.1101/gr.134635.111.

Degiacomi, M. T. (2015). Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure*, 1034-1040.e1033.

del Sol, A., Fujihashi, H., Amoros, D., & Nussinov, R. (2006). Residue Centrality, Functionally Important Residues, and Active Site Shape: Analysis of Enzyme and Non-Enzyme Familes. *Protein Sci.*, 15, 2120-2128.

del Sol, A., Fujihashi, H., Amoros, D., & Nussinov, R. (2006). Residues Crucial for Maintaining Short Paths in Network Communication Mediate Signaling Proteins. *Mol. Sys. Biol.*, 2006.0019.

Deng, M., Bragelmann, J., Schultze, J. L., & Perner, S. (2016). Web-TCGA: An Online Platform for Integrated Analysis of Molecular Cancer Data Sets. *BMC Bioinformatics*, 17, 72.

Deng, Q., Xie, B., Wu, L., Ji, X., Li, C., Feng, L., & et al. (2018). Competitive evolution of NSCLC tumor clones and the drug resistance mechanism of first-generation EGFR-TKIs in Chinese NSCLC patients. *Heliyon*, 4:e01031. doi: 10.1016/j.heliyon.2018.e01031.

Dimitrov, T., Kreisbeck, C., Becker, J. S., Aspuru-Guzik, A., & Saikin, S. K. (2019). Autonomous Molecular Design: Then and Now. *ACS Appl Mater Interfaces*, 11(28), 24825-24836.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., & et al. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, 173, 305-320.e10.

Ding, L., Wendl, M. C., McMichael, J. F., & Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.*, 15, 556–570. doi: 10.1038/nrg3767.

Dixit, A., & Verkhivker, G. M. (2011). The energy landscape analysis of cancer mutations in protein kinases. *PLoS ONE*, 6:13. doi: 10.1371/journal.pone.0026071.

Dixit, A., Yi, L., Gowthaman, R., Torkamani, A., Schork, N. J., & Verkhivker, G. M. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS ONE*, 4:14. doi: 10.1371/journal.pone.0007485.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, 24, 2125–2137. doi: 10.1093/hmg/ddu733.

Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P. D., & et al. (2013). CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics*, 29, 647–648. doi: 10.1093/bioinformatics/btt017.

Eck, M. J., & Yun, C. H. (2010). Structural and mechanistic underpinnings of the differential drug sensitivity of EGFR mutations in non-small cell lung cancer. *Biochim. Biophys. Acta*, 1804, 559–566. doi: 10.1016/j.bbapap.2009.12.010.

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., & et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, 6, 271–281.e277. doi: 10.1016/j.cels.2018.03.002.

Emamzadah, S., Tropia, L., & Halazonetis, T. (2011). Crystal Structure of a Multidomain Human p53 Tetramer Bound to the Natural CDKN1A (p21) p53-Response Element. *Mol. Cancer. Res.*, 9, 1493-1499.

Engin, H. B., Kresiberg, J. F., & Carter, H. (2016). Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLos One*, 11, e0152929.

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., & Philbrick, K. (2017). Toolkits and libraries for deep learning. *J. Digit Imag.*, 30, 400–405. doi: 10.1007/s10278-017-9965-6.

Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. . *J Cheminform* , 1, 8https://doi.org/10.1186/1758-2946-1-8.

Fan, Y. X., Wong, L., Ding, J., Spiridonov, N. A., Johnson, R. C., & Johnson, G. R. (2008). Mutational activation of ErbB2 reveals a new protein kinase autoinhibition mechanism. *J. Biol. Chem.*, 283, 1588–1596. doi: 10.1074/jbc.M708116200.

Fan, Y., Xi, L., Hughes, D. S., Zhang, J., Futreal, P. A., Wheeler, D. A., & et al. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, 17:178. doi: 10.1186/s13059-016-1029-6.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., & et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, 43, D805–811. doi: 10.1093/nar/gku1075.

Fraczkiewicz, R., & Braun, W. (1998). Exact and Efficient Analytical Calcuation of the Accessible Surface Areas and their Gradients for Macromolecules. *J. Comput. Chem.*, 19, 319-333.

Frazier, P. I. (2018). A Tutorial on Bayesian Optimization. *arXiv [stat.ML]*, arXiv:1807.02811 .

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., & et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer*, 4, 177–183. doi: 10.1038/nrc1299.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., & et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, 6:pl1. doi: 10.1126/scisignal.2004088.

Gao, J., Chang, M. T., Johnsen, H. C., Gao, S. P., Sylvester, B. E., Sumer, S. O., & et al. (2017). 3D Clusters of Somatic Mutations in Cancer Reveal Numerous Rare Mutations as Functional Targets. *Genome Med.*, 9, 4.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25, i54–62. doi: 10.1093/bioinformatics/btp190.

Gauthier, N. P., Reznik, E., Gao, J., Sumer, S. O., Schultz, N., Sander, C., & et al. (2016). MutationAligner: a resource of recurrent mutation hotspots in protein domains in cancer. *Nucleic Acids Res.*, 44, D986–991. doi: 10.1093/nar/gkv1132.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., & Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, 14 (Suppl 3):S7. doi: 10.1186/1471-2164-14-S8-S7.

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.*, 38, 1291–1307. doi: 10.1002/jcc.24764.

Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernandez-Lobato, J. M., Sanchez-Lengeling, B., Sheberla, D., & al., e. (2017). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.*, *4*, 268-276 doi: 10.1021/acscentsci.7b00572.

Gonzalez-Perez, A., & Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, 88, 440–449. doi: 10.1016/j.ajhg.2011.03.004.

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, 40:e169. doi: 10.1093/nar/gks743.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., & et al. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, 10, 723–729. doi: 10.1038/nmeth.2562.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *arXiv [stat.ML]*, arXiv:1406.2661.

Grebner, C., Matter, H., Plowright, A. T., & Hessler, G. (2020). Automated De Novo Design in Medicinal Chemistry: Which Types of CHemistry Does a Generative Neural Network Learn? *J Med Chem*.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., & et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446, 153–158. doi: 10.1038/nature05610.

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320, 369–387. doi: 10.1016/S0022-2836(02)00442-4.

Gupta, A., Muller, A. T., Huisman, B. H., Fuchs, J. A., Schneider, P., & Schneider, G. (2018). Generatie Recurrent Neural Networks for De Novo Drug Design. *Mol Inform*, 37(1-2).

Gymnopoulos, M., Elsliger, M. A., & Vogt, P. K. (2007). Rare Cancer-Specific Mutations in PIK3CA Show Gain of Function. *Proc. Natl. Acad. Sci. U.S.A*, 104, 5569-5574.

Haber, D. A., & Settleman, J. (2007). Cancer: drivers and passengers. *Nature*, 446, 145–146. doi: 10.1038/446145a.

Han, R., Chen, K., & Tan, C. (2020). Curiosity-driven recommendation strategy for adaptive learning via deep reinforement learning. *Br J Math Stat Psychol*.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artifical Intelligence. *Neuron*, 95(2), 245-258.

Hernandez, C. X., Wayment-Steele, H. K., Sultan, M. M., Husic, B. E., & Pande, V. S. (2018). Variational encoding of complex dynamics. *Phys Rev E*, 97(6-1), 062412.

Hespenheide, B. M., Rader, A. J., Thorpe, M. F., & Kuhn, L. A. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.*, 21, 195–207. doi: 10.1016/S1093-3263(02)00146-8.

Hey, T., Butler, K., Jackson, S., & Thiyagalingam, J. (2020). Machine learning and big scientific data. *Philos Trans A Math Phys Eng Sci*, 378(2166), 20190054.

Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Kerlavage, A. R., & Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell. Dev. Biol.*, 5:83. doi: 10.3389/fcell.2017.00083.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*, 504–507 doi: 10.1126/science.1127647.

Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge Discovery and interactive Data Mining in Bioinformatics -- State-of-the-Art, future challenges and research directions. *BMC Bioinformatics*, 15, Suppl6, I1.

Huang, C. H., Mandelker, D., Schmidt-Kittler, O., Samuels, Y., Velculescu, V. E., Kinzler, K. W., & et al. (2007). The Structure of a Human p110alpha/p85alpha Complex Elucidates the Effects of Oncogenic 913Kalpha Mutations. *Science*, 318, 1744-1748.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., & et al. (2010). International network of cancer genome projects. *Nature*, 464, 993–998. doi: 10.1038/nature08987.

Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *J Am Chem Soc*, 140(7), 2386-2396.

Imagawa, T., Terai, T., Yamada, Y., Kamada, R., & Sakaguchi, K. (2008). Evaluation of Transcriptional Activity of p53 in Individual Living Mammalian Cells. *Anal. Biochem.*, 387, 249-256.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*, *45*(1), 177-182 doi: 10.1021/ci049714.

Jacobs, D. J., Rader, A. J., Kuhn, L. A., & Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, 44, 150–165. doi: 10.1002/prot.1081.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., & Castaneda, A. G. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443), 859-865.

James, K. A., & Verkhivker, G. M. (2014). Structure-based network analysis of activation mechanisms in the ErbB family of receptor tyrosine kinases: the regulatory spine residues are global mediators of structural stability and allosteric interactions. *PLoS ONE*, 9:e113488. doi: 10.1371/journal.pone.0113488.

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, 130, 453–459. doi: 10.1182/blood-2017-03-735654.

Jiang, R., Yang, H., Zhou, L., Kuo, C. C., Sun, F., & Chen, T. (2007). Sequence-Based Prioritization of Nonsynonymous Single-Nucleotide Polymorphisms for the Study of Disease Mutations. *Am. J. Hum. Gene.t*, 81, 346-360.

Jin, W., Barzilay, R., & Jaakkola, T. (2019). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv [cs.LG]*, arXiv:1802.04364v4.

Jing, Y., Bian, Y., Hu, Z., Wang, L., & Xie, X. S. (2018). Deep Learning for Drug Design: an Artifical Intelligence Paradigm for Drug DIscovery in the Big Data Era. *AAPS J.*, 20, 58.

Joerger, A. C., & Fersht, A. R. (2007). Structural Biology of the Tumor Suppressor p53 and Cancer-Associated mutants. *Adv Cancer Res*, 97, 1-23.

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., & Khrabrov, K. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7), 10883-10890.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics, 14*(9), 3098-3104 https://doi.org/10.1021/acs.molpharmaceut.7b00346.

Kamada, R., Nomura, T., Anderson, C. W., & Sakaguchi, K. (2011). Cancer-Associated p53 Tetramerization Domain Mutants: Quantitative Analysis Reveals a Low Threshold for Tumor Suppressor Inactivation. *J. Biol. Chem.*, 286, 252-258.

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., & et al. (2015). Comprehensive Assessment of Cancer Missense Mutation Clustering in Protein Structures. *Proc Natl Acad Sci U S A*, 112, E5486-E5495.

Kawaguchi, T., Kato, S., Otsuka, K., Watanabe, G., Kumabe, T., Tominaga, T., . . . Ishioka, C. (2005). The Relationship among p53 Oligomer Formation, Structure and Transcriptional Activity Using a Comprehensive Missense Mutation Library. *Oncogene*, 24, 6976-6981.

Kiel, C., Benisty, H., Llorens-Rico, V., & Serrano, L. (2016). The yin-yang of kinase activation and unfolding explains the peculiarity of Val600 in the activation segment of BRAF. *Elife*, 5:e12814. doi: 10.7554/eLife.12814.

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., & Bae, H. J. (2019). Deep Learning in MEdical Imaging. *Neurospine*, 16(4), 657-668.

Kircher, M., Written, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Schendure, J. (2014). A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.*, 46, 310-315.

Klonowska, K., Czubak, K., Wojciechowska, M., Handschuh, L., Zmienko, A., Figlerwicz, M., & et al. (2016). Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget*, 7, 176–192. doi: 10.18632/oncotarget.6128.

Kobayashi, S., Boggon, T. J., Dayaram, T., Janne, P. A., Kocher, O., Meyerson, M., & et al. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.*, 352, 786–792. doi: 10.1056/NEJMoa044238.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., & et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22, 568–576. doi: 10.1101/gr.129684.111.

Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol Pharm*, 14(12), 4462-4475.

Kovacs, E., Zorn, J. A., Huang, Y., Barros, T., & Kuriyan, J. (2015). A Structural Perspective on the Regulation of the Epidermal Growth Factor Receptor. *Annu. Rev. Biochem.*, 84, 739-764.

Kruger, D. M., Rathi, P. C., Pfleger, C., & Gohlke, H. (2013). CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res.*, 41, W340–W348. doi: 10.1093/nar/gkt292.

Kuzminykh, D., Polykovskiy, D., Artur Kadurin, A., Zhebrak, A., Baskov, I., Nikolenko, S., . . . Zhavoronkov, A. (2018). 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharmaceutics, 15*(10), 4378–4385 https://doi.org/10.1021/acs.molpharmaceut.7b01134.

La Sala, G., Decherchi, S., De Vivo, M., & Rocchia, W. (2017). Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent Sci*, 3(9), 949-960.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., & et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, 44:e108. doi: 10.1093/nar/gkw227.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbot, T. E., Dooling, D. J., & et al. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28, 311–317. doi: 10.1093/bioinformatics/btr665.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., & et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499, 214–218. doi: 10.1038/nature12213.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436-444.

Lehmann, K. V., & Chen, T. (2013). Exploring Functional Variant Discovery in Non-Coding Regions with SInBaD. *Nucleic Acids Res.*, 41, e7.

Lemmon, M. A., & Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell*, 141, 1117–1134. doi: 10.1016/j.cell.2010.06.011.

Li, J., Drubay, D., Michiels, S., & Gautheret, D. (2015). Mining the coding and non-coding genome for cancer drivers. *Cancer Lett.*, 369, 307–315. doi: 10.1016/j.canlet.2015.09.015.

Li, Z., Kermode, J. R., & De Vita, A. (2015). Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys Rev Lett*, 114(9), 096405.

Lipinski, C. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies, 1*(4), 337–341 doi:10.1016/j.ddtec.2004.11.007.

Littlefield, P., & Jura, N. (2013). EGFR lung cancer mutants get specialized. *Proc. Natl. Acad. Sci. U.S.A.*, 110, 15169–15170. doi: 10.1073/pnas.1314719110.

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, 34, E2393–2402. doi: 10.1002/humu.22376.

Liu, X., Jian, X., & Boerwinkle, K. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, 32, 894–899. doi: 10.1002/humu.21517.

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, 37, 235–241. doi: 10.1002/humu.22932.

Luo, P., Ding, Y., Lei, X., & Wu, F. X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.*, 10:13. doi: 10.3389/fgene.2019.00013.

Madhunapantula, S. V., & Robertson, G. P. (2009). The PTEN-AKT3 Signaling Cascase as a Therapeutic Target in Melanoma. *Pigment Cell Melanoma Res.*, 22, 400-419.

Mandelker, D., Gabelli, S. B., Schmidt-Kittler, O., Zhu, J., Cheong, I., Huang, C. H., & et al. (2009). A Frequent Kinase Domain Mutation that Changes the Interaction between P13Kalpha and the Membrane. *Proc. Natl. Acad. Sci. U.S.A*, 106, 16996-17001.

Mankoo, P. K., Sukumar, S., & Karchin, R. (2009). PIK3CA Somatic Mutations in Breast Cancer: Mechanistic Insights from Langevin Dynamics Simulations. *Proteins*, 75, 499-508.

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G. B., & Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS ONE*, 8:e77945. doi: 10.1371/journal.pone.0077945.

Martd, A., Pasquali, L., Wu, H., & Noe, F. (2018). VAMPnets for deep learning of molecular kinetics. *Nat Commun*, 9(1), 5.

Martelotto, L. G., Ng, C. K., De Filippo, M. R., Zhang, Y., Piscuoglio, S., Lim, R. S., & et al. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol*, 15:484. doi: 10.1186/s13059-014-0484-1.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., & et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 27, 240-248.

Masica, D. L., Douville, C., Tokheim, C., Bhattacharya, R., Kim, R., Moad, K., & et al. (2017). CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Res.*, 77, e35–e38. doi: 10.1158/0008-5472.CAN-17-0338.

Mater, A. C., & Coote, M. L. (2017). Deep Learning in Chemistry. *J Chem Inf Model*, 59(6), 2545-2559.

Maziarka, Ł., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., & Warchoł, M. (2020). Mol-CycleGAN: a generative model for molecular optimization. J Cheminform 12, 2 (2020). https://doi.org/10.1186/s13321-019-0404-1. *Journal of Chemoinformatics*, J. Cheminform. 12:2. doi: 10.1186/s13321-019-0404-1.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimations of Word Representations in Vector Space. *arXiv:1301.3781 [cs.CL]*, Available online at: https://arxiv.org/abs/1301.3781.

Miller, M. S., Schmidt-Kittler, O., Bolduc, D. M., Brower, E. T., Chaves-Moreira, D., Allaire, M., & et al. (2014). Structural Basis of nSH2 Regulation and Lipid Binding in P13Kalpha. *Oncotarget*, 5, 5198-5208.

Min, S., Lee, B., & Yoon, S. (2017). Deep Learning in Bioinformatics. *Brief. Bioinform.*, 18, 851-869.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., & Bellemare, M. G. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., & Lopez-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, 17:128. doi: 10.1186/s13059-016-0994-0.

Ng, P. K., Li, J., Jeong, K. J., Shao, S., Chen, H., Tsang, Y. H., & et al. (2018). Systematic functional annotation of somatic mutations in cancer. *Cancer Cell*, 33, 450–462.e410. doi: 10.1016/j.ccell.2018.01.021.

Nguyen, D. D., Xia, K., & Wei, G. W. (2016). Generalized Flexibility-Rigidity Index for Protein-Nucleic Acid Flexibility and Fluctuation Analysis. *J. Comput. Chem.*, 37, 1283-1295.

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., & et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.*, 48, 827–837. doi: 10.1038/ng.3586.

Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J Cheminform*, 9(1), 48.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: a large benchmark suite for machine learning evaluation comparison. *BioData Min*, 10, 36.

Opron, K., Xia, K., & Wei, G. W. (2014). Fast and Anisotropic Flexibility-Rigidity Index for Protein Flexibility and Fluctuation Analysis. *J. Chem. Phys.*, 140, 234105.

Opron, K., Xia, K., Burton, Z., & Wei, G. W. (2016). Flexibility-Rigidity Index for Protein-Nucleic Acid Flexibility and Fluctuation Analysis. *J. Comput Chem*, 37, 1283-1295.

Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., & et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304, 1497–1500. doi: 10.1126/science.1099314.

Paladino, A., Morra, G., & Colombo, G. (2015). Structural stability and flexibility direct the selection of activating mutations in epidermal growth factor receptor kinase. *J. Chem. Inf. Model.*, 55, 1377–1387. doi: 10.1021/acs.jcim.5b00270.

Palazon-Bru, A., Folgado-de la Rosa, D. M., Cortes-Castell, E., Lopez-Cascales, M. T., & GilGuillen, V. F. (2017). Sample Size Calculation to Externally Validate Scoring Systems Based on Logistic Regression Models. *PLoS One*, 12, e0176726.

Parthiban, V., Gromiha, M. M., & Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.*, 34, W239–242. doi: 10.1093/nar/gkl190.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, 12, 2825–2830. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pfleger, C., Radestock, S., Schmidt, E., & Gohlke, H. (2013a). Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.*, 34, 220–233. doi: 10.1002/jcc.23122.

Pfleger, C., Rathi, P. C., Klein, D. L., Radestock, S., & Gohlke, H. (2013b). Constraint Network Analysis (CNA): a python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.*, 53, 1007–1015. doi: 10.1021/ci400044m.

Piraino, S. W., & Furney, S. J. (2016). Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann. Oncol.*, 27, 240–248. doi: 10.1093/annonc/mdv561.

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., & Mamoshina, P. (2018). Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol Pharm*, 15(10), 4398-4405.

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., & et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, 36, 983–987. doi: 10.1038/nbt.4235.

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci Adv*, 4(7), eaap7885.

Poulos, R. C., & Wong, J. W. (2018). Finding cancer driver mutations in the era of big data research. *Biophys. Rev.*, 11, 21–29. doi: 10.1007/s12551-018-0415-6.

Racz, A., Bajusz, D., & Heberger, K. (2019). Multi-level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules*, 24(15).

Rader, A. J., Hespenheide, B. M., Kuhn, L. A., & Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U.S.A.*, 99, 3540–3545. doi: 10.1073/pnas.062492699.

Rajakulendran, T., Sahmi, M., Lefrancois, M., Sicheri, F., & Therrien, M. (2009). A Dimerization-Dependent Mechanism Drivers RAF Catalytic Activation. *Nature*, 461, 542-545.

Raphael, B. J., Dobson, J. R., Oesper, L., & Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.*, 6:5. doi: 10.1186/gm524.

Red Brewer, M., Yun, C., Lai, D., Lemmon, M. A., Eck, M. J., & Pao, W. (2013). Mechanism for activation of mutated epidermal growth factor receptors in lung cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 110, E3595–3604. doi: 10.1073/pnas.1220050110.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, 39:e118. doi: 10.1093/nar/gkr407.

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J. M., Kim, J., & et al. (2017). Recurrent and Functional Regulartory Mutations in Breast Cancer. *Nature*, 547, 55-60.

Ritchie, G. R., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional Annotation of Noncoding Sequence Variants. *Nat. Methods*, 11, 294-296.

Rodriguez-Escudero, I., Oliver, M. D., Andres-Pons, A., Molina, M., Cid, V. J., & Pulido, R. (2011). A Comprehensive Functional Analysis of PTEN Mutations: Impl,ications in Tumor- and Autism-Related Syndromes. *Hum. Mol. Genet.*, 20, 4132-4142.

Roskoski, R. (2014). The ErbB/HER family of protein-tyrosine kinases and cancer. *Pharmacol. Res.*, 79, 34–74. doi: 10.1016/j.phrs.2013.11.002.

Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, *52*(11), 2864-2875 DOI: 10.1021/ci30041.

Rupp, M., Tkatchenko, A., Muller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett*, 108(5), 058301.

S. (n.d.).

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., & et al. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science*, 304:554. doi: 10.1126/science.1096502.

Scheffler, K. S., Halpern, A. L., Bekritsky, M. A., Noh, E., Kallberg, M., & et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, 15, 591–594. doi: 10.1038/s41592-018-0051-x.

Schwarz, J. M., Rodelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, 7, 575–576. doi: 10.1038/nmeth0810-575.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.*, 33, W382–388. doi: 10.1093/nar/gki387.

Seidinger, A. L., Fortes, F. P., Mastellaro, M. J., Cardinalli, I. A., Zambaldi, L. G., Aguinar, S. S., & Yunes, J. A. (2015). Occurrence of Neuroblastoma among TP53 p.R337H Carriers. *PLoS One*, 10, e0140356.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., & Green, T. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins*, 87(12), 1141-1148.

Sethi, A., Eargle, J., Black, A. A., & Luthey-Schulten, Z. (2009). Dynamical networks in tRNA: protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 106, 6620–6625. doi: 10.1073/pnas.0810961106.

Shamsi, Z., Cheng, K. J., & Shukla, D. (2018). Reinforcement Learning Based Adaptive Sampling: REAPing Rewards by Exploring Protein Conformational Landscapes. *J Phys Chem B*, 122(35), 8386-8395.

Shan, Y., Arkhipov, A., Kim, E. T., Pan, A. C., & Shaw, D. E. (2013). Transitions to Catalytically Inactive Conformations in EGFR Kinase. *Proc. Natl. Acad. Sci. U.S.A.*, 110, 7270-7275.

Shan, Y., Eastwood, M. P., Zhang, X., Kim, E. T., Arkhipov, A., Dror, R., & et al. (2012). Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell*, 149, 860-870.

Shcherbinin, D., Veselovsky, A., Robtsova, M., Grigorenko, V., & Egorov, A. (2019). The impact of long-distance mutations in the Omega-loop conformation in TEM type beta-lactamases. *J Biomol Struc Dyn*, 1-8.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., & et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, 34, 57–65. doi: 10.1002/humu.22225.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., & Guez, A. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354-359.

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, 40, W452–457. doi: 10.1093/nar/gks539.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., & et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*, 314, 268–274. doi: 10.1126/science.1133427.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv [stat.ML]*, arXiv:1206.2944.

Sorin, V., Barash, Y., Konen, E., & Klang, E. (2020). Creating Artifical Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review. *Acad Radiol*.

Spinella, J. F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., & et al. (2016). SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, 17:912. doi: 10.1186/s12864-016-3281-2.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., & et al. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.*, 37, 590–592. doi: 10.1038/ng1571.

Stephens, P., Hunter, C., Bignell, G., Edkins, S., Davies, H., Teague, J., & et al. (2004). Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature*, 431, 525–526. doi: 10.1038/431525b.

Stetz, G., & Verkhivker, G. M. (2015). Dancing through Life: Molecular Dynamics Simulations and Network-Centric Modeling of Allosteric Mechanisms in Hsp70 and Hsp110 Chaperone Proteins. *PLoS One*, 10, e0143752.61.

Stetz, G., & Verkhivker, G. M. (2016). Probing Allosteric Inhibition Mechanisms of the Hsp70 Chaperone Proteins Using Molecular Dynamics SImulations and Analysis of the Residue Interaction Networks. *J. Chem. Inf. Model.*, 1490-1517.

Stetz, G., & Verkhivker, G. M. (2017). Computational analysis of residue interaction networks and coevolutionary relationships in the Hsp70 chaperones: a community-hopping model of allosteric regulation and communication. *PLoS Comput. Biol.*, 13:e1005299. doi: 10.1371/journal.pcbi.1005299.

Stetz, G., Tse, A., & Verkhivker, G. M. (2017). Ensemble-based modeling and rigidity decomposition of allosteric interaction networks and communication pathways in cyclin-dependent kinases: differentiating kinase clients of the Hsp90-Cdc37 chaperone. *PLoS ONE*, 12:e0186089. doi: 10.1371/journal.pone.0186089.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Physchol Rev*, 88(2), 135-170.

Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29, 2238–2244. doi: 10.1093/bioinformatics/btt395.

Thevakumaran, N., Lavoie, H., Critton, D. A., Tebben, A., Marinier, A., Sicheri, F., & Therrien, M. (2015). Crystal Structure of a BRAF Kinase Domain Monomer Explains Basis for Allosteric Regulation. *Nat. Struct. Mol. Biol.*, 22, 37-43.

Thorpe, L. M., Spangle, J. M., Ohlson, C. E., Cheng, H., Roberts, T. M., Cantley, L. C., & Zhao, J. J. (2017). P13K-p110alpha Mediates the Oncogenic Activity Induced by Loss of the Novel Tumor Suppresor P13K-p85alpha. *Proc. Natl. Acad. Sci. U.S.A.*, 114, 7095-7100.

Tokheim, C. J., Papadoulos, N., Kinzler, K. W., Vokelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.*, 113, 14330–14335. doi: 10.1073/pnas.1616440113.

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., & et al. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.*, 76, 3719–3731. doi: 10.1158/0008-5472.CAN-15-3190.

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., & Tawfik, D. S. (2007). The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.*, 369, 1318–1332. doi: 10.1016/j.jmb.2007.03.069.

Toledano, D. T., Fernandez-Gallego, M. P., & Lozano-Diez, A. (2018). Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT. *PLoS One*, 13(10), e0205355.

Tse, A., & Verkhivker, G. M. (2016). Exploring Molecular Mechanisms of Paradoxical Activation in the BRAF Kinase Dimers: Atomistic Simulations of Conformational Dynamics and Modeling of Allosteric Communication Networks and Signaling Pathways. *PLoS One*, 11, e0166583.

Tsuchiya, Y., Taneishi, K., & Yonezawa, Y. (2019). Autoencoder-Based Detection of Dynamic Allostery Triggered by Ligand Binding Based on Molecular Dynamics. *J Chem Inf Model*, 59(9), 4043-4051.

Tvorogov, D., Sundvall, M., Kurppa, K., Hollmen, M., Repo, S., Johnson, M. S., & et al. (2009). Somatic mutations of ErbB4: selective loss-of-function phenotype affecting signal transduction pathways in cancer. *J. Biol. Chem.*, 284, 5582–5591. doi: 10.1074/jbc.M805438200.

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., & et al. (2017). The UCSC genome browser database: 2017 update. *Nucleic Acids Res.*, 45, D626–D634. doi: 10.1093/nar/gkw1134.

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., & Rousseau, F. (2011). A graphical interface for the FoldX forcefield. *Bioinformatics*, 27, 1711–1712. doi: 10.1093/bioinformatics/btr254.

Vega, A., Torres, J., Camaselle-Teijeiro, J., Macia, M., Carracedo, A., & Pulido, R. (2003). A Novel Loss-of-Function Mutation (N48K) in the PTEN Gene in a Spanish Patient with Cowden Disease. *J. Invest. Dermatol*, 121, 1356-1359.

Vijayabaskar, M. S., & Visheshwara, S. (2010). Interaction energy based protein structure networks. *Biophys. J.*, 99, 3704–3715. doi: 10.1016/j.bpj.2010.08.079.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339, 1546–1558. doi: 10.1126/science.1235122.

Vogenberg, R. F., Barash, C. I., & Pursel, M. (2010). Personalized Medicine. *P T*, 35(10): 560-562, 565-567, 576.

Wang, Z., Longo, P. A., Tarrant, M. K., Kim, K., Head, S., Leahy, D. J., & et al. (2011). Mechanistic insights into the activation of oncogenic forms of EGF receptor. *Nat. Struct. Mol. Biol.*, 18, 1388–1393. doi: 10.1038/nsmb.2168.

Wang, Z., Shen, D., Parsons, D. W., Bardelli, A., Sager, J., Szabo, S., & et al. (2004). Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science*, 304, 1164–1166. doi: 10.1126/science.1096096.

Watson, I. R., Takahashi, K., Futreal, P. A., & Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, 14, 703–718. doi: 10.1038/nrg3539.

Weininger, D. (1988). SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences, 31*(6), doi:10.1021/ci00057a005.

Weinstein, J. N., Collisson, E., Mills, G., Shaw, K., Ozenberger, B. A., Ellrott, K., & et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, 45, 1113–1120. doi: 10.1038/ng.2764.

Welling, M., & Kingma, D. P. (2013). Auto-Encoding Variational Bayes. *arXiv [stat.ML]*, https://arxiv.org/abs/1312.6114.

Wildman, S. A., & Crippen, G. M. (1999). Prediction of Physicochemical Parameters by Atomic Contributions 39,. *J. Chem. Inf. Comput. Sci.*, 868– 873 DOI: 10.1021/ci990307.

Wood, D. E., White, J. R., Georgiadis, A., Van Emburgh, B., Parpart-Li, S., Mitchell, J., & et al. (2018). A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.*, 10:eaar7939. doi: 10.1126/scitranslmed.aar7939.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R., & et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, 318, 1108–1113. doi: 10.1126/science.1145720.

Wu, J., Wu, M., Li, L., Liu, Z., Zeng, W., & Jiang, R. (2016). dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. *Database*, 2016:baw024. doi: 10.1093/database/baw024.

Wu, J., Yilmaz, E., Zhang, M., Li, H., & Tan, K. C. (2020). Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition. *Front Neurosci*, 14, 199.

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *arXiv*, arXiv:1609.05473.

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., & Feng, D. D. (2016). DeepGene: An Advanced Cancer Type Classifier Based on Deep Learning and Somatic Point Mutations. *BMC Bioinformatics*, 17, 476.

Zhang, J., Bharan, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., & et al. (2011). International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database*, 2011:bar026. doi: 10.1093/database/bar026.

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery. *Drug Discov. Today*, 22, 1680-1685.

Zhong, G., Gao, W., Liu, Y., Yang, Y., Wang, D. H., & Huang, K. (2020). Generative adversarial networks with decoder-encoder output noises. *Neural Netw*, 127, 19-28.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nat. Methods*, 12, 931-934.

Zhou, W., Ercan, D., Chen, L., Yun, C. H., Li, D., Capalleti, M., & et al. (2009). Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature*, 462, 1070–1074. doi: 10.1038/nature08622.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv [cs.CV]*, arXiv:1703.10593.

Zimmerman, M. I., & Bowman, G. R. (2015). FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J Chem Theory Comput*, 11(12), 5747-5757.

Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R., & Bowman, G. R. (2018). Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities and the Apparent Mechanism of Conformational Changes. *J Chem Theory Comput*, 14(11), 5459-5475.