Computational methods for the analysis and prediction of drug resistance in EGFR-mutated non-small cell lung cancer patients

Preprint · April 2021

DOI: 10.36227/techrxiv.14370317

Rizwan Qureshi
University of Texas MD Anderson Cancer Center

Computational methods for the analysis and prediction of drug resistance in EGFR-mutated non-small cell lung cancer patients

SUBMISSION DATE / POSTED DATE

04-04-2021 / 09-04-2021

DOI

10.36227/techrxiv.14370317.v1

Computational methods for the analysis and prediction of drug resistance in EGFR-mutated non-small cell lung cancer patients

Rizwan Qureshi[1], Bin Zou[1], Victor H.F. Lee[3], and Hong Yan[1]
[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, [3]Department of Clinical Oncology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, China
Email: engr.rizwanqureshi786@gmail.com, h.yan@cityu.edu.hk

Abstract—**Lung cancer is the major cause of cancer deaths worldwide, and has a very low survival rate. Non-small cell lung cancer (NSCLC) is the largest subset of lung cancers, which accounts for about 85% of all the cases. It has been well established that the mutation in the epidermal growth factor receptor (EGFR) can lead to lung cancer. EGFR tyrosine kinase inhibitors (TKIs) are developed to target the kinase domain of EGFR. These TKIs produce promising results at the initial stage of therapy, but the efficacy becomes limited due to the development of drug resistance. In this paper, we provide a comprehensive review of computational methods, for understanding drug resistance mechanism. The important EGFR mutants and different generations of EGFR–TKIs, with the survival and response rates are discussed. Next, we evaluate the role of important parameters in drug resistance, including structural dynamics, hydrogen bonds, stability, binding free energies, and signaling pathways. Personalized drug resistance prediction models and the use of deep learning and big data analytics in drug design, and drug resistance analysis are also highlighted. In addition, we discuss limitations in the current methodologies, and explore potential research avenues for the development of future therapies, using computational methods. We hope, that this review will serve as a reference for early and advanced stage researchers, to apply computational techniques for drug resistance analysis and prediction in lung cancer patients.**

Index Terms—**Non-small cell lung cancer (NSCLC), Epidermal growth factor receptor (EGFR), Molecular modeling, Computational methods, Molecular dynamics (MD) simulation, Deep Learning**

I. Backgroud

Cancer is the second largest cause of deaths worldwide, resulting in a loss of 9.6 million lives in 2018 [1]. About 13% of all the new cancer cases are lung cancer and more than 65% of them are diagnosed at advanced cancer stages.

The over-expression of epidermal growth factor receptor (EGFR) is found in about 60% of non-small cell lung carcinomas (NSCLCs) [2]. EGFR Tyrosine kinase inhibitors (TKIs) are developed to target the kinase domain of EGFR. These inhibitors produce promising results at the initial stage of therapy, but the drug resistance develops in most of the cases after about a year [2]. Drug resistance is currently one of the biggest challenges in targeted cancer therapy.

The discovery of small molecule inhibitors targeting the kinase EGFR has generated much hope for the further advances in the treatment of lung cancer, but the limitations of their efficacy are apparent [3]. The upregulation and engagement of efflux pumps to remove the drug from the binding site is one of the primary reason for the acquired resistance [4]. Another reason is the genomic variations in the kinase domain of EGFR, such as T790M or C797S mutations. Other mechanisms include c-Met amplification, activation of alternate signaling pathways, and the co-activation of multiple receptor tyrosine kinases that can reduce the dependence of tumor cells on EGFR-mediated signaling [5].

One of the major causes of drug resistance to the 1st and 2nd generation drugs is a secondary acquired gatekeeper T790M mutation. The 3rd generation drug Osimertinib is designed specially to inhibit the T790M mutation through covalent binding of CYS 797 [6], but the efficacy of Osimertinib is lost upon the emergence of C797S mutation (Figure 1). The 4th generation drug EAI045 appeared in 2016, targets both T790M and C797S giving hope to NSCLC patients. Table I presents different generations of EGFR inhibitors with response and progression free survival rate.

EGFR mutations usually increase the activity of kinase domain, resulting in uncontrolled cell division that can eventually lead to lung cancer. The COSMIC database [7] is the largest open access database for EGFR mutations, which shows that about 93% of EGFR mutations are found in the four exons 19 – 21 of the kinase domain. These mutations include inframe deletion in exon 19, insertion in exon 20, and a single point substitution in exon 21. Nucelotide substitution in exon 18 (G719S or G719C) accounts for about 5% of the EGFR mutations [8]. The statistics of different EGFR mutations is shown in Figure 1.

Experiment methods are expensive and time consuming, due to the necessity of multiple experiment conditions. Computational methods are also applied to quantitatively investigate the drug resistance process, virtual screening of compound, visualize the drug-ligand interactions [9] and the identification of drug binding sites [10].

Although, there are reviews on EGFR-mutated drug resistance in NSCLC [18]–[20], these studies mainly focus on clinical data and random trials. The methods using CT scan lung cancer images are also covered in [21]. The Nanotechnology based intelligent cancer drug design is discussed

TABLE I: EGFR Tyrosine Kinase Inhibitors with response and survival rates

Reference	Generation	Drug	Mutation	PFS (Months)	ORR
IPASS [11]	1st	Gefitinib	EGFR del 19, L858R	9.5	17.2%
EURTAC [12]	1st	Erlotinib	EGFR del 19, L858R	9.7	64%
Lux – Lung [13]	2nd	Afatinib	EGFR del 19, L858R, rare mutations, HER2/4	11	66.9%
NA	2nd	Neratinib	EGFR G719X, HER2/4	10	NA
FLAURA [14]	3rd	Osimertinib	mEGFR, T790M	18.9	80
AURA2 [15]	3rd	Osimertinib	mEGFR	8.6	71%
AURA1/2	3rd	Olmutinib	mEGFR/T790M	7	56%
ARCTIC	3rd	durvalumab	mEGFR	NA	NA
TATTON [16]	3rd	Osimertinib	MET, BRAF, C797S	NA	NA
TIGER-X [17]	3rd	Rociletinib	T790M	9.6	45%
NA	4th	EAI045	T790M/C797S	NA	NA

in [22]. This article presents a comprehensive review of computational studies on EGFR-mutated NSCLC patients and personalized drug resistance prediction models. The paper is organized as follows, we discuss computational methods for drug resistance analysis in Section 2. Personalized drug resistance prediction models are described in Section 3. The use of deep learning and big data analytics is highlighted in Section 4. The challenges and opportunities in the current methodologies are discussed in Section 5. Finally, we conclude this review with potential future research directions in Section 6. A list of acronyms with definitions, used in this paper is given in supplementary Table 1.

II. Computational Methods for EGFR-mutated Drug Resistance

Computational methods are widely used for the drug resistance analysis, due to its flexibility, low cost, easy implementation, and the ability to process a large amount of data [23]. Computational methods can provide deeper insights, generate novel hypothesis and devise new promising strategies. These methods can broadly be classified into structure based and ligand based. In this paper, we mainly focus on structure based drug design and analysis methods.

A. Molecular Dynamics Simulations

Molecular dynamics (MD) simulation is a powerful computational method for analyzing the interactions between the drug and the target at atomic scale [24]. MD simulation enables to examine the structural changes occurred due to genetic mutations, and it is widely used for drug resistance analysis, prediction and discovery [25]. The interactions between drug and protein is a dynamic process, involving several residues. MD simulations can reveal the atomic level details of drug-protein interactions.

Since its beginning, MD simulations have advanced from simulating several hundred atoms to systems with biological relevance, including entire protein with solvent, protein with membrane embedded, and large macromolecular complexes like nucleosomes [26] or ribosomes [27]. This immense improvement is due to the high performance computing (HPC) and basics of MD simulation algorithm [28]. The trajectories of positions, accelerations and velocities of each atom can be obtained using Newtons's second law of motion. The MD simulation generates a large amount of dynamics data, which can be used to understand the structure to function relationship.

Shan *et al.* [29] used long time scale MD simulation and showed that N-lobe dimerization of the wildtype (WT) EGFR is disordered, and it becomes ordered only upon dimerization. They also showed that some cancer specific mutations L858R/L834R may facilitate the dimerization by suppressing this local disorder. The L858R causes abnormally high kinase activity by promoting EGFR dimerization. They further performed unbiased, all-atom MD simulations of EGFR kinase domain [30], and showed that EGFR monomer is more stable in its inactive state than its active state.

Shunzhou *et al.* [31] used 10 micro-second long MD simulation for the investigation of conformational dynamics and interactions. The simulation result shows that the mutant type L858R binds to Gefitinib, rather than ATP. They used the MD simulation of [32] and applied PCA on the MD trajectories. The number of correct bound association were compared, which showed the preference of WT EGFR binding with ATP and L858R with Gefitinib.

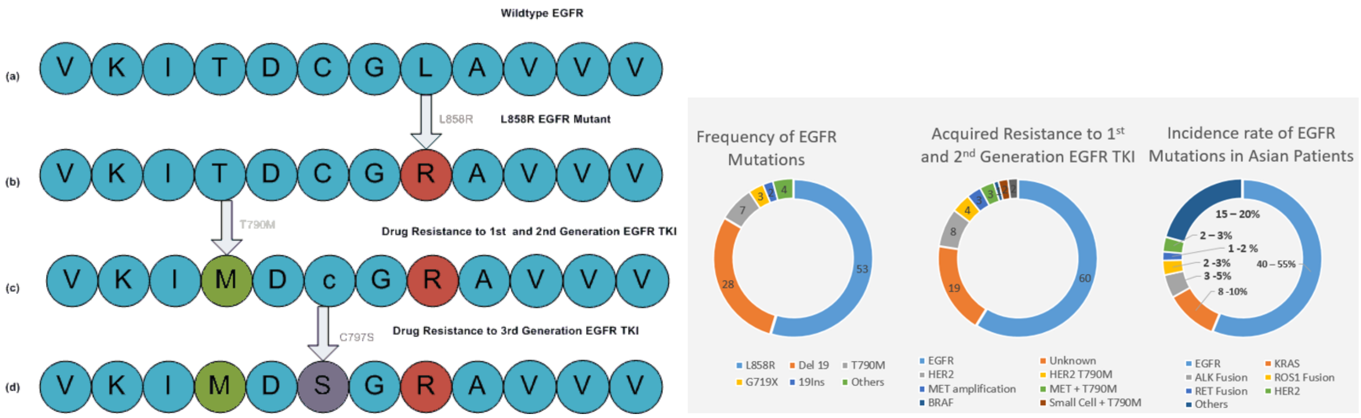

Fig. 1: EGFR frequent mutations (right). L85R drug sensitive mutant, T790M drug resistive mutant to the 1st and 2nd generation drugs, and C79S drug resistive mutant to 3rd generation drug.
The frequency of EGFR mutations, acquired resistance and mutation rate in Asian patients [33], [34]. (a) L858R is the most common type of EGFR mutation. (b) The EGFR mutation are the most common mechanism associated with the drug resistance to 1st and 2nd generation of EGFR TKI. (c) Mutation rates in Asian patients.

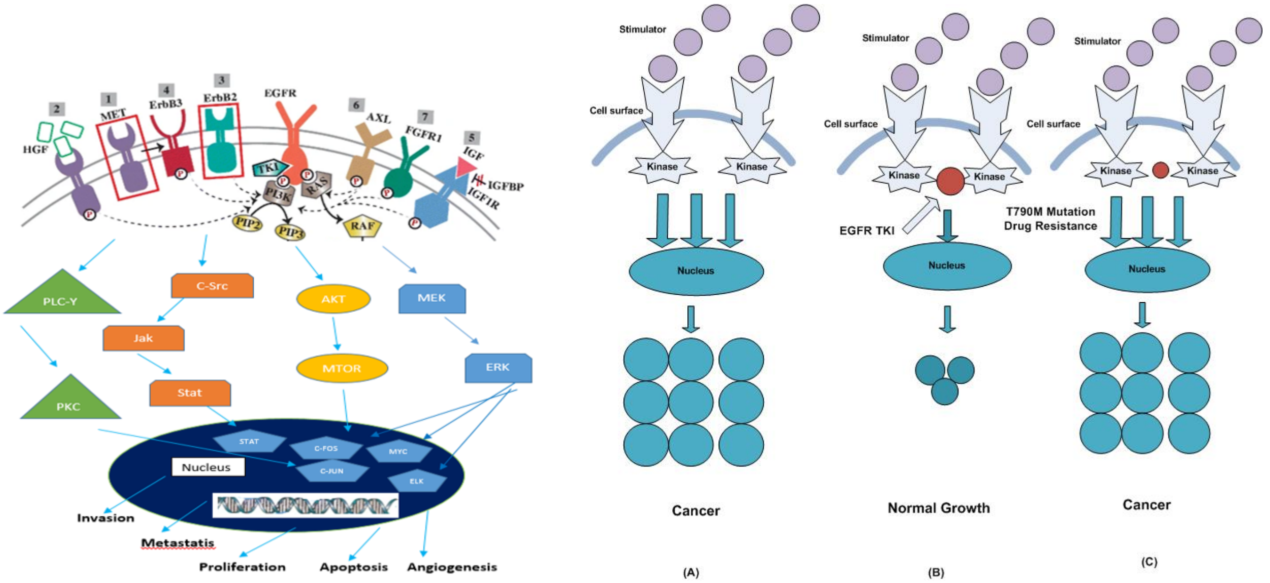

Fig. 2: EGFR RTK Dimerization and downstream signaling mechanism (left) [35].
EGFR TKI Binding mechanism, the WT EGFR shows the cancer growth, L858R mutant binds with the TKI, due to T790m mutation TKI is unable to bind the kinase (right most).

B. Binding Free Energy EGFR

The energy released due to bond formation, ligand and protein interactions is known as the binding free energy. The free energy of a favourable reaction is negative. The energetic contribution of individual residues provide useful quantitative information about the binding mode of a ligand and the protein [36]. Wang *et al.* [37] investigated the structural and energetic features of both active and inactive states of WT EGFR and L858R mutant using the molecular mechanics Generalized Born solvent accessible surface area MMGBSA [38] tool in Amber [39]. They further analyzed how the mutations affects the stability of several conformational states. They mapped the free energy landscape on to the principal components to identify population changes in various conformations on mutations. The study reveals that the L858R mutant induces conformational changes in active and inactive states, which affect the relative stability.

In another study, Wang *et al.* [40] profiled the correlation between the EGFR mutations and EGFR-TKI Afatinib (second generation drug). The progression free survival rate and drug-response level was recorded. The L858R and the complex mutation L858R - T790M showed a response level of 3, 2 and a PFS rate of 15.87 and 9.59 months respectively, with the lower values showing low response. The WT and the single T790M mutation showed no response and no survival. Results of this study verifies the higher potency of Afatanib to classical EGFR mutation L858R and exon 19 deletion, and a lower one to T790M mutation, which are consistent with the clinical values.

C. Molecular Docking

Molecular docking is another important tool in structure-based drug design, due to its low-cost, and simplicity [41]. In molecular docking, the position and orientation of a molecule are predicted against the other molecule, when they are bound to each other to form a complex. The prediction can be useful for estimating the strength of binding affinity.

In [42], molecular docking and MD simulation are used for the investigation of structural and chemical features of EGFR inhibitors and binding pocket mechanism. They found that the binding pocket consists of three regions (P1, P2, P3). The Met 793 and ASP 855 may be the reason for the binding recognition through H-bond interactions with Phe 856.

Rajith *et al.* [43] investigated the G719S-T790M double mutation using 50-ns MD simulation and molecular docking. They observed the escalation in distance between P-loop and functional loop in T790M mutation compared with the G719S. They also verified that G719S mutant causes the ligand and hinge region to come closer and T790M mutant caused the ligand to escape from the binding pocket. This may be the reason for the aberrant function of EGFR-TKIs in T790M mutant.

D. Computational Modeling of EGFR Mutations

Structural information for only a few mutants is available, and a variety of rare EGFR mutations account for about 10 – 20% of NSCLC patients. It is very difficult to treat patients harboring such rare mutations with EGFR TKIs. Robust prediction models are needed for such rare EGFR mutants to existing EGFR TKIs [44].

Starting from the structures found in the Protein Data Bank (PDB) [45] (www.rcsb.org), Ma *et al.* [46] created a 3D structure database of EGFR mutants with binding free energies of 112 EGFR mutation types collected from 942 NSCLC patients, using computational modeling and MD simulations. For residue substitution mutants, the Rosetta ddg - monomer protocol was employed. The side-chain of the mutated residue is first replaced and the rotamers of all residues are then optimized by the Rosetta's standard side-chain optimization module. For the mutations of amino acids, Rosetta comparative (CM) protocol was applied. The framework for building the EGFR mutant structure database is shown in Figure 3. The predicted models may not be accurate, and several computational tools, including Q-MEAN Z-score [47], Verify3D [48], and Ramachandaran plot [49] can be used to validate the predicted models. It will be interesting to compare the computationally predicted mutants with multiple methods, to increase the confidence level [50].

E. EGFR Dimerization and Signaling Pathways

When the extracellular ligand-binding domain is activated by its growth factor, a homo-dimer or a hetero-dimer is formed with another member of the ErbB family [51]. EGFR dimerization is an essential event in the EGF-signal transduction [52]. The dimerization stimulates the catalytic activity of tyrosine kinase domain, and promotes the autophosphorylation of several residues in the kinase domain [53]. These residues provide docking sites for downstream signaling molecules such as Shc, Grb2 and P13k. It is also shown that the EGFRs can form dimer on the cell surface independent of ligand binding [54]. Tumors, that are sensitive to EGFR TKIs are characterized by a rapid decrease in the Akt activity [55]. The failure to irregulate Akt causes drug insensitivity [56].

Wang *et al.* [35] investigated the contribution of EGFR and ErbB-3 heterodimerization based on binding free energy. The EGFR dimerization and downstream signaling mechanism is shown in Figure 2 (left). They characterized the EGFR mutations for 168 clinical subjects using molecular interactions for three EGFR dimers (ErbB-2, IGF-1R, c-Met) and two EGFR inhibitors, Gefitinib and Erlotinib. Simulation results show that the mutant – partner interactions increased in the L858R and L858R - T790M. The mutant delL747 - P753insS has the largest difference between the mutant interactions and inhibitors, and this may be the reason for the shorter progression free survival in this mutant type. They also investigated ErbB-3 and its interactions with EGFR mutants, IGF-1R, ErbB-2 and c-Met and found that the binding was remarkably strong between ErbB-3 and c-Met, which shows its role in regulating ErbB-3 signaling.

F. Long Range Communication Capability in EGFR

Allosteric Communication is an important phenomenon in many biological processes and considered to be a useful parameter in governing molecular motion and signal transduction [57], [58]. Recent studies [59], [60] show that allosteric networks of cooperative protein motion may be formed by sparsely connected group of residues.

In [57], Dixit and Verkhivker used (MD) simulation, PCA and signal propagation in protein to identify the allosteric communication in ABL and EGFR kinase domain with cancer mutations. They used the concept of absolute and relative long-range communication capability (LRCC) in residues for tracing the signal propagation in proteins. According to their algorithm, two remote protein residues (residue clusters) have strong communication if the mean square fluctuation of inter-residues remain in a specific threshold over long time MD simulations.

The efficient long-range communication is possible not only because of the thermal fluctuations, but also a dynamic long-range interaction exists between the regions that are important in coordinating inter-lobe and inter domain motion.

It has been shown in [31] that free energy landscape is populated by conformational isomers, and extended sampling of the landscape indicates the flexibility of EGFR. First two principal components are used to describe the system dynamics [31]. Simulation results show that L858R has higher binding preference to Gefitinib than the ATP.

G. Hydrogen Bond Analysis

Hydrogen bonds are important for the analysis of biological and chemical interactions of molecules. Ghosh and Yan [61]

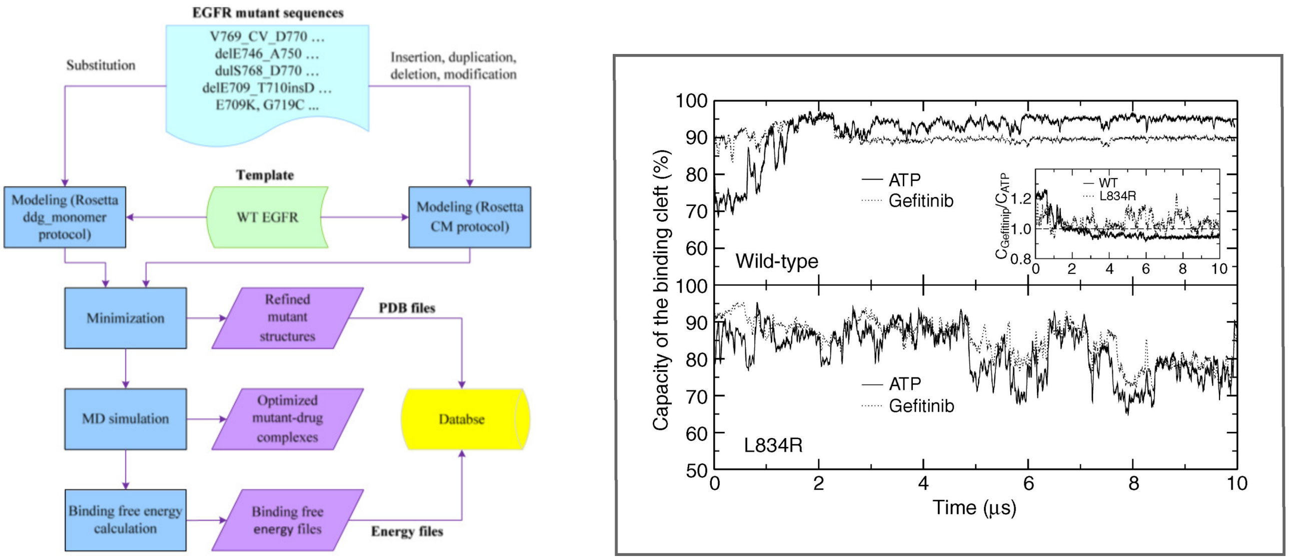

Fig. 3: The framework for predicting the mutant structure of EGFR and their binding energies [46]. The Binding preference for ATP in WT EGFR and Gefitinib L858R mutant [31]

investigated the EGFR and ErbB-3 heterodimers and their mutant structures. They performed 3-ns MD simulation of three EGFR dimers (WT, L858R, and L858R-T790M). They calculated the hydrogen bonds and found that the number of hydrogen bond changes throughout the structure. The mean value was 481 for WT, 484 for L858R and 477 in L858R - T790M structure. They concluded that the T790m mutant structure have a smaller number of hydrogen bonds, which changes the conformational stability of the system.

H. Drug Response Curves

The drug / dose response curve shows the response of an organism as a function of exposure to a drug after certain time [62]. The drug′*s* half maximum inhibitory concentration IC_{50} [63] of cancer cell viability is widely used to measure the potency. Another commonly used parameter is the half maximal effective concentration EC_{50}, which measures the concentration of the drug, which induces a response halfway between the baseline and maximum, after a specified exposure time [64]. Computational methods, including support vector machine (SVM), neural networks (NN) and quantitative structure activity representation (QSAR) are widely used to determine the drug sensitivity and design [65]. Jang *et al.* [66] carried out a systematic assessment of analytical methods for drug sensitivity prediction using CCLE data. They evaluated more than 110,00 models, based on multifactorial experiment design. They found that the model input data, including molecule features and compound are important factor for model performance, followed by the type of algorithms.

In another study, Zou *et al.* [67] analyzed the 30 most common EGFR mutants. They used MD simulation, and protein-ligand interactions footprint (IFP) to analyze the binding modes of EGFR-Gefitinib complexes [68]. Multilinear Principal Component Analysis (MPCA) was used for dimensionality reduction and feature selection. Target projection pursuit [69] was used to show drug sensitivities. The findings show that the IFP features of EGFR-mutant complexes and MPCA based tensor are useful candidates for prediction of drug sensitivity. They used five classifiers for predicting the drug sensitivity, and achieved greater than 90% accuracy.

I. PCA-3D QSAR

Quantitative structural activity representation (QSAR) is applied for establishing a relationship between chemical properties of a compound and their biological activities [70]. Principal component analysis (PCA) is a method to determine the essential dynamics of a protein. It is used to express the dominant motions in the protein system or a MD trajectory. PCA can be used for reducing the dimension of QSAR data matrix [71].

Sukrita *et al.* [72] performed MD simulations and applied PCA to WT EGFR and mutant T854A. 3-D QSAR model was built using the step forward multiple regression and advanced variable selection. The proposed model was validated using the statistical parameters.

PCA was used to simplify the motion and extract the important component of motion. For evaluation purpose, co-variance matrix was created from all the trajectories and diagonalized to identify set of Eigenvalues and Eigenvectors that correspond to displacement of atom and show the concentrated motion.

The simulation results show higher flexibility in mutant T854A compared to WT. Another parameter known as radius of gyration Rg was calculated and results show higher values of Rg for T854A mutant as well. The WT structure become flexible upon T854A mutation and losses stability in RMSD, RMSF, and Rg.

Computational methods [29], [30], [57], [72], [73] provide useful information about the structure and dynamics of EGFR mutants. The L858R genetic mutations in the kinase domain suppress the local disorder, and provides a series of conformations states, capable of binding Gefitinib. This provides an explanation for effectiveness of Gefitinib in L858R EGFR mutated NSCLC patients. While, the T790M mutant increases the ATP affinity and the escalation in the p-loop and long range communication capability of residues are some of the factors for the failure of the EGFR TKIs in this mutant. The loss of hydrogen bond between threonine and arginine in T790M mutant is also a reason for the drug resistance.

T854A residue is located at the bottom of ATP binding site on C–lobe and in contact with Erlotinib and Gefitinib. The substitution T854A results in the loss of contact and binding affinity to the inhibitors. The loss of stability in T854A and increase in the hinge region further explains the aberrant function. In the docking analysis, the G719S mutant causes the ligand and hinge region to come closer, and T790M mutant causes the ligand to escape from the binding pocket. In Table II, we list several important computational studies on EGFR mutants related to drug resistance.

III. Personalized Drug Resistance prediction

Personalized medicine is a growing field of healthcare. It is an individual treatment approach based on the patient's unique clinical, genetic, epigenetic, and environmental information [85]. Disease are heterogeneous, and the ultimate objective of the personalized therapy is to define the disease at molecule level, so that therapeutic agents are targeted towards right population of the people [86], [87], [87]. Keeping this in view, Wang *et al.* [79] used binding energy of EGFR mutant complex and personal features (age, sex, smoking history, medical history) to build a personalized drug resistance prediction model. They used extreme learning machines (ELMS) [88] to predict the drug resistance level. EGFR-TKI interaction pattern and personal features are used in the prediction model, and an overall accuracy of 95% was achieved. The accuracy of the models is significantly increased with the addition of personal features.

Kureshi *et al.* [36] established the relationship between patient's personal characteristic and tumor response level. The drug response is associated with EGFR mutation status type and personal features. They applied four classifiers to predict the outcome of EGFR TKI, and achieved an overall accuracy of 76.54% with area under the curve (AUC) equal to 0.76. They showed that the SVM and the decision trees are potential candidates for personalized drug resistance prediction.

1) Geometric Analysis of Drug Binding Site: Analysis of geometrical shape at the drug binding site can also reveal interesting insights about the drug resistance mechanism and structure based drug design (SBDD). The geometrical properties of the protein-drug complex can also be a useful feature for drug response prediction [89]. The concave shape has a higher drug molecule affinity than the convex. Low convex degree shapes bind tightly than the high convex degree shapes.

Ma *et al.* [80] analyzed the properties of EGFR mutants using structural information. They computed the alpha shape model [90] and solid angle [91] to evaluate the properties of the atom at the binding sites.

MD simulations were performed using Amber [92] suite, and the (CGAL) [93] library is employed to compute the shapes of the EGFR mutants. They normalized the curvature value to [-1, 1], with values falling in the range of [-1, 0] is defined as concave and for [0, 1], the shape is defined as convex. To simplify the problem, they calculated the average convex degree at the drug binding site and obtained the average knob level of mutant by averaging the convex degrees for 200 frames. It is shown that the 90% of the mutant can be grouped together by the knob threshold. To validate the model, they compared the results with the clinical data, to be specific the L858R - T790M mutant showed no drug response which is consistent with our knowledge.

These studies provide useful reference for the personalized drug design for EGFR-mutated NSCLC patients. The simulation results show that the accuracy of the prediction model is significantly increased by adding the personal features. The geometrical features, drug binding site distance, energy and personal features may be combined to construct an efficient personalized drug resistance prediction model.

IV. Deep Learning

Deep learning has shown remarkable results in solving many challenging problems of computer vision, natural language processing, financial model analysis, and bioinformatics [94], [95]. The use of deep learning has also shown promise in pharmaceutical research such as bio-activity prediction, and drug discovery. The convolutional neural network (CNN) has provided excellent results in image recognition problems [96], CNN can also be used to design protein-ligand scoring functions. Recently, a new antibiotics have been discovered from the pool of 100 million molecules using deep learning [97]

A deep learning model is proposed to predict the mutation status form CT-scan lung images in [98]. The proposed model predicts the probability of a tumor being EGFR mutant using the CT-scan tumor image as an input. The model consists of two sub-networks, sub-network 1 shares the same structure as first 20 layers of Densenet [99] using transfer learning [100], and sub-network 2 was trained on the EGFR mutation dataset. The proposed model achieved high accuracy (AUC 0.85, 95% CI 0.83–0.88) and also revealed an association between high dimensional CT-scan images and EGFR genotype.

Deep learning can also be used for virtual screening, and to design novel EGFR inhibitors [101]. Recently, a multi-input deep neural network based on attention mechanism [102] is proposed for biological activity prediction of EGFR inhibitors [82]. The proposed model uses the dataset [103] of 3492 compounds labeled as inhibitors or non-inhibitors. The method uses SMILES (Simplified molecule input line entry system [104] which is a way to represent compound molecular structures in *in silico* study, as an input to the CNN. The CNN generates a 120 dimensional vector, which generates the attention map. The proposed model achieves state of the art accuracy with AUC equal to 90% by cross-validation. Another contribution of this model is the integration of attention layer, which may explain the contribution of each atom on the overall biological activity.

The drug response prediction is related to precision medicine. A 1-dimensional CNN (1D CNN) DeepIC50 [83]

TABLE II: Important computational studies related to EGFR mutation induced drug resistance

Reference	Mutant Sequences	Algorithms	Applications
[2]	EGFR mutants	Review	Review
[31]	WT and L858R	MD and PCA	Drug efficacy to Gefitinib
[35]	EGFR, ErbB-3, IGF, cMet	MD and structural analysis	Contribution of hetrodimers in drug resistance
[36]	Clinical data	pattern minning and univariate analysis	Personalized predictive models
[43]	WT, T790M, G719S, and T790M-G719S	Molecular dynamics and docking analysis	Structural investigation
[44]	Rare mutations & clinical data of 3779 patients	MD & Statistical analysis	EGFR TKI sensitivity prediction
[46]	112 EGFR mutants	MD & Clinical data of 942 patients	Binding energies
[61]	EGFR ErbB-3 hetrodimers	Hydrogen bonds	Stability analysis
[67]	30 EGFR mutants	Interaction footptint matrix and PCA	EGFR-TKI sensitivity prediction
[74]	WT, L858R, T790M	computational methods	Affinity for Gefitinib and ATP
[72]	WT & T854A	MD & PCA	Structural investigation
[75]	L858R, T790M, G719C and L861Q	MD &NMA	Impact of point mutants to Geftinib
[76]	WT, L858R, L858R-T790M	Parametric methods	EGFR domain's analysis
[77]	HER2, BRAF and EGFR	Molecular modeling	effect of in frame deletion
[78]	WT and T790M	Attribute ranking	Mechanism of T70M mutation
[79]	32 EGFR mutants	MD and extreme learning machines (Elm	Personlized drug resistance prediction
[80]	EGFR and 30 mutants	Alpha shape dynamics	Drug resistance prediction
[81]	EGFR extracellular domain	Energetics and PSN	Ligand binding effects
[82]	3492 Compounds of EGFR	Deep Learning attention mechanism	EGFR drug discovery
[83]	CCLE data	DeepIC$_{50}$	Drug sensitivity prediction
[84]	110 mutations	Machine learning and time-series	Mutation impact prediction

is proposed to predict the three-class drug response. The model used genomics profile and drug molecular features from massive drug data on cancer cell lines. The proposed model achieved better accuracy than the baseline methods. Training a deep neural network with a small number of points is a very active research topic in deep learning. One example is the one shot learning [105]. Such a network can be exploited for drug resistance prediction model with small number of clinical samples.

A. *Big Data Analytics*

Big data analysis and Artificial Intelligence are changing the drug discovery pipeline. The Cancer Therapeutics Response Portal (CTRP) [106] provides a resource to develop new insights into small molecule action mechanism, generate new hypothesis, and personalized drug discovery based on predictive bio-markers. Several big data projects, including Cancer Cell Line Encyclopedia (CCLE) [107], and Genomics of Drug Sensitivity in Cancer (GDSC) [108] have performed large scale molecule screens on panels of hundreds of molecularly characterized cancer cell lines. These projects demonstrates the potential of modern machine learning algorithms to develop drug response predictors based on molecular profiling measurements [97].

However, it is important to acknowledge the challenges of big data analytics. The current cancer data resources are not enough for providing an adequate answer to drug resistance and response. In fact, independent procedures analyzing the clinical data may reach different conclusions, aiming to answer the same biological question [109]. Another problem is the inconsistency between datasets and missing clinical information [110]. One solution to missing values is to use data imputation techniques [111].

V. Challenges and Opportunities in the Current Methodologies

Most of the studies discussed in this review are based on MD simulations. MD simulations are a valuable tool in structure-based drug design (SBDD), and play a vital role in understanding the molecular interactions, and conformational changes. However, it is important to acknowledge the limitations of MD simulations, including time limitation, force - field inadequacy and quantum effects [112]. The enormous amount of computational power has made it possible to carry out MD simulations for several microseconds for systems containing hundreds of millions of atoms, yet the time resolution may not be enough for relaxing the system to certain quantities.

Moreover, several biological properties such as protein folding, ligand binding and unbinding may occur at longer time scales. The issue of selecting the force fields remain a significant challenge and MD simulation results are reliable only if the force fields mimic the same force as experienced by the real-world systems. It is important to mention that force field files are parameterized, and they describe varied situation of the same atom-type.

Classical MD-simulations cannot model the chemical reaction of drug substrate and the bonding process of certain covalently bonded ligands. In such cases, quantum mechanics becomes a viable option, which models the system at the electron level. Nevertheless, they require more computational power than the classical method. The reactive force fields are developed to model the chemical activities.

Another challenge faced by the MD simulations is electronic polarization, and quantum effect. The bio-molecules are polarizable, and the electron clouds around the atom constantly changes with the chemical environment. In such a case, we can use quantum mechanics (QM) MD. QM-MD are computationally expensive and limited to small number of atoms. Computationally efficient QM approaches are needed to model large protein-ligand systems at the electron-level.

The binding free energies calculated are also not accurate, and there are errors reported around 1k/cal on average in applicable scenarios [109]. However, there is still merit in calculating binding free energies, as they allow one to distinguish between weak and tight binding. In the personalized models, the data dimensionality remains a significant challenge. Most cancer studies of anticancer drug response have small sample size (less than 100 patients) compared to variable size (human genes $>$ 20,000). Computational methods might not produce robust results, and non-algorithmic solutions are necessary. The high dimensionality problem can be addressed by using feature filtering techniques or sparse principal component analysis [113]. Another solution is the data integration [114]. Integrating all studies together may increase the confidence in results.

Deep neural networks are considered black boxes outside the machine learning community, and often domain experts are needed to interpret the model output [115]. Since most of the studies are connected to patient's health, the logical reasoning of the model will provide useful information [116]. Furthermore, transforming the deep learning black-box to a white-box is an active research topic. Different methods have been developed to interpret the model including back-propagation, exaggeration of hidden representations, and activation maximization [117].

VI. Conclusion

In this paper, we explained the drug resistance mechanism in EGFR-mutated NSCLC patients. We discussed several important EGFR mutants and their interactions with EGFR-TKIs. The mutation changes the stability, binding free energy, dynamics and structure of EGFR. In our opinion, the stability is one of the most crucial factor for analyzing the drug response. The stability is the change in the net energy in the unfolded and folded states [118]. It will be interesting to see whether a state-transition matrix can be propose to model the changing states of a protein [119]. Computational methods have shown promise in analyzing the EGFR properties, and produced useful insights about drug resistance mechanisms.

There are various reasons for the failure of EGFR-TKIs. The increase in the ATP affinity, flexibility, loss of stability and the loss of hydrogen bonds at the site 790 are some of the reasons for the T790M mutation. However, some mutants are drug sensitive. The L858R mutant prefers binding to Gefitinib than ATP. The Del 19 mutant has a high drug sensitivity, and L858R and L861Q have moderate, and T790M and T790M-L858R have low drug sensitivity. The personal features also play an important role, and it is observed that the drug sensitivity also depends upon the personal characteristics. We believe that deep understanding of personalized models may result in age specific or gender specific treatment plans.

A variety of rare mutations occur in about 10 – 20% of the NSCLC patients, due to higher diversity, proper medication for such mutants is difficult in the daily clinic. A little is known about the effect of these mutations on downstream signaling and interactome [120]. Robust prediction models are needed to predict the sensitivities to such mutants. Predicting the impact of mutations on protein-ligand affinity, structural changes is of great importance. Deep learning architectures based on time-series features can be used to predict the impact of mutations, which can help understand the mutation induced drug-resistance mechanism.

Diagnosing the EGFR-mutant using CT-scan images and deep learning provides non-invasive and an easy solution [121]. By using feature visualization or other method, interesting insights can be obtained. Despite of all these studies and findings, there are still unknown reasons of drug resistance and further studies are required to investigate, and validate the findings. The rise of deep learning, and the enormous amount of digital data, and large computational resources can provide efficient pipeline to improve drug discovery, understand the drug resistance process and provide optimal decision making in treating EGFR-mutated NSCLC patients.

A. *Future Work*

More clinical data is required to refine the prediction models and deep neural networks can be exploited to increase the prediction accuracy. The interpretation of deep neural networks may produce useful insights about the drug resistance mechanisms. The small dataset problem may be addressed by using the matching networks [122]. Moreover, data augmentation can be used to create virtual clinical samples. Combination of long MD simulations, and usage of recently developed tools, such as, DeepMD [123] with large clinical data may pave the way for further studies.

The high arithmetic and intrinsic parallelism of graphical processing units (GPUs), tensor processing units (TPUs) can be exploited to perform longer MD simulations. ACEMD is a

MD simulation program, capable of providing supercomputer level performance of 40-ns/ day for protein systems with more than 23,000 atoms [124]. AceCloud [125] is another cloud-based MD program, capable of running hundreds of MD simulations. Moreover, the development of efficient open source libraries for the analysis of molecular dynamics trajectories, such as MDAnalysis [126], MDTraj [127], and online web applications [128] provides opportunity for flexible and fast framework for complex analysis programs. In addition, online freely drug-databases, such as drug-bank [129] can be exploited for future drug design. Another important research direction is to design decision support systems, based on clinical features. Such systems can help doctors in devising optimal treatment strategies.

We believe that data analysis methods will play a vital role in future cancer research [130]. Several Bio-tech companies are using artificial intelligence to enhance the drug development process, from candidate screening to trial management [131]. We speculate that decades from now, personalized drug models will be used in treatment of NSCLC patients. We further believe that the combination of computational methods and clinical studies can provide useful recommendations, for precision and personalized medicine.

VII. Acknowledgement

This work is supported by the Hong Kong Research Grants Council (CityU 11200818 and C1007-15G) and City University of Hong Kong (Projects 9610034).

VIII. Author Biographies

Rizwan Qureshi is working towards his Ph.D. degree in Electrical Engineering at City University of Hong Kong. His research interests include bioinformatics, computational biology, signal and image processing and machine learning.

Bin Zou received PhD degree in Electronic Engineering from City University of Hong Kong. His research interests include bioinformatics, computational biology and pattern recognition.

Dr. Victor Lee is currently Clinical Assistant Professor of the Department of Clinical Oncology, Li Ka Shing Faculty of Medicine, The University of Hong Kong. He graduated in the University of Hong Kong in 2002. He obtained the fellowship in Royal College of Radiologists in Clinical Oncology in 2007. Then he joined the Department of Clinical Oncology, Li Ka Shing Faculty of Medicine, The University of Hong Kong as Clinical Assistant Professor in 2008. He became a specialist in clinical oncology in 2010. His research interests include novel radiation therapy, head and neck cancer, lung cancer, liver cancer, colorectal cancer and brain tumours, as well as genetic and molecular studies on nasopharyngeal cancer and lung cancer

Hong Yan received his PhD degree from Yale University. He was Professor of Imaging Science at the University of Sydney and is currently Chair Professor of Computer Engineering at City University of Hong Kong. His research interests include image processing, pattern recognition and bioinformatics, and he has over 600 journal and conference publications in these areas. Professor Yan is an IEEE Fellow and IAPR Fellow, and he received the 2016 Norbert Wiener Award from the IEEE SMC Society for contributions to image and biomolecular pattern recognition techniques.

IX. Highlights

- Computational methods for drug resistance analysis in EGFR-mutated NSCLC patient are discussed in detail.
- Personalized drug resistance prediction models are also evaluated.
- Limitations in the current computational methodologies and the way forward are also highlighted.
- Properties of EGFR mutant structures, including stability, binding free energy and dynamics and structure are critically evaluated.
- This article provides useful information for understanding the drug resistance mechanism in EGFR-mutated NSCLC patients, using computational methods and sheds some light on the future.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] S. V. Sharma, D. W. Bell, J. Settleman, *et al.*, "Epidermal growth factor receptor mutations in lung cancer," *Nature Reviews Cancer*, vol. 7, no. 3, p. 169, 2007.

[3] D. J. Leahy, "A molecular view of anti-erbb monoclonal antibody therapy," *Cancer cell*, vol. 13, no. 4, pp. 291–293, 2008.

[4] J. Li, G. Cusatis, J. Brahmer, A. Sparreboom, R. W. Robey, S. E. Bates, M. Hidalgo, and S. Baker, "Association of variant abcg2 and the pharmacokinetics of epidermal growth factor receptor tyrosine kinase inhibitors in cancer patients," *Cancer biology & therapy*, vol. 6, no. 3, pp. 432–438, 2007.

[5] J. M. Stommel, A. C. Kimmelman, H. Ying, R. Nabioullin, A. H. Ponugoti, R. Wiedemeyer, A. H. Stegh, J. E. Bradner, K. L. Ligon, C. Brennan, *et al.*, "Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies," *Science*, vol. 318, no. 5848, pp. 287–290, 2007.

[6] K. Politi, D. Ayeni, and T. Lynch, "The next wave of egfr tyrosine kinase inhibitors enter the clinic," *Cancer cell*, vol. 27, no. 6, pp. 751–753, 2015.

[7] S. Bamford, E. Dawson, S. Forbes, *et al.*, "The cosmic (catalogue of somatic mutations in cancer) database and website," *British journal of cancer*, vol. 91, no. 2, p. 355, 2004.

[8] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, *et al.*, "Egf receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib," *Proceedings of the National Academy of Sciences*, vol. 101, no. 36, pp. 13306–13311, 2004.

[9] R. Qureshi, M. Zhu, and H. Yan, "Visualization of protein-drug interactions for the analysis of lung cancer drug resistance," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[10] D. Prada-Gracia, S. Huerta-Yépez, and L. M. Moreno-Vargas, "Application of computational methods for anticancer drug discovery, design, and optimization," *Boletín Médico Del Hospital Infantil de México (English Edition)*, vol. 73, no. 6, pp. 411–423, 2016.

[11] M. Fukuoka, Y. Wu, Thongprasert, *et al.*, "Biomarker analyses and final overall survival results from a phase iii, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non–small-cell lung cancer in asia (ipass)," *Journal of clinical oncology*, vol. 29, no. 21, pp. 2866–2874, 2011.

[12] R. Rosell, E. Carcereny, R. Gervais, *et al.*, "Erlotinib versus standard chemotherapy as first-line treatment for european patients with advanced egfr mutation-positive non-small-cell lung cancer (eurtac): a multicentre, open-label, randomised phase 3 trial," *The lancet oncology*, vol. 13, no. 3, pp. 239–246, 2012.

[13] J. C. H. Yang, Y.-L. Wu, Schuler, *et al.*, "Afatinib versus cisplatin-based chemotherapy for egfr mutation-positive lung adenocarcinoma (lux-lung 3 and lux-lung 6): analysis of overall survival data from two randomised, phase 3 trials," *The lancet oncology*, vol. 16, no. 2, pp. 141–151, 2015.

[14] J.-C. Soria, Y. Ohe, J. Vansteenkiste, T. Reungwetwattana, B. Chewaskulyong, K. H. Lee, A. Dechaphunkul, F. Imamura, N. Nogami, T. Kurata, *et al.*, "Osimertinib in untreated egfr-mutated advanced non–small-cell lung cancer," *New England journal of medicine*, vol. 378, no. 2, pp. 113–125, 2018.

[15] K. S. Thress, C. P. Paweletz, E. Felip, *et al.*, "Acquired egfr c797s mutation mediates resistance to azd9291 in non–small cell lung cancer harboring egfr t790m," *Nature medicine*, vol. 21, no. 6, p. 560, 2015.

[16] G. Oxnard, J.-H. Yang, H. Yu, S.-W. Kim, H. Saka, L. Horn, K. Goto, Y. Ohe, H. Mann, K. Thress, *et al.*, "Tatton: A multi-arm, phase ib trial of osimertinib combined with selumetinib, savolitinib or durvalumab in egfr-mutant lung cancer," *Annals of Oncology*, 2020.

[17] K. L. Reckamp, V. O. Melnikova, C. Karlovich, *et al.*, "A highly sensitive and quantitative test platform for detection of nsclc egfr mutations in urine and plasma," *Journal of Thoracic Oncology*, vol. 11, no. 10, pp. 1690–1700, 2016.

[18] M. Juchum, M. Günther, and S. A. Laufer, "Fighting cancer drug resistance: Opportunities and challenges for mutation-specific egfr inhibitors," *Drug Resistance Updates*, vol. 20, pp. 12–28, 2015.

[19] E. P. Allain, M. Rouleau, E. Lévesque, and C. Guillemette, "Emerging roles for udp-glucuronosyltransferases in drug resistance and cancer progression," *British Journal of Cancer*, pp. 1–11, 2020.

[20] G. da Cunha Santos, F. A. Shepherd, and M. S. Tsao, "Egfr mutations and lung cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 6, pp. 49–69, 2011.

[21] G. Zhang, S. Jiang, Z. Yang, L. Gong, X. Ma, Z. Zhou, C. Bao, and Q. Liu, "Automatic nodule detection for lung cancer in ct images: A review," *Computers in biology and medicine*, vol. 103, pp. 287–300, 2018.

[22] Y. Gao, J. Xie, H. Chen, S. Gu, R. Zhao, J. Shao, and L. Jia, "Nanotechnology-based intelligent drug design for cancer metastasis treatment," *Biotechnology advances*, vol. 32, no. 4, pp. 761–777, 2014.

[23] C. D. Fjell, J. A. Hiss, R. E. Hancock, *et al.*, "Designing antimicrobial peptides: form follows function," *Nature reviews Drug discovery*, vol. 11, no. 1, p. 37, 2012.

[24] B. Knapp, S. Demharter, R. Esmaielbeiki, *et al.*, "Current status and future challenges in t-cell receptor/peptide/mhc molecular dynamics simulations," *Briefings in bioinformatics*, vol. 16, no. 6, pp. 1035–1044, 2015.

[25] C. M. Song, S. J. Lim, and J. C. Tong, "Recent advances in computer-aided drug design," *Briefings in bioinformatics*, vol. 10, no. 5, pp. 579–591, 2009.

[26] D. Roccatano, A. Barthel, and M. Zacharias, "Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation," *Biopolymers: Original Research on Biomolecules*, vol. 85, no. 5-6, pp. 407–421, 2007.

[27] I. Tinoco Jr and J.-D. Wen, "Simulation and analysis of single-ribosome translation," *Physical biology*, vol. 6, no. 2, p. 025006, 2009.

[28] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: advances and applications," *Advances and applications in bioinformatics and chemistry: AABC*, vol. 8, p. 37, 2015.

[29] Y. Shan, M. P. Eastwood, X. Zhang, *et al.*, "Oncogenic mutations counteract intrinsic disorder in the egfr kinase and promote receptor dimerization," *Cell*, vol. 149, no. 4, pp. 860–870, 2012.

[30] Y. Shan, E. T. Arkhipov, Anton Kim, *et al.*, "Transitions to catalytically inactive conformations in egfr kinase," *Proceedings of the National Academy of Sciences*, vol. 110, no. 18, pp. 7270–7275, 2013.

[31] S. Wan, D. W. Wright, and P. V. Coveney, "Mechanism of drug efficacy within the egf receptor revealed by microsecond molecular dynamics simulation," *Molecular cancer therapeutics*, vol. 11, no. 11, pp. 2394–2400, 2012.

[32] S. Wan and P. V. Coveney, "Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor," *Journal of computational chemistry*, vol. 32, no. 13, pp. 2843–2852, 2011.

[33] J. G. Paez, P. A. Jänne, J. C. Lee, *et al.*, "Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, no. 5676, pp. 1497–1500, 2004.

[34] Y. Shi, J. S. K. Au, Thongprasert, *et al.*, "A prospective, molecular epidemiology study of egfr mutations in asian patients with advanced non–small-cell lung cancer of adenocarcinoma histology (pioneer)," *Journal of thoracic oncology*, vol. 9, no. 2, pp. 154–162, 2014.

[35] D. D. Wang, L. Ma, M. P. Wong, *et al.*, "Contribution of egfr and erbb-3 heterodimerization to the egfr mutation-induced gefitinib-and erlotinib-resistance in non-small-cell lung carcinoma treatments," *PloS one*, vol. 10, no. 5, p. e0128360, 2015.

[36] N. Kureshi, S. S. R. Abidi, and C. Blouin, "A predictive model for personalized therapeutic interventions in non-small cell lung cancer," *IEEE journal of biomedical and health informatics*, vol. 20, no. 1, pp. 424–431, 2014.

[37] S. Wan and P. V. Coveney, "Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor," *Journal of computational chemistry*, vol. 32, no. 13, pp. 2843–2852, 2011.

[38] I. Massova and P. A. Kollman, "Combined molecular mechanical and continuum solvent approach (mm-pbsa/gbsa) to predict ligand binding," *Perspectives in drug discovery and design*, vol. 18, no. 1, pp. 113–135, 2000.

[39] A. R. Brice and B. N. Dominy, "Analyzing the robustness of the mm/pbsa free energy calculation method: application to dna conformational transitions," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1431–1440, 2011.

[40] D. D. Wang, V. H. Lee, G. Zhu, *et al.*, "Selectivity profile of afatinib for egfr-mutated non-small-cell lung cancer," *Molecular BioSystems*, vol. 12, no. 5, pp. 1552–1563, 2016.

[41] N. Huang, B. K. Shoichet, and J. J. Irwin, "Benchmarking sets for molecular docking," *Journal of medicinal chemistry*, vol. 49, no. 23, pp. 6789–6801, 2006.

[42] Q.-H. Liao, Q.-Z. Gao, J. Wei, *et al.*, "Docking and molecular dynamics study on the inhibitory activity of novel inhibitors on epidermal growth factor receptor (egfr)," *Medicinal Chemistry*, vol. 7, no. 1, pp. 24–31, 2011.

[43] B. Rajith, C. Chakraborty, N. Naga Sundaram, *et al.*, "Structural signature of the g719s-t790m double mutation in the egfr kinase domain and its response to inhibitors," *Scientific reports*, vol. 4, p. 5868, 2014.

[44] S. Ikemura, H. Yasuda, Matsumoto, *et al.*, "Molecular dynamics simulation-guided drug sensitivity prediction for lung cancer with rare egfr mutations," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 10025–10030, 2019.

[45] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, *et al.*, "The rcsb protein data bank: integrative view of protein, gene and 3d structural information," *Nucleic acids research*, p. gkw1000, 2016.

[46] L. Ma, D. D. Wang, Y. Huang, *et al.*, "Egfr mutant structural database: computationally predicted 3d structures and the corresponding binding free energies with gefitinib and erlotinib," *BMC bioinformatics*, vol. 16, no. 1, p. 85, 2015.

[47] M. Srivastava, S. K. Gupta, P. Abhilash, and N. Singh, "Structure prediction and binding sites analysis of curcin protein of jatropha curcas using computational approaches," *Journal of molecular modeling*, vol. 18, no. 7, pp. 2971–2979, 2012.

[48] D. Eisenberg, R. Lüthy, and J. U. Bowie, "[20] verify3d: assessment of protein models with three-dimensional profiles," in *Methods in enzymology*, vol. 277, pp. 396–404, Elsevier, 1997.

[49] V. D. Prasasty, U. S. F. Tambunan, and T. J. Siahaan, "Homology modeling and molecular dynamics studies of ec1 domain of ve-cadherin to elucidate docking interaction with cadherin-derived peptide," *OnLine Journal of Biological Sciences*, vol. 14, no. 2, p. 155, 2014.

[50] N. J. Rollins, K. P. Brock, F. J. Poelwijk, M. A. Stiffler, N. P. Gauthier, C. Sander, and D. S. Marks, "Inferring protein 3d structure from deep mutation scans," *Nature genetics*, vol. 51, no. 7, p. 1170, 2019.

[51] A. B. Singh and R. C. Harris, "Autocrine, paracrine and juxtacrine signaling by egfr ligands," *Cellular signalling*, vol. 17, no. 10, pp. 1183–1193, 2005.

[52] D. Dhar, L. Antonucci, H. Nakagawa, J. Y. Kim, E. Glitzner, S. Caruso, S. Shalapour, L. Yang, M. A. Valasek, S. Lee, *et al.*, "Liver cancer

initiation requires p53 inhibition by cd44-enhanced growth factor signaling," *Cancer Cell*, vol. 33, no. 6, pp. 1061–1077, 2018.

[53] F. Morgillo, J. K. Woo, E. S. Kim, W. K. Hong, and H.-Y. Lee, "Heterodimerization of insulin-like growth factor receptor/epidermal growth factor receptor and induction of survivin expression counteract the antitumor action of erlotinib," *Cancer research*, vol. 66, no. 20, pp. 10100–10111, 2006.

[54] X. Yu, K. D. Sharma, T. Takahashi, R. Iwamoto, and E. Mekada, "Ligand-independent dimer formation of epidermal growth factor receptor (egfr) is a step separable from ligand-induced egfr signaling," *Molecular biology of the cell*, vol. 13, no. 7, pp. 2547–2557, 2002.

[55] N. G. Anderson, T. Ahmad, K. Chan, R. Dobson, and N. J. Bundred, "Zd1839 (iressa), a novel epidermal growth factor receptor (egfr) tyrosine kinase inhibitor, potently inhibits the growth of egfr-positive cancer cell lines with or without erbb2 overexpression," *International journal of cancer*, vol. 94, no. 6, pp. 774–782, 2001.

[56] R. Sordella, D. W. Bell, D. A. Haber, and J. Settleman, "Gefitinib-sensitizing egfr mutations in lung cancer activate anti-apoptotic pathways," *Science*, vol. 305, no. 5687, pp. 1163–1167, 2004.

[57] A. Dixit and G. M. Verkhivker, "Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in abl and egfr kinases," *PLoS computational biology*, vol. 7, no. 10, p. e1002179, 2011.

[58] R. Qureshi, A. Ghosh, and H. Yan, "Correlated motions and dynamics in different domains of egfr with l858r and t790m mutations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2020.

[59] N. M. Goodey and S. J. Benkovic, "Allosteric regulation and catalysis emerge via a common route," *Nature chemical biology*, vol. 4, no. 8, p. 474, 2008.

[60] W. Zheng, B. R. Brooks, and D. Thirumalai, "Allosteric transitions in biological nanomachines are described by robust normal modes of elastic networks," *Current Protein and Peptide Science*, vol. 10, no. 2, pp. 128–132, 2009.

[61] A. Ghosh and H. Yan, "Hydrogen bond analysis of the egfr-erbb3 heterodimer related to non-small cell lung cancer and drug resistance," *Journal of theoretical biology*, vol. 464, pp. 63–71, 2019.

[62] L. Wang, H. L. McLeod, and R. M. Weinshilboum, "Genomics and drug response," *New England Journal of Medicine*, vol. 364, no. 12, pp. 1144–1153, 2011.

[63] A. Bag, "Dft based computational methodology of ic50 prediction.," *Current computer-aided drug design*, 2020.

[64] J. Sebaugh, "Guidelines for accurate ec50/ic50 estimation," *Pharmaceutical statistics*, vol. 10, no. 2, pp. 128–134, 2011.

[65] M. Zhao, L. Wang, L. Zheng, M. Zhang, C. Qiu, Y. Zhang, D. Du, and B. Niu, "2d-qsar and 3d-qsar analyses for egfr inhibitors," *BioMed research international*, vol. 2017, 2017.

[66] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," in *Biocomputing 2014*, pp. 63–74, World Scientific, 2014.

[67] B. Zou, V. H. Lee, and H. Yan, "Prediction of sensitivity to gefitinib/erlotinib for egfr mutations in nsclc based on structural interaction fingerprints and multilinear principal component analysis," *BMC bioinformatics*, vol. 19, no. 1, p. 88, 2018.

[68] J. L. Medina-Franco, O. Méndez-Lucio, and K. Martinez-Mayorga, "The interplay between molecular modeling and chemoinformatics to characterize protein–ligand and protein–protein interactions landscapes for drug discovery," in *Advances in protein chemistry and structural biology*, vol. 96, pp. 1–37, Elsevier, 2014.

[69] S.-S. Chiang, C.-I. Chang, and I. W. Ginsberg, "Unsupervised target detection in hyperspectral images using projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1380–1391, 2001.

[70] S. L. Dixon, A. M. Smondyrev, Knoll, *et al.*, "Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. methodology and preliminary results," *Journal of computer-aided molecular design*, vol. 20, no. 10-11, pp. 647–671, 2006.

[71] C. Yoo and M. Shahlaei, "The applications of pca in qsar studies: A case study on ccr5 antagonists," *Chemical biology & drug design*, vol. 91, no. 1, pp. 137–152, 2018.

[72] S. Goyal, S. Jamal, A. Shanker, *et al.*, "Structural investigations of t854a mutation in egfr and identification of novel inhibitors using structure activity relationships," *BMC genomics*, vol. 16, no. 5, p. S8, 2015.

[73] S. Wan and P. V. Coveney, "Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor," *Journal of computational chemistry*, vol. 32, no. 13, pp. 2843–2852, 2011.

[74] C.-H. Yun, K. E. Mengwasser, Toms, *et al.*, "The t790m mutation in egfr kinase causes drug resistance by increasing the affinity for atp," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 2070–2075, 2008.

[75] B. Liu, B. Bernard, and J. H. Wu, "Impact of egfr point mutations on the sensitivity to gefitinib: insights from comparative structural analyses and molecular dynamics simulations," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 2, pp. 331–346, 2006.

[76] R. Qureshi, M. Nawaz, A. Ghosh, *et al.*, "Parametric models for understanding atomic trajectories in different domains of lung cancer causing protein," *IEEE Access*, 2019.

[77] S. A. Foster, D. M. Whalen, Özen, *et al.*, "Activation mechanism of oncogenic deletion mutations in braf, egfr, and her2," *Cancer cell*, vol. 29, no. 4, pp. 477–493, 2016.

[78] B. Zou, V. H. Lee, L. Chen, *et al.*, "Deciphering mechanisms of acquired t790m mutation after egfr inhibitors for nsclc by computational simulations," *Scientific reports*, vol. 7, no. 1, p. 6595, 2017.

[79] D. D. Wang, W. Zhou, H. Yan, *et al.*, "Personalized prediction of egfr mutation-induced drug resistance in lung cancer," *Scientific reports*, vol. 3, p. 2855, 2013.

[80] L. Ma, B. Zou, and H. Yan, "Identifying egfr mutation-induced drug resistance based on alpha shape model analysis of the dynamics," *Proteome science*, vol. 14, no. 1, p. 12, 2016.

[81] Q. Shao and W. Zhu, "Ligand binding effects on the activation of the egfr extracellular domain," *Physical Chemistry Chemical Physics*, vol. 21, no. 15, pp. 8141–8151, 2019.

[82] H. N. Pham and T. H. Le, "Attention-based multi-input deep learning architecture for biological activity prediction: An application in egfr inhibitors," *arXiv preprint arXiv:1906.05168*, 2019.

[83] M. Joo, A. Park, K. Kim, W.-J. Son, H. S. Lee, G. Lim, J. Lee, D. H. Lee, J. An, J. H. Kim, *et al.*, "A deep learning model for cell growth inhibition ic50 prediction and its application for gastric cancer patients," *International Journal of Molecular Sciences*, vol. 20, no. 24, p. 6276, 2019.

[84] L. Ou-Yang, X. Haoran, Z. Mengxu, Y. Hong, *et al.*, "Predicting the impacts of mutations on protein-ligand affinity based on molecular dynamics simulations and machine learning methods," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 439–454, 2020.

[85] B. Yadav, "Quantitative modeling and analysis of drug screening data for personalized cancer medicine," *University of Helsinki*, 2017.

[86] X. Xu, M. C. Farach-Carson, and X. Jia, "Three-dimensional in vitro tumor models for cancer research and drug evaluation," *Biotechnology advances*, vol. 32, no. 7, pp. 1256–1268, 2014.

[87] R. Qureshi, "Personalized drug-response prediction model for lung cancer patients using machine learning," 2020.

[88] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[89] M. Zhu, R. Qureshi, and H. Yan, "Geometrical features of epidermal growth factor receptor-related dimers reveal the mechanisms of drug resistance in lung cancer patients," *IEEE Access*, 2020.

[90] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics (TOG)*, vol. 13, no. 1, pp. 43–72, 1994.

[91] W. Zhou, H. Yan, and Q. Hao, "Analysis of surface structures of hydrogen bonding in protein–ligand interactions using the alpha shape model," *Chemical Physics Letters*, vol. 545, pp. 125–131, 2012.

[92] D. A. Case, T. E. Cheatham III, Darden, *et al.*, "The amber biomolecular simulation programs," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[93] K. Fischer *et al.*, "Computational geometry algorithms library (cgal), reference manual, bounding volumes reference," 2016.

[94] A. Voulodimos, N. Doulamis, A. Doulamis, *et al.*, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[95] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.

[96] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[97] A. Valavanidis, "Scientific reviews artificial intelligence application with machine-learning algorithm identified a powerful broad-spectrum antibiotic,"

[98] S. Wang, J. Shi, Z. Ye, *et al.*, "Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning," *European Respiratory Journal*, vol. 53, no. 3, p. 1800986, 2019.

[99] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[100] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[101] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter, "Deep learning as an opportunity in virtual screening," in *Proceedings of the deep learning workshop at NIPS*, vol. 27, pp. 1–9, 2014.

[102] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[103] H. Singh, S. Singh, D. Singla, S. M. Agarwal, and G. P. Raghava, "Qsar based model for discriminating egfr inhibitors and non-inhibitors using random forest," *Biology direct*, vol. 10, no. 1, p. 10, 2015.

[104] A. A. Toropov and E. Benfenati, "Smiles in qspr/qsar modeling: Results and perspectives," *Current drug discovery technologies*, vol. 4, no. 2, pp. 77–116, 2007.

[105] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[106] N. Pozdeyev, M. Yoo, R. Mackie, R. E. Schweppe, A. C. Tan, and B. R. Haugen, "Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies," *Oncotarget*, vol. 7, no. 32, p. 51619, 2016.

[107] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[108] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, *et al.*, "Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic acids research*, vol. 41, no. D1, pp. D955–D961, 2012.

[109] W. Hugo, H. Shi, L. Sun, M. Piva, C. Song, X. Kong, G. Moriceau, A. Hong, K. B. Dahlman, D. B. Johnson, *et al.*, "Non-genomic and immune evolution of melanoma acquiring mapki resistance," *Cell*, vol. 162, no. 6, pp. 1271–1285, 2015.

[110] P. Jiang, W. R. Sellers, and X. S. Liu, "Big data approaches for modeling response and resistance to cancer drugs," *Annual review of biomedical data science*, vol. 1, pp. 1–27, 2018.

[111] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial intelligence in medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[112] J. Lee, B. T. Miller, Damjanovic, *et al.*, "Constant ph molecular dynamics in explicit solvent with enveloping distribution sampling and hamiltonian exchange," *Journal of chemical theory and computation*, vol. 10, no. 7, pp. 2738–2750, 2014.

[113] M. C. Rendleman, J. M. Buatti, T. A. Braun, B. J. Smith, C. Nwakama, R. R. Beichel, B. Brown, and T. L. Casavant, "Machine learning with the tcga-hnsc dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality," *BMC bioinformatics*, vol. 20, no. 1, p. 339, 2019.

[114] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, *et al.*, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," 2012.

[115] A. A. Kalinin, G. A. Higgins, N. Reamaroon, S. Soroushmehr, A. Allyn-Feuer, I. D. Dinov, K. Najarian, and B. D. Athey, "Deep learning in pharmacogenomics: from gene regulation to patient stratification," *Pharmacogenomics*, vol. 19, no. 7, pp. 629–650, 2018.

[116] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[117] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[118] W. J. Becktel and J. A. Schellman, "Protein stability curves," *Biopolymers: Original Research on Biomolecules*, vol. 26, no. 11, pp. 1859–1877, 1987.

[119] C.-C. Chang, M.-S. Cheng, Y.-C. Su, and L.-S. Kan, "A first-order-like state transition for recombinant protein folding," *Journal of Biomolecular Structure and Dynamics*, vol. 21, no. 2, pp. 247–255, 2003.

[120] P. T. Harrison, S. Vyse, and P. H. Huang, "Rare epidermal growth factor receptor (egfr) mutations in non-small cell lung cancer," in *Seminars in Cancer Biology*, Elsevier, 2019.

[121] S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, M. Zhou, Y. Liu, O. Gevaert, K. Wang, Y. Zhu, *et al.*, "Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning," *European Respiratory Journal*, vol. 53, no. 3, p. 1800986, 2019.

[122] O. Vinyals, C. Blundell, T. Lillicrap, *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, pp. 3630–3638, 2016.

[123] H. Wang *et al.*, "Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Computer Physics Communications*, vol. 228, pp. 178–184, 2018.

[124] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, "Acemd: accelerating biomolecular dynamics in the microsecond time scale," *Journal of chemical theory and computation*, vol. 5, no. 6, pp. 1632–1639, 2009.

[125] M. J. Harvey and G. De Fabritiis, "Acecloud: molecular dynamics simulations in the cloud," 2015.

[126] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domanski, D. L. Dotson, S. Buchoux, I. M. Kenney, *et al.*, "Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

[127] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: a modern open library for the analysis of molecular dynamics trajectories," *Biophysical journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[128] L. Skjærven, S. Jariwala, X.-Q. Yao, and B. J. Grant, "Online interactive analysis of protein structure ensembles with bio3d-web," *Bioinformatics*, vol. 32, no. 22, pp. 3510–3512, 2016.

[129] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[130] J. Li, S. Zheng, B. Chen, *et al.*, "A survey of current trends in computational drug repositioning," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 2–12, 2015.

[131] R. Kneller, "The importance of new companies for drug discovery: origins of a decade of new drugs," *Nature Reviews Drug Discovery*, vol. 9, no. 11, pp. 867–882, 2010.