Computational Methods for the Analysis and Prediction of EGFR-Mutated Lung Cancer Drug Resistance: Recent Advances in Drug Design, Challenges and Future Prospects

Rizwan Qureshi, Bin Zou, Tanvir Alam, Jia Wu, Victor H. F. Lee, and Hong Yan

Abstract—Lung cancer is a 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 cases. It has been well established that a 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 overview of computational methods, for understanding drug resistance mechanisms. The important EGFR mutants and the different generations of EGFR–TKIs, with the survival and response rates are discussed. Next, we evaluate the role of important EGFR parameters in drug resistance mechanism, including structural dynamics, hydrogen bonds, stability, dimerization, binding free energies, and signaling pathways. Personalized drug resistance prediction models, drug response curve, drug synergy, and other data-driven methods are also discussed. Recent advancements in deep learning; such as AlphaFold2, deep generative models, big data analytics, and the applications of statistics and permutation are also highlighted. We explore limitations in the current methodologies, and discuss strategies to overcome them. We believe this review will serve as a reference for researchers; to apply computational techniques for precision medicine, analyzing structures of protein-drug complexes, drug discovery, and understanding the drug response and resistance mechanisms 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, AlphaFold2, deep learning

1 Introduction

Cancer is the second largest cause of deaths worldwide, resulting in a loss of about 10 million lives in 2020 [1]. About 13% of all new cancer cases are lung cancer and more than 65% of them are diagnosed at advanced cancer stages [2].

Epidermal Growth Factor Receptor (EGFR) is a protein that controls cell proliferation and survival. It has been linked to a variety of human malignancies, where over-expression or mutations promote its oncogenicity. EGFR activity and medication sensitivity have been linked to mutations. The over-expression of EGFR is found in about 60% of non-small cell lung carcinomas (NSCLCs) [3]. 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 drug resistance develops in most cases after about an year [3], which limits drug efficacy.

The discovery of small molecule inhibitors targeting the kinase EGFR has generated much hope for the further advances in the treatment of lung cancer. Still, the limitations of their efficacy are apparent [4]. The up-regulation and engagement of efflux pumps to remove the drug from the binding site is one of the primary reasons for the acquired resistance [5]. 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 [6].

- *Rizwan Qureshi is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, and with the Fast School of Computing, National University of Computer and Emerging Sciences, Karachi 75160, Pakistan, and also with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. E-mail: engr.rizwanqureshi786@gmail.com.*
- *Bin Zou and Hong Yan are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. E-mail: binzou2-c@my.cityu.edu.hk, h.yan@cityu.edu.hk.*
- *Tanvir Alam is with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. E-mail: talam@hbku.edu.qa.*
- *Jia Wu is with the Department of Imaging Physics, University of Texas MD Anderson Cancer Center, Houston, TX 77030 USA. E-mail: JWu11@mdanderson.org.*
- *Victor H. F. Lee is with the Department of Clinical Oncology, Li Ka Shing Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong, China. E-mail: rrrizwan2-c@my.cityu.edu.hk.*

Manuscript received 6 Apr. 2021; revised 31 Dec. 2021; accepted 2 Jan. 2022. Date of publication 10 Jan. 2022; date of current version 3 Feb. 2023.
This work was supported in part by the Hong Kong Research Grants Council under Grant CityU 11204821, Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), and the City University of Hong Kong under Grant 9610034.
(Corresponding author: Rizwan Qureshi.)
Digital Object Identifier no. 10.1109/TCBB.2022.3141697

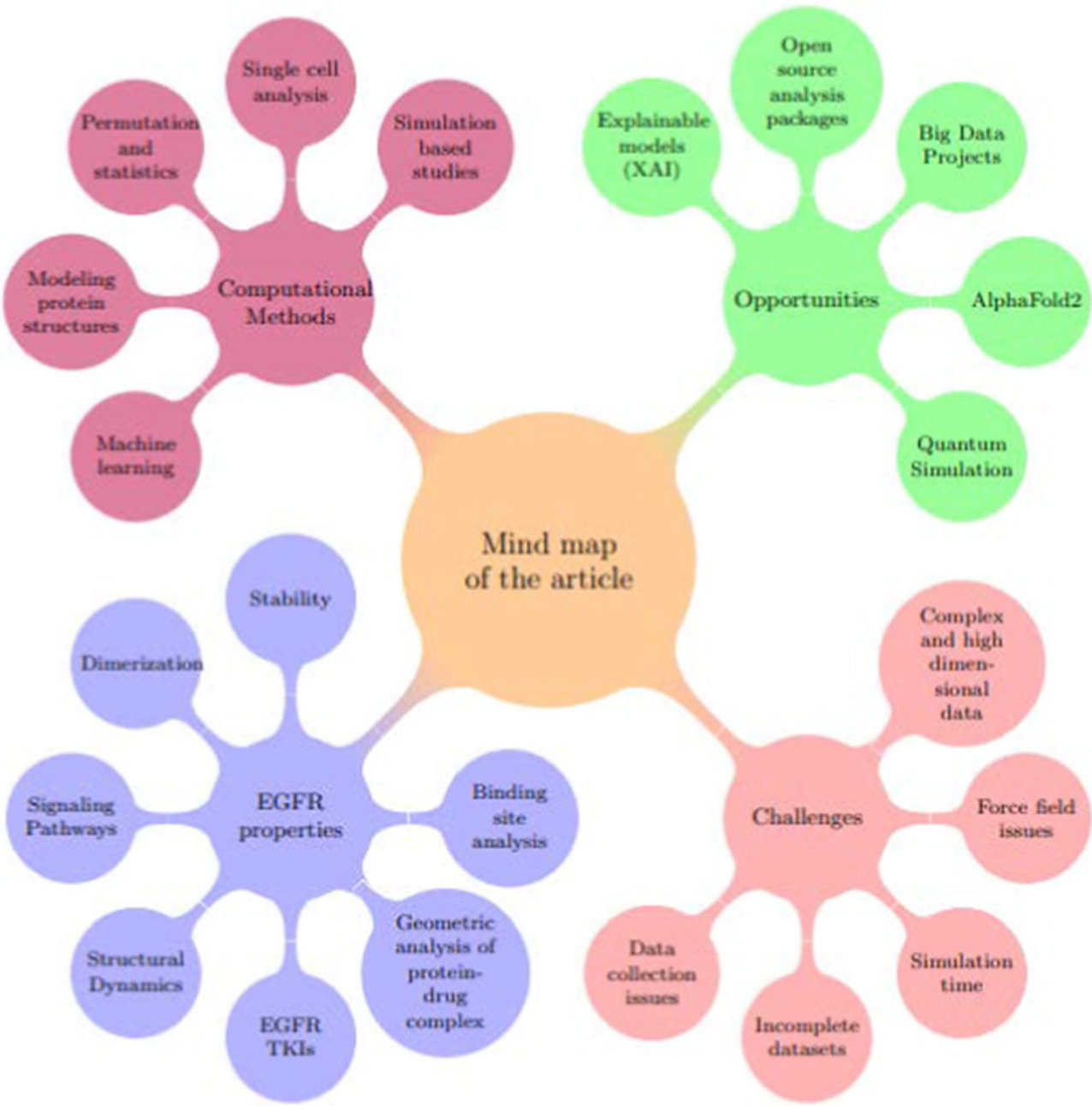


Fig. 1. A mind map of the article. We discuss important computational studies of EGFR mutants, comparison of EGFR properties, including stability, dynamics, signaling pathways binding free energy, and mechanisms of drug resistance. Computational methods, including simulation based and data-driven methods for drug discovery, response, and resistance analysis are explained. Recent advancements, such as AlphaFold2, Big data projects, open source software analysis packages, and quantum simulation are also discussed. Challenges in the current methods including time limits, force field approximations, incomplete datasets and lack of interpretability of machine learning methods are presented.

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 (TagrissoTM, [AZD9291] AstraZeneca, Cambridge, UK) is an EGFR-TKI that specifically targets activating EGFR mutations, as well as the T790M-resistance mutation. It forms a covalent link to the C797 position in mutant EGFR's ATP-binding region [7], but the efficacy of Osimertinib is lost upon the emergence of C797S mutation (Fig. 3). Osimertinib was also approved in 2018 as a first-line treatment for advanced EGFR-mutated NSCLC, independent of the presence or absence of the T790M mutation [8]. There was no significant difference in Ordinary Response Rate (ORR) between osimertinib and conventional EGFR-TKI therapy (80 percent versus 76 percent). Despite these findings, sequencing treatment with first- or second-generation EGFR-TKIs followed by osimertinib could be a viable alternative with consistent data, but only for patients with T790M mutations. The 4th generation drug EAI045 appeared in 2016, targets both T790M and C797S giving hope to NSCLC patients. EAI045 is the first allosteric inhibitor to target the EGFR mutations T790M and C797S. It is only effective, when taken alongside cetuximab [9]. To overcome the limitation, more bench research and biochemical optimization are required. Clinical trials are needed to confirm its efficacy in advanced NSCLC patients [10]. Table 1

TABLE 1
EGFR Tyrosine Kinase Inhibitors With Response and Survival Rates

Reference	Generation	Drug	Mutation	PFS (Months)	ORR
IPASS [18]	1st	Gefitinib	EGFR del 19, L858R	9.5	17.2%
EURTAC [19]	1st	Erlotinib	EGFR del 19, L858R	9.7	64%
Lux – Lung [20]	2nd	Afatinib	EGFR del 19, L858R, rare mutations, HER2/4	11	66.9%
NA	2nd	Neratinib	EGFR G719X, HER2/4	10	NA
FLAURA [8]	3rd	Osimertinib	mEGFR, T790M	18.9	80
AURA2 [21]	3rd	Osimertinib	mEGFR	8.6	71%
AURA1/2	3rd	Olmutinib	mEGFR/T790M	7	56%
ARCTIC	3rd	durvalumab	mEGFR	NA	NA
TATTON [22]	3rd	Osimertinib	MET, BRAF, C797S	NA	NA
TIGER-X [23]	3rd	Rociletinib	T790M	9.6	45%
NA	4th	EAI045	T790M/C797S	NA	NA

PFS: Progression free survival, ORR: Ordinary response rate

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 [11] is the largest open access database for EGFR mutations, which shows that about 93% of EGFR mutations are found in the four exons 18 – 21 of the kinase domain. These mutations include in-frame deletion in exon 19, insertion in exon 20, and a single point substitution in exon 21. Although variation in medication sensitivity within a single exon has been found, clinical trial design and treatment of individuals with atypical EGFR mutations often rely on mutated-exon location to predict treatment. Nucleotide substitution in exon 18 (G719S or G719C) accounts for about 5% of the EGFR mutations [12]. The statistics of different EGFR mutations are shown in Fig. 2.

In [13], the EGFR mutational landscape is characterized in 16715 NSCLC patients, and a structure-function relationship is developed based on drug sensitivity. It was found that the EGFR mutations can be divided into four subgroups based on drug sensitivity and structural changes. These findings support a structure-based approach to define functional categories or molecular descriptors of EGFR mutations, which can help patients with EGFR-mutant make informed treatment and clinical trial decisions.

Besides genetic mutations, there are other mechanisms of drug resistance [14]. Tumor epigenetic alterations can alter gene expression patterns, allowing them to adapt to targeted therapy and develop acquired drug resistance [15]. Post-translational activation of signaling pathways can escape the therapeutic target, and may lead to changes in gene patterns, causing drug resistance [16]. Cancer stem cell (CSCs) have also been hypothesised as a mechanism for explaining tumour initiation, development, and more crucially, therapeutic resistance. Many biological processes and pathways have been discovered to be involved in the drug resistance, causing CSC-mediated cellular mechanisms [17].

Experiment methods are usually used for drug resistance analysis. These 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 compounds, visualize the protein-ligand interactions [26], de novo molecule design and optimization, and the identification of drug binding sites [27]. In addition, the massive computational power and the advanced data analysis methods provide opportunities to study the complex biological and chemical systems at atomic level. One such method is molecular dynamics (MD) simulation, for analyzing the physical movements of atoms, molecules and protein-drug complexes. MD simulation are widely used in structure based drug design, drug resistance and response analysis. Other methods include molecular docking, quantitative structure activity representation and deep learning based methods for precision medicine, drug discovery, design and analysis. This article presents a review of computational methods for drug discovery, resistance and response analysis in lung cancer patients.

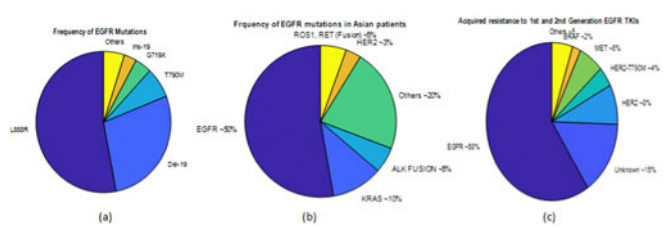


Fig. 2. The frequency of EGFR mutations, mutation rate in Asian patients, and acquired resistance [24], [25]. L858R is the most common type of EGFR mutation (a). Mutation rates in Asian patients (b). The EGFR mutation are the most common mechanism associated with the drug resistance to 1st and 2nd generation of EGFR TKI (c).

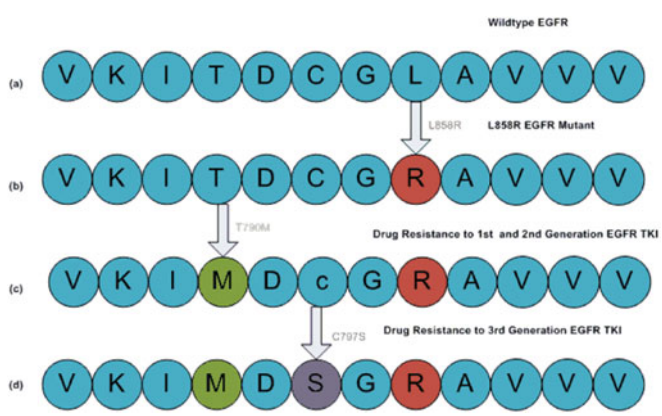


Fig. 3. Wildtype EGFR structure (a), L85R drug sensitive mutant (b), T790M drug resistive mutant to the 1st and 2nd generation drugs (c), and C79S drug resistive mutant to 3rd generation drug (d).

Although there are reviews on EGFR-mutated drug resistance in NSCLC [28], [29], [30], these studies mainly focus on clinical data and random trials. The methods using CT scan lung cancer images are also covered in [31]. The Nanotechnology based intelligent cancer drug design is discussed in [32].

This article presents a comprehensive review of computational studies on EGFR-mutated NSCLC patients, including simulation based studies and data-driven methods. Challenges in the current methodology and potential future opportunities are also presented. The rest of this paper is organized as follows. Section 2 presents computational methods for drug resistance analysis. Personalized drug response prediction models are described in Section 3. The use of deep learning and big data analytics is highlighted in Section 4. Challenges and opportunities in the current methodologies are given in Section 5, and finally, we conclude this review with potential future research directions in Section 6. A mind-map of the paper is given in Fig. 1.

2 Computational Methods for EGFR-Mutated Drug Resistance

Computational methods are widely used for the drug resistance analysis, due to the flexibility, low cost, easy implementation, and the ability to process a large amount of data [33]. These methods can provide deeper insights, generating novel hypothesis and devise new promising strategies. These methods can broadly be classified into structure-based and ligand-based methods. Computer aided drug design and analysis methods [34], relying on the 3-dimensional structure of protein is known as structure-based methods, whereas the ligand-based methods, relying on the information coming from known ligands, binding to the target protein. In this paper, we mainly focus on structure-based drug design and analysis methods.

2.1 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 [35]. MD simulation enables to examine the structural changes occurred due to genetic mutations, and is widely used for drug resistance analysis, prediction and discovery [36]. The interactions between drug and protein are a dynamic process, involving several residues. MD simulations can reveal the atomic level details of drug-protein interactions.

Since its beginning, MD simulations have progressed from simulating a few hundred atoms to systems with biological relevance, including entire protein with solvent, protein with membrane embedded, and large macromolecular complexes like nucleosomes [37] or ribosomes [38]. This immense improvement is due to the high performance computing (HPC) and basics of MD simulation algorithm [39]. 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*. [40] used long time scale MD simulation and showed that N-lobe of the wildtype (WT) EGFR is disordered (partial folding), and it becomes ordered only upon dimerization, as shown in Fig. 4 (leftmost). They also showed that some cancer-specific mutations L858R/L834R may facilitate the dimerization by suppressing this local disorder. The L858R mutation causes abnormally high kinase activity by promoting EGFR dimerization. They further performed unbiased, all-atom MD simulations of EGFR kinase domain [41], and showed that EGFR monomer is more stable in its inactive state than its active state.

Shunzhou *et al*. [43] used 10 micro-second long MD simulation for the investigation of conformational dynamics and interactions of EGFR mutants. The simulation result shows that the mutant type L858R binds to Gefitinib, rather than ATP, Fig. 4 (rightmost). 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. They used the MD simulation of [43] 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.

EGFR mutants L858R, L858R - T790M, and C797S were simulated with Gefitinib and Osimertinib in [26]. The binding site roughly included 14 amino-acids, and the Gefitinib was simplified into its essential pharmacophore, represented by the center positions of two carbon rings. The distance was calculated between the 14 binding-site residues and two center atoms of the the drug molecule, which can be represented by a 14×2 matrix. For the drug-sensitive disease mutation L858R, the distance between Gefitinib and the binding site residues is less than the WT structure, which is consistent with good evidence, whereas in the drug-resistant L858R-T790M double mutant the drug is further away in the mutant, indicating weak binding. A third-generation medication, Osimertinib, was developed to overcome drug resistance to Gefitinib in the L858R-T790M double mutant. The distance between Osimertinib and EGFR or its mutants shows that the drug remains close to EGFR in the L858R-T790M mutation, consistent with its treatment efficacy, while the distance between the drug and EGFR increases with the additional mutation, C797S, (Fig. 5).

In binding free energy calculation, compared to L858R, a second mutation T790M (co-expressed L858R and T790M mutations) leads in a higher binding affinity (lower binding free energy) with receptor tyrosine kinase (RTKs), suggesting that a tighter L858R-T790M-RTK, when crosstalk is involved in drug resistance with Gefitinib. Osimertinib, an irreversible third generation drug, used for patients with L858R-T790M mutations, the binding free energy is high, indicating tighter binding. However, the additional C797S mutation results in lower binding free energy. These findings are consistent with the current clinical literature.

The atomic level details of interaction between the binding sites residues and the ligand is analyzed in [44]. The analysis is carried out using the contact frequency of residues and the ligand during the entire MD simulation. The analysis identifies 39 amino acids, close to the binding site. Most of the residues are located at the β sheet and helixes. L718, V726, A743, M793 and L844 have the highest contact frequencies, and are located at $\beta 1 - 3$, hinge region, and $\beta - 6$, forming the binding pocket. Most co-crystallized EGFR

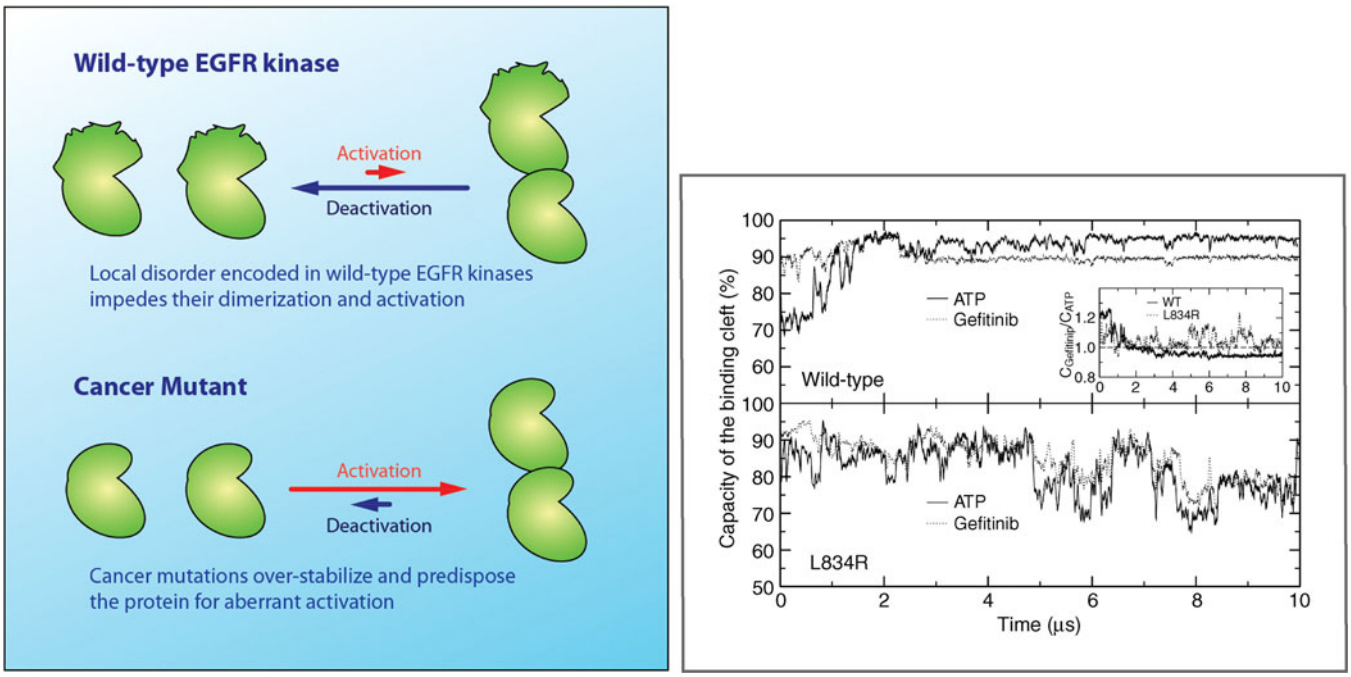


Fig. 4. EGFR kinase is intrinsically disordered in the dimerization, L858R cancer mutant stabilize the interface [40], (Leftmost). The binding preference for WT EGFR is ATP, and for L858R, the binding preference is Gefitinib [42] (Rightmost), Figure taken from [40].

kinase inhibitors are ATP-competitive, based on the high contact frequencies in the core binding site.

Molecular docking is another important tool in structure-based drug design, due to its low-cost and simplicity [46]. In molecular docking, the position and orientation of a molecule are predicted against another 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 [45], molecular docking and MD simulation are used for the investigation of structural and chemical features of EGFR inhibitors and binding pocket mechanisms. It was found that the binding pocket consists of three regions (P1, P2, P3), composed of mostly hydrophobic residues. The 7-aniline substituent is located in a hydrophobic pocket (P1 site) that includes Met766, Leu777, Phe856, and Leu858 side chains. The side chains of Val726, Ala743, Lys745, Leu788, Thr790, Thr854, and Asp855 (named P2 site) sandwich the thiazolopyrimidine ring. The surrounding residues involved in hydrophobic interactions with the phenyl group, off the C2 position of the thiazolopyrimidine ring, are Val726, Cys797, Leu844, and Thr854 (named P3 site), as shown in Fig. 6. These residues can accommodate the lipophilic arms of the compounds. The interaction with Phe856 is hydrophobic interaction, and the presence of residues Met793 and Asp855 may also be responsible for the binding recognition through H-bond interactions, with Phe856 through a T-shape $\pi - \pi$ stacking interaction.

Rajith *et al.* [47] investigated the G719S-T790M double mutation with ligand Gefitinib, using 50-ns MD simulation and molecular docking. They observed the escalation in the distance between P-loop and functional loop in T790M mutation compared with the G719S. They also verified that the 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 another reason for the aberrant function of EGFR-TKIs in T790M mutant.

2.2 Computational Modeling of EGFR Mutations and Binding Free Energy Calculation

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 [48].

Starting from the structures found in the Protein Data Bank (PDB) [49], Ma *et al.* [50] 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 (Fig. 7 leftmost). 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 other types of amino-acid mutations, such as; insertion, deletion, and duplication, Rosetta comparative (CM) protocol was applied. The framework for building the EGFR mutant structure database is shown in Fig. 7 (leftmost), and a computationally predicted EGFR-mutant dimer with Gefitinib drug is shown in Fig. 7 (rightmost). The predicted models may not be accurate, and several other computational tools, including Q-MEAN Z-score [51], Verify3D [52], and Ramachandran plot [53] can be used to validate the predicted models.

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. EGFR-TKI binding mechanism and emergence of drug resistance

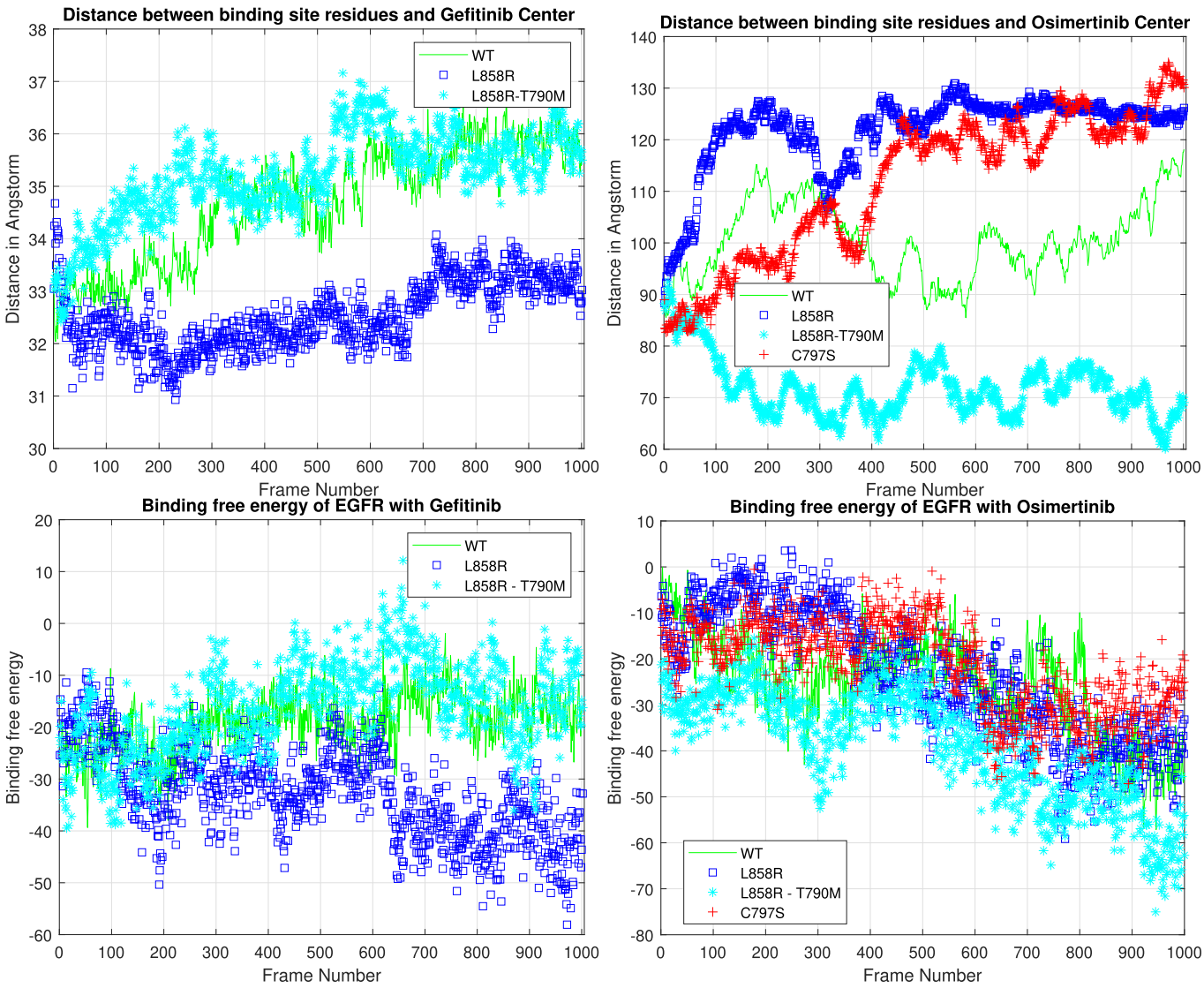


Fig. 5. Distance and binding free energy of EGFR protein drug complexes with Gefitinib and Osimertinib [26].

to genetic mutations, is shown in Fig. 8. The energetic contribution of individual residues provide useful quantitative information about the binding mode of a protein-ligand [55]. Wang *et al.* [43] 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 [70] tool in Amber [71]. They further analyzed how the mutations affect 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.

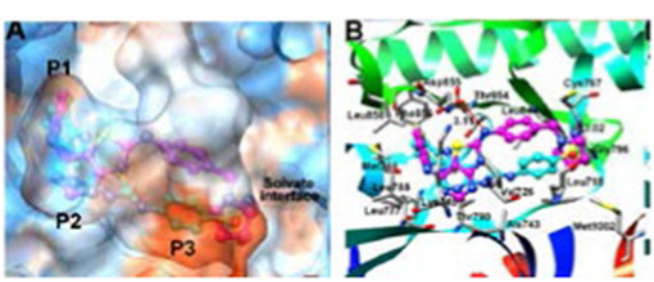


Fig. 6. Binding pocket consists of three regions (P1, P2, and P3 (A). Residues involved in the hydrophobic interactions (B) [45].

The shapes and sizes of WT and mutants EGFR show that they sample different structural regions. In the active state, the most visible difference between WT and mutant EGFR arises, with the L858R mutant sampling additional conformational space that is most energetically favourable for the mutant kinase. The active conformations observed in X-ray structures roughly lie in a line in the space of the first two PCs. In the L858R simulation, the additional conformation is located on a branch of this line. The orientation of the αC-helix in the inactive state and an expanded A-loop in the active state define this extra conformation. This conformation could be one of the intermediate states between active and inactive states in the transformation pathway. The study reveals that the L858R mutant induces conformational changes in active and inactive states, which affect relative stability.

In another study, Wang *et al.* [72] profiled the correlation between the EGFR mutations and the potency of EGFR-TKI Afatinib (second-generation drug). The progression-free survival (PFS) rate and drug-response level were recorded, as the end point of the study. The drug response is

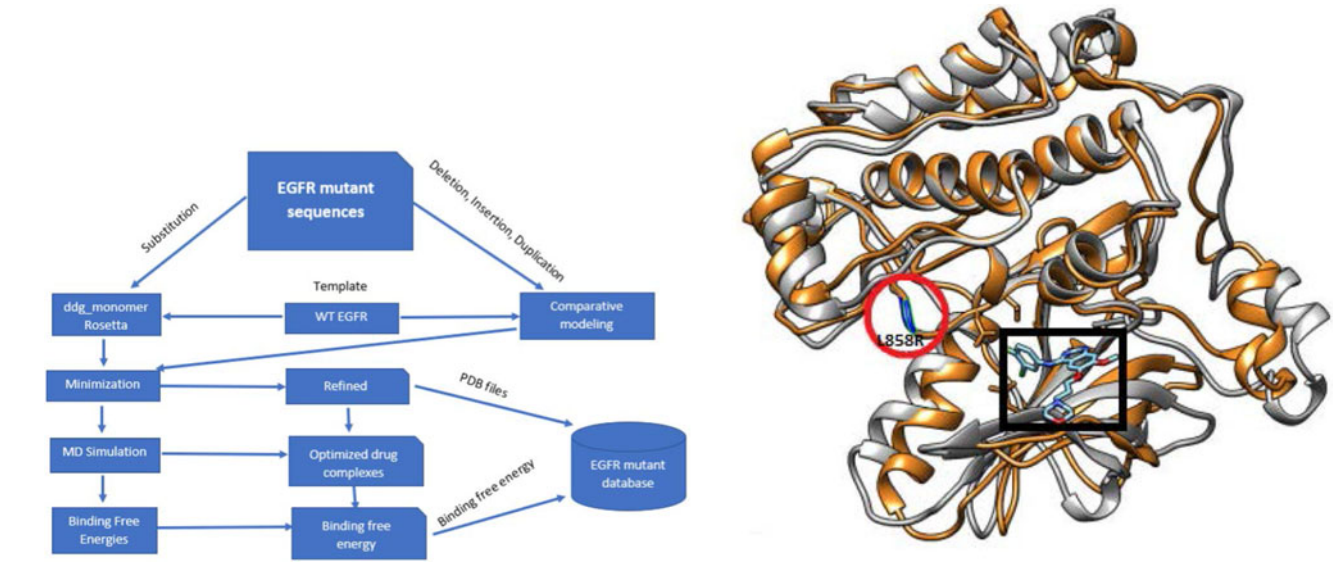


Fig. 7. The framework for predicting the mutant structure of EGFR and their binding energies [50] (Leftmost). A computationally predicted EGFR-L858R mutant dimer with Gefitinib drug. The wildtype EGFR is colored grey, L858R mutant is orange, mutation site is shown in blue, and the drug is shown in black square. (Rightmost).

determined based on RECIST [73], (Response evaluation criteria in solid tumors). It was divided into four types (1, 2, 3, 4), with the lower values of drug response show weak binding and less efficacy. 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. The WT and the single T790M mutation showed no response and no survival. Results of this study verify the higher potency of Afatanib to classical EGFR mutation L858R and exon -19 deletion, and a lower one to T790M mutation, which is consistent with the clinical values.

2.3 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 [75]. EGFR dimerization is an essential event in the EGF-signal transduction [76]. The dimerization stimulates the catalytic activity of tyrosine kinase domain, and promotes the autophosphorylation (ATP) of several residues in the kinase domain [77]. These residues provide docking sites for downstream signaling molecules, such as Shc, Grb2 and P13k. It is also known that the EGFRs can form dimer on the cell surface, independent of ligand binding [78]. AKT signaling is moderated by a complex network of protein, and is commonly deregulated in cancer [79]. Tumors, that are sensitive to EGFR TKIs are characterized by a rapid decrease in the Akt activity [80], (Fig. 9). The Akt pathways shut down and lung cancer cells suffer apoptosis after treatment with EGFR inhibitors [81]. The failure to irregulate Akt [82] or reactivation of Akt pathway causes drug resistance [83]. In Table 2, we list several important computational studies on EGFR mutants related to drug resistance.

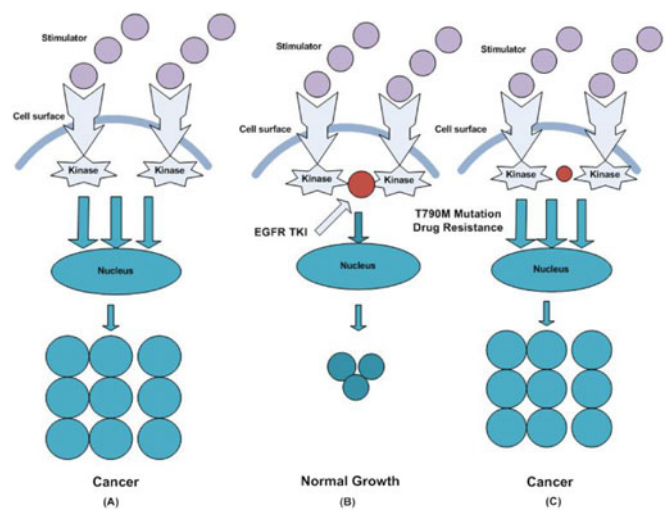


Fig. 8. EGFR TKI Binding mechanism, the leftmost panel (a), shows the EGFR mutations and uncontrolled cell growth, that leads to cancer; L858R mutant binds tightly with the EGFR-TKI; the over-expression is stopped and normal growth is recovered (b). The large red circle shows tight binding with the EGFR kinase. The T790m mutation causes drug resistance; EGFR-TKI is unable to bind the kinase, indicated by small red circle the binding becomes weak (c).

Wang *et al.* [54] investigated the contribution of EGFR and ErbB-3 heterodimerization based on their 3D structures. Based on the molecular dynamics of these systems, the binding free energy and its components were used to characterise both the mutant-inhibitor (protein-ligand) and mutant-partner (protein-protein) interactions. They characterized the EGFR mutations for 168 clinical subjects using molecular interactions for three EGFR dimers (ErbB-2, IGF-1R, c-Met) (mutant partners), and with two EGFR inhibitors, Gefitinib and Erlotinib.

Simulation results showed that the mutant–partner interactions increased in the L858R - T790M system and has a weaker connection with inhibitor Gefitinib or Erlotinib. L858R has the highest binding free energy with Gefitinib, lowest with the Erbb-2, and average with either IGF-1R, c-Met or Erlotinib. The mutant-partner interactions can have a negative impact on inhibitor efficacy while mutant-inhibitor interactions have a beneficial impact. The mutant delL747 - P753insS has the largest difference in binding free energy between the mutant interactions and inhibitors, and this may be the reason for the shorter progression free

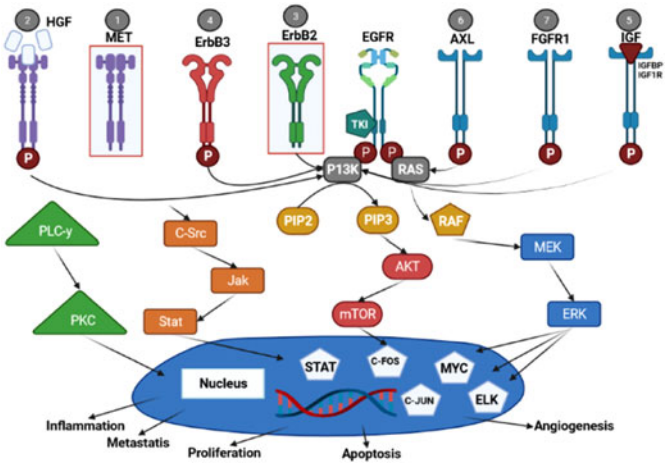


Fig. 9. EGFR RTK Dimerization and downstream signaling mechanism [54]. In downstream signaling, Ras/Raf/Mek/Erk and PI3K/Akt/mTOR are two general pathways [74]. Importantly, the causes of AKT pathway activation will influence the possibility of PI3K/AKT inhibition providing therapeutic benefit.

survival in this mutant type. As a supplementary investigation, interactions between ErbB-3 and prospective partners were evaluated and compared, again using binding free energies as a basis. c-Met had a greater interaction with ErbB-3 than EGFR mutants, implying that it plays a major role in ErbB-3 signalling.

2.4 Single-Cell RNA Sequencing (scRNA-seq) and Cancer Cells

By offering a high resolution of transcriptome alterations at the single-cell level, single-cell RNA sequencing (scRNA-seq) technology is a potential approach for characterising individual cells and understanding molecular pathways [84]. The tumour microenvironment and its cells play an important role in tumour evolution, and the development of cancer may be understood by the complex network of intracellular and intercellular signals. A computational pipeline is developed in [85] by combining scRNA-seq data with clinical bulk gene expression data, to identify prognostic and predictive biomarker, between cancer cell and micro-environmental cells. The pipeline identified a tumor associated with microglia/macrophage-mediated EGFR/ERBB2 feedback-crosstalk signaling module, defined as multi-layer network biomarker (MNB). The biomarker (MNB) is used to predict the survival outcome, therapeutic response and drug sensitivity or resistance to molecularly targeted therapeutics. The results on publicly available datasets showed improved performance of MNB, compared with the other gene biomarkers. The proposed MNB technique could aid in the development of more effective biomarkers for predicting cancer patient prognosis and therapy resistance.

2.5 Long Range Communication Capability and Stability of EGFR-Mutants

Allosteric Communication is an important phenomenon in many biological processes, which is considered to be a useful parameter in governing molecular motion and signal transduction [86], [87]. Recent studies [88], [89] show that allosteric networks of cooperative protein motion may be formed by a sparsely connected group of residues.

In [86], Dixit and Verkhivker used (MD) simulation, PCA and signal propagation in protein to identify the allosteric

TABLE 2
Important Computational Studies Related to EGFR Mutation Induced Drug Resistance

Reference	Mutant Sequences	Algorithms	Applications
[3]	EGFR mutants	Review	Review
[42]	WT and L858R	MD and PCA	Drug efficacy to Gefitinib
[54]	EGFR, ErbB-3, IGF, cMet	MD and structural analysis	Contribution of hetrodimers in drug resistance
[55]	Clinical data	pattern minning and univariate analysis	Personalized predictive models
[47]	WT, T790M, G719S, and T790M-G719S	Molecular dynamics and docking analysis	Structural investigation
[48]	Rare mutations & clinical data of 3779 patients	MD & Statistical analysis	EGFR TKI sensitivity prediction
[50]	112 EGFR mutants	MD & Clinical data of 942 patients	Binding energies
[56]	EGFR ErbB-3 hetrodimers	Hydrogen bonds	Stability analysis
[57]	30 EGFR mutants	Interaction footptint matrix and PCA	EGFR-TKI sensitivity prediction
[58]	WT, L858R, T790M	computational methods	Affinity for Gefitinib and ATP
[59]	WT & T854A	MD & PCA	Structural investigation
[60]	L858R, T790M, G719C and L861Q	MD &NMA	Impact of point mutants to Geftinib
[61]	WT, L858R, L858R-T790M	Parametric methods	EGFR domain's analysis
[62]	HER2, BRAF and EGFR	Molecular modeling	effect of in frame deletion
[63]	WT and T790M	Attribute ranking	Mechanism of T70M mutation
[64]	32 EGFR mutants	MD and extreme learning machines (Elm	Personalized drug resistance prediction
[65]	EGFR and 30 mutants	Alpha shape dynamics	Drug resistance prediction
[13]	Exon 18 - 21 Mutants	Structure function relationship	Drug response
[66]	EGFR extracellular domain	Energetics and PSN	Ligand binding effects
[67]	3492 Compounds of EGFR	Deep Learning attention mechanism	EGFR drug discovery
[68]	CCLE data	DeepIC_{50}	Drug sensitivity prediction
[69]	110 mutations	Machine learning and time-series	Mutation impact prediction

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 [42] 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 [42], showing that L858R has higher binding preference to Gefitinib than the ATP.

Hydrogen bonds are another important parameter, for the analysis of biological and chemical interactions of molecules and the stability of protein structures. Ghosh and Yan [56] 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. They found L858R drug-sensitive mutant has the highest number of hydrogen bonds. The mean value was 481 for WT, 484 for L858R and 477 in L858R - T790M (complex mutation) structure. They concluded that the T790M mutant structure has a smaller number of hydrogen bonds, which changes the conformational stability of the system, causing the development of drug resistance.

2.6 Quantitative Structural Activity in EGFR-Drug Complexes and Pharmacophore Modeling

QSAR is applied for establishing a relationship between chemical properties of a compound and their biological activities [90]. QSAR structural activity began with 0-D and has progressed to 6-D. Each dimension evolved from a desire to transcend the constraints of previous dimensions and to outperform them. The molecule properties such as electrical, hydrophobic, steric, and so on were computed using 1-D QSAR. Geometric characteristics, topological indices, molecular fingerprints, and polar surface area are all considered in 2-D QSAR, but steric features are not. The spatial features of the compound are the focus of the 3-D QSAR approach. As a result, the shortcomings of the 2D QSAR method are resolved in the 3D QSAR method [91].

Sukrita *et al.* [59] performed MD simulations and applied principal component analysis (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. 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 becomes flexible upon T854A mutation and losses stability in Root mean square deviation (RMSD), Root mean square fluctuation (RMSF), and Radius of gyration (Rg). The RMSD is a commonly used metric for comparing predicted and observed protein-drug complexes, whereas RMSF is a measure of the difference between the position of residues during i^{th} index and a reference position.

In the Computer-aided drug design method, a structure and ligand-based pharmacophore model may be used to find similar active compounds against a specific target protein, and the binding affinity of a large size chemical with a target macro-molecule can be easily assessed using an in-silico molecular docking technique [92]. Next generation EGFR inhibitors were identified using 3D-pharmacophore modeling and virtual screening. The pharmacophore model consists of seven 3D points, one hydrogen bond donor, three hydrogen bond acceptors, and three aromatic rings. CUDC101, a multi-targeted kinase inhibitor, outperformed the well-known EGFR inhibitors Gefitinib and Erlotinib, in terms of binding free energy and 3D pharmacophore fit value. This computational study may serve as a foundation for identifying and designing more potent EGFR next-generation kinase inhibitors.

The L858R mutation in the kinase domain suppresses the local disorder, and provides a series of conformation states, which is capable of binding to Gefitinib. This provides an explanation for effectiveness of Gefitinib in L858R EGFR mutated NSCLC patients. Nevertheless, 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 Theonine 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. The deletion of five amino acids (ELREA) in exon-19 is also a cause for abberant function [93].

Computational methods [35], [41], [43], [46], [85], [90], [94] provide useful information about the structure and dynamics of EGFR mutants, and may help designing of new drugs, as well as analyzing the drug response. Different simulation packages, protocols for structure predictions, may yield different results. Multiple runs of simulation with multiple software packages, and predicting structures with multiple methods, may increase the confidence level [95] in these simulation-based studies.

3 Personalized Drug Response Prediction Models

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 [96]. 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 [97], [98], [98]. It is plausible to argue that individualised therapy for NSCLC patients should include a genetic assessment of their EGFR mutation status.

Wang *et al.* [64] used binding energy of EGFR mutant complex and personal features (age, sex, smoking-history, medical-history) to build a personalized drug resistance prediction model. Extreme learning machines (ELMS) [99] were applied to predict the drug resistance level. EGFR-TKI interaction pattern and personal features are used in the prediction model, and overall accuracy of 95% was achieved. The accuracy of the model is significantly increased with the addition of personal features.

Kureshi *et al.* [55] 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 overall accuracy of 76.54% with area under the curve (AUC) equal to 0.76. They showed that the Support Vector Machine (SVM) and the decision trees are potential candidates for personalized drug resistance prediction.

3.1 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 [100]. 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.* [65] analyzed the properties of EGFR mutants using structural information. Alpha shape model [101] and solid angle [102] were computed to evaluate the properties of the atom at the binding sites.

MD simulations were performed using Amber [103] suite, and the Computational Geometry Algorithms Library (CGAL) [104] 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 values in the range [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 the mutant by averaging the convex degrees for 200 frames. It is shown that about 90% of the mutants 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 efficacy to Gefitinib, which is consistent with our knowledge.

In a recent study [98], a combination of geometric features, protein-drug complex binding energy, and clinical information is fed to predict the four class drug response, the proposed model achieves state of the performance. Previous studies [105] have also shown that combining clinical and genetic or genomics information improves the capability to predict lung cancer risk stratification. These studies show that a combination of smoking-status, age, tumour tissue histology, and EGFR mutation status can be used to predict EGFR-TKI therapy outcome in advanced NSCLC, with EGFR mutation status being the most powerful feature, followed by tumor histology.

3.2 Drug Response Curves

The drug-/dose-response curve shows the response of an organism as a function of exposure to a drug after a certain time [106]. The drug's half maximum inhibitory concentration IC_{50} [107] 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 [108].

Jang *et al.* [109] carried out a systematic assessment of analytical methods for drug sensitivity prediction using Cancer Cell Line Encyclopedia (CCLE) [110] data. The cancer cell lines are most widely used models for researching cancer biology, confirming cancer targets, and determining therapeutic efficacy. They evaluated more than 110,00 models, based on multifactorial experiment design. They found that the model input data, including molecular features and compound are an important factor for model performance, followed by the type of algorithms.

In another study, Zou *et al.* [57] 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 [111]. Multilinear Principal Component Analysis (MPCA) was used for dimensionality reduction and feature selection. Target projection pursuit [112] was used to show drug sensitivities. The findings show that the IFP features of EGFR-mutant complexes and MPCA based tensors are useful candidates for prediction of drug sensitivity. They used five classifiers for predicting the drug sensitivity and achieved greater than 90% accuracy.

These studies provide a 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 dynamic distance, energy and personal features may be combined to construct a discriminative composite feature, for an efficient personalized drug response prediction model.

3.3 Drug Synergy

Treatment combinations or synergistic combinations are critical for combating drug resistance and minimising recurrence caused by a small number of cancer cells that persist [113]. Due to time limits and the possibility for harmful exposure to toxic combinations without boosting efficacy,

trial and error combination design has limited relevance in the clinic. Adaptive trials via reinforcement learning [114] are one possible path for a context, where the goal is to examine a confined set of possibilities in order to build an ideal treatment plan for a patient. A probabilistic ranking based method is proposed in [115], that identifies which drug combinations can lead to drug resistance. This work learns more complex yet effective strategies in terms of survival, than what is currently available in the clinic. Nevertheless, it is unclear how to avoid the possible hazards of reinforcement learning exploration space, and requires further investigations.

4 Data-Driven Methods

4.1 Deep Learning

Deep learning has shown remarkable results in solving many challenging problems of computer vision, natural language processing, financial model analysis, and bioinformatics [116], [117]. Deep neural networks are considered as universal approximator, and have the potential to learn the complex and evolutionary relationship in high dimensional datasets. Predicting the 3D structure of a protein solely from the 1D amino acid sequence is a very complex and challenging problem in computational biology [118]. Alphafold2 [119], a deep learning system for predicting the 3D structure of protein, achieves accuracy competitive with experiment methods. The key idea behind AlphaFold2 is the combination of bioinformatics and physical approaches. It learns from PDB data with minimal handcrafted features using a physical and geometric inductive bias (for example, AlphaFold builds hydrogen bonds effectively without a hydrogen bond score function). By adding unique neural network topologies and training processes based on the evolutionary, physical, and geometric constraints of protein structures, AlphaFold dramatically enhances the accuracy of structure prediction. AlphaFold2 uses multiple sequence alignments (MSAs) and pairwise features to jointly embed them. A new output representation and associated loss that enables accurate end-to-end structure prediction. A new equivariant attention architecture, and intermediate losses are used to achieve iterative refinement of predictions.

The use of deep learning has also shown promise in pharmaceutical research such as bio-activity prediction, and drug discovery [94]. The convolutional neural network (CNN) has provided excellent results in image recognition problems, CNN can also be used to design protein-ligand scoring functions. .

A deep learning model is proposed to predict the mutation status form CT-scan lung images in [120]. 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 [121] using transfer learning [122], 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. This shows that the features extracted from CT scans of lung cancer are related to gene expression patterns. Image-driven methods can provide additional information, that is complementary to biopsies [123].

Deep generative models [124] can be used to explore the high dimensional chemical search space, and to design novel molecules or compounds. Every dataset, no matter how big is actually a subset of the universe. The generative models can learn the probability distribution of the chemical search space, based on a training dataset. After extracting the representative features, the model can generate new chemical, EGFR inhibitors or structures by sampling from the probability distribution.

Deep learning can also be used for virtual screening, and to design novel EGFR inhibitors [125]. Recently, a multi-input deep neural network based on attention mechanism [126] is proposed for biological activity prediction of EGFR inhibitors [67]. The proposed model uses the dataset [127] of 3492 compounds labeled as inhibitors or non-inhibitors. The method uses SMILES (Simplified molecule input line entry system [128], 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 [68] 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 [129]. Such a network can be exploited for drug resistance prediction model with small number of clinical samples.

Understanding the relationship between a cell's genome and its phenotype is a critical issue in precision medicine [130]. Deep learning algorithms are being used to create models that link genomic variations to phenotypes. Nonetheless, genotype-to-phenotype prediction poses significant challenges for deep learning algorithms, limiting their application in this context. Algorithms that produce interpretable models are preferable because models must rely on a decision process that can be evaluated by domain experts. Such models can be used to predict the impacts of EGFR mutations on the protein structure and drug response.

4.2 Big Data Analytics

Big data analysis and Artificial Intelligence are changing the drug discovery pipeline. The Cancer Therapeutics Response Portal (CTRP) [131] provides a resource to develop new insights into small molecule action mechanism, generate new hypothesis, and personalized drug discovery based on predictive biomarkers. Several big data projects, including Cancer Cell Line Encyclopedia (CCLE) [110], and Genomics of Drug Sensitivity in Cancer (GDSC) [132] have performed large scale molecule screens on panels of hundreds of molecularly characterized cancer cell lines. These projects

demonstrate the potential of modern machine learning algorithms to develop drug response predictors based on molecular profiling measurements [133].

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 or response. In fact, independent procedures analyzing the clinical data may reach different conclusions, aiming to answer the same biological question [134]. Another problem is the inconsistency between datasets and missing clinical information [135].

One solution to the missing value problem is to use data imputation techniques [136]. Consider the well-known Netflix recommendation problem, because users only rate a few items from a large sets of items, one would like to deduce their preference for unrated items, using matrix factorization [137]. Similarly, in a large protein-drug complex, only a subset of genetic mutations affects the protein structure and function. Thus, such problems are also well suited for low rank matrix completion [138]. Other data imputation techniques include, mean value substitution, where missing value is replaced with the variable mean score [139], least square estimation [140], which imputes each missing value with a linear combination of similar genes [141]. Machine learning approaches coupled with communication infrastructure [142] also provides an opportunity for treating patients remotely [143], [144].

4.3 Applications of Statistics and Permutation

In Bioinformatics, Big multi-omics data often consists of a large number of features and a small number of samples [145]. The data from genome-wide association studies (GWAS), for example, contains at least thousands of samples and hundreds of thousands of single nucleotide polymorphisms (SNPs) [146]. Furthermore, depending on whether they are from genomes, proteomics, transcriptomics or metabolomics, features from multi-omics data have their own unique characteristics. Standard statistical analyses based on parametric assumptions may be failed to yield correct asymptotic results, due to these diverse high-dimensional data [147].

To address this issue, Permutation tests [148] can be a means to precisely examine multi−omics data, as they are distribution-free and flexible to apply. Under the strong null hypothesis, that a set of genetic variants/mutations has no effect on the result, a permutation test provides a simple technique to determine the sampling distribution for any test statistic. The null hypothesis in permutations testing is that the labels used to allocate samples to classes are interchangeable [149], and shuffling and randomization has no effect on probability distribution. The P values are used to determine the significance of a permutation test. The P-value is calculated by conducting all possible permutations and calculating the fraction of permutation values that are at least as extreme as the unpermuted data's test statistic. In practice, however, performing all potential permutations is (by far) not viable, due to massive computations and permutations. As a result, the P-value is approximated by computing a small number of permutations, say N, and then computing the fraction of those N values that are at least as extreme as the test statistic. The permutation test can be used to check the quality assessment of 3D drug protein-complexes, predicted by computational tools, effects of different EGFR mutations on drug response or drug resistance, or different combination of drug responses, to better plan the treatment strategies.

5 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 [150]. 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 [134]. 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 much larger or variable human genes size. Computational methods might not produce robust results, and non-algorithmic solutions are necessary. The high dimensionality problem can be addressed by using feature filtering or feature selection techniques, or sparse principal component analysis [151]. Another solution is the data integration [152].

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 [153]. Since most of the studies are connected to patient's health, the logical reasoning of the model will provide useful information [94]. Many more handcrafted features are required to increase trust in interpretability. 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 [154].

6 Conclusion and Future Work

In this paper, we explained the drug resistance mechanisms 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 factors for analyzing the drug response. The stability is the change in the net energy in the unfolded and folded states [155]. It will be interesting to see whether a state-transition matrix can be proposed to model the changing states of a protein [156]. 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 [157]. However, there is still a need to develop a robust system for predicting drug sensitivity for patients with EGFR mutations, as well as for clinical trial design. In addition, AlphaFold2 can be used to predict or verify the mutant EGFR structures. 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-mutants using CT-scan images and deep learning provides non-invasive and an easy solution [158]. By using feature visualization or other method, interesting insights can be obtained. Despite of all these studies and findings, there are still unknown reasons for 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.

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 (MN) [159], a neural network-based architecture built with an attention mechanism and memory (non-parametric structure). The advantage of MN is that it can easily map a small number of labelled support set and an unlabelled example to its label, without the need of fine-tuning to adapt new class types. MN also speeds up the learning process by training the model using only few samples per class.

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 [160] 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 [161]. AceCloud [162] is another cloud-based MD program, capable of running hundreds of MD simulations. Anton [163] is a special-purpose system for MD simulations, using application specific integrated circuits (ASIC), for observing large scale conformational dynamics of protein system. Moreover, the development of efficient open source libraries for the analysis of molecular dynamics trajectories, such as MDAnalysis [164], MDTraj [165], and online web applications [166] provides opportunity for flexible and fast framework for complex analysis programs. In addition, online freely drug-databases, such as drug-bank [167] 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.

Machine learning approaches, applied to biological, clinical, and chemical data may improve the drug discovery and analysis pipeline, as well as treatment strategies. Recently, new antibiotics have been discovered from the pool of 100 million molecules using deep learning [133]. We believe that simulation-based studies, coupled with data analysis methods will play a vital role in future cancer research [168]. Several Bio-tech companies are using artificial intelligence to enhance the drug development process, from candidate screening to trial management [169]. We speculate that decades from now, personalized drug models will be used in treatment of NSCLC patients. We further hope that the combination of computational methods and clinical studies can provide useful recommendations, for precision and personalized medicine.

REFERENCES

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] D. R. Youlden, S. M. Cramb, and P. D. Baade, "The international epidemiology of lung cancer: Geographical distribution and secular trends," *J. Thoracic Oncol.*, vol. 3, no. 8, pp. 819–831, 2008.

[3] S. V. Sharma *et al.*, "Epidermal growth factor receptor mutations in lung cancer," *Nature Rev. Cancer*, vol. 7, no. 3, 2007, Art. no. 169.

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

[5] J. Li *et al.*, "Association of variant ABCG2 and the pharmacokinetics of epidermal growth factor receptor tyrosine kinase inhibitors in cancer patients," *Cancer Biol. Ther.*, vol. 6, no. 3, pp. 432–438, 2007.

[6] J. M. Stommel *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.

[7] 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.

[8] J.-C. Soria *et al.*, "Osimertinib in untreated EGFR-mutated advanced non–small-cell lung cancer," *New Engl. J. Med.*, vol. 378, no. 2, pp. 113–125, 2018.

[9] J. Tabernero *et al.*, "Phase II trial of cetuximab in combination with fluorouracil, leucovorin, and oxaliplatin in the first-line treatment of metastatic colorectal cancer," *J. Clin. Oncol.*, vol. 25, no. 33, pp. 5225–5232, 2007.

[10] S. Wang, Y. Song, and D. Liu, "EAI045: The fourth-generation EGFR inhibitor overcoming T790M and C797S resistance," *Cancer Lett.*, vol. 385, pp. 51–54, 2017.

[11] S. Bamford *et al.*, "The COSMIC (catalogue of somatic mutations in cancer) database and website," *Brit. J. Cancer*, vol. 91, no. 2, 2004, Art. no. 355.

[12] W. Pao *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," *Proc. Nature Acad. Sci. USA*, vol. 101, no. 36, pp. 13306–13311, 2004.

[13] J. P. Robichaux *et al.*, "Structure-based classification predicts drug response in EGFR-mutant NSCLC," *Nature*, vol. 597, no. 7878, pp. 732–737, 2021.

[14] X. Sun and B. Hu, "Mathematical modeling and computational prediction of cancer drug resistance," *Brief. Bioinf.*, vol. 19, no. 6, pp. 1382–1399, 2018.

[15] R. Brown, E. Curry, L. Magnani, C. S. Wilhelm-Benartzi , and J. Borley, "Poised epigenetic states and acquired drug resistance in cancer," *Nature Rev. Cancer*, vol. 14, no. 11, pp. 747–753, 2014.

[16] H.-J. Lee, G. Zhuang, Y. Cao, P. Du, H.-J. Kim, and J. Settleman, "Drug resistance via feedback activation of Stat3 in oncogene-addicted cancer cells," *Cancer Cell*, vol. 26, no. 2, pp. 207–221, 2014.

[17] A. Singh and J. Settleman, "EMT, cancer stem cells and drug resistance: An emerging axis of evil in the war on cancer," *Oncogene*, vol. 29, no. 34, pp. 4741–4751, 2010.

[18] M. Fukuoka *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)," *J. Clin. Oncol.*, vol. 29, no. 21, pp. 2866–2874, 2011.

[19] R. Rosell *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," *Lancet Oncol.*, vol. 13, no. 3, pp. 239–246, 2012.

[20] J. C. H. Yang *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," *Lancet Oncol.*, vol. 16, no. 2, pp. 141–151, 2015.

[21] K. S. Thress *et al.*, "Acquired EGFR C797S mutation mediates resistance to AZD9291 in non–small cell lung cancer harboring EGFR T790M," *Nature Med.*, vol. 21, no. 6, 2015, Art. no. 560.

[22] G. Oxnard *et al.*, "TATTON: A multi-arm, phase IB trial of osimertinib combined with selumetinib, savolitinib or durvalumab in EGFR-mutant lung cancer," *Ann. Oncol.*, vol. 31, pp. 507–516, 2020.

[23] K. L. Reckamp *et al.*, "A highly sensitive and quantitative test platform for detection of NSCLC EGFR mutations in urine and plasma," *J. Thoracic Oncol.*, vol. 11, no. 10, pp. 1690–1700, 2016.

[24] J. G. Paez *et al.*, "EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy," *Science*, vol. 304, no. 5676, pp. 1497–1500, 2004.

[25] Y. Shi *et al.*, "A prospective, molecular epidemiology study of EGFR mutations in asian patients with advanced non–small-cell lung cancer of adenocarcinoma histology (PIONEER)," *J. Thoracic Oncol.*, vol. 9, no. 2, pp. 154–162, 2014.

[26] R. Qureshi, M. Zhu, and H. Yan, "Visualization of protein-drug interactions for the analysis of lung cancer drug resistance," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1839–1848, May 2021.

[27] 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.

[28] 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.

[29] E. P. Allain, M. Rouleau, E. Lévesque, and C. Guillemette, "Emerging roles for UDP-glucuronosyltransferases in drug resistance and cancer progression," *Brit. J. Cancer*, vol. 122, pp. 1277–1287, 2020.

[30] G. da Cunha Santos, F. A. Shepherd, and M. S. Tsao, "EGFR mutations and lung cancer," *Annu. Rev. Pathol., Mechanisms Dis.*, vol. 6, pp. 49–69, 2011.

[31] G. Zhang *et al.*, "Automatic nodule detection for lung cancer in CT images: A review," *Comput. Biol. Med.*, vol. 103, pp. 287–300, 2018.

[32] Y. Gao *et al.*, "Nanotechnology-based intelligent drug design for cancer metastasis treatment," *Biotechnol. Adv.*, vol. 32, no. 4, pp. 761–777, 2014.

[33] C. D. Fjell *et al.*, "Designing antimicrobial peptides: Form follows function," *Nature Rev. Drug Discov.*, vol. 11, no. 1, 2012, Art. no. 37.

[34] M. Hassan Baig *et al.*, "Computer aided drug design: Success and limitations," *Curr. Pharm. Des.*, vol. 22, no. 5, pp. 572–581, 2016.

[35] B. Knapp *et al.*, "Current status and future challenges in T-cell receptor/peptide/MHC molecular dynamics simulations," *Brief. Bioinf.*, vol. 16, no. 6, pp. 1035–1044, 2015.

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

[37] D. Roccatano, A. Barthel, and M. Zacharias, "Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation," *Biopolym., Original Res. Biomol.*, vol. 85, no. 5/6, pp. 407–421, 2007.

[38] I. Tinoco Jr and J.-D. Wen, "Simulation and analysis of single-ribosome translation," *Phys. Biol.*, vol. 6, no. 2, 2009, Art. no. 025006.

[39] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: Advances and applications," *Adv. Appl. Bioinf. Chem.*, vol. 8, 2015, Art. no. 37.

[40] Y. Shan *et al.*, "Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization," *Cell*, vol. 149, no. 4, pp. 860–870, 2012.

[41] Y. Shan *et al.*, "Transitions to catalytically inactive conformations in EGFR kinase," *Proc. Nature Acad. Sci. USA*, vol. 110, no. 18, pp. 7270–7275, 2013.

[42] S. Wan, D. W. Wright, and P. V. Coveney, "Mechanism of drug efficacy within the EGF receptor revealed by microsecond molecular dynamics simulation," *Mol. Cancer Ther.*, vol. 11, no. 11, pp. 2394–2400, 2012.

[43] 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," *J. Comput. Chem.*, vol. 32, no. 13, pp. 2843–2852, 2011.

[44] Z. Zhao, L. Xie, and P. E. Bourne, "Structural insights into characterizing binding sites in epidermal growth factor receptor kinase mutants," *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 453–462, 2018.

[45] Q.-H. Liao *et al.*, "Docking and molecular dynamics study on the inhibitory activity of novel inhibitors on epidermal growth factor receptor (EGFR)," *Med. Chem.*, vol. 7, no. 1, pp. 24–31, 2011.

[46] N. Huang, B. K. Shoichet, and J. J. Irwin, "Benchmarking sets for molecular docking," *J. Med. Chem.*, vol. 49, no. 23, pp. 6789–6801, 2006.

[47] B. Rajith *et al.*, "Structural signature of the G719S-T790M double mutation in the EGFR kinase domain and its response to inhibitors," *Sci. Rep.*, vol. 4, 2014, Art. no. 5868.

[48] S. Ikemura *et al.*, "Molecular dynamics simulation-guided drug sensitivity prediction for lung cancer with rare EGFR mutations," *Proc. Nature Acad. Sci. USA*, vol. 116, no. 20, pp. 10025–10030, 2019.

[49] P. W. Rose *et al.*, "The RCSB protein data bank: Integrative view of protein, gene and 3D structural information," *Nucleic Acids Res.*, vol. 45, 2016, Art. no. gkw1000.

[50] L. Ma *et al.*, "EGFR mutant structural database: Computationally predicted 3D structures and the corresponding binding free energies with gefitinib and erlotinib," *BMC Bioinf.*, vol. 16, no. 1, 2015, Art. no. 85.

[51] 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," *J. Mol. Model.*, vol. 18, no. 7, pp. 2971–2979, 2012.

[52] D. Eisenberg, R. Lüthy, and J. U. Bowie, "VERIFY3D: Assessment of protein models with three-dimensional profiles," *Methods Enzymol.*, vol. 277, pp. 396–404, 1997.

[53] 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 J. Biol. Sci.*, vol. 14, no. 2, 2014, Art. no. 155.

[54] D. D. Wang *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, 2015, Art. no. e0128360.

[55] N. Kureshi, S. S. R. Abidi, and C. Blouin, "A predictive model for personalized therapeutic interventions in non-small cell lung cancer," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 424–431, Jan. 2016.

[56] A. Ghosh and H. Yan, "Hydrogen bond analysis of the EGFR-ErbB3 heterodimer related to non-small cell lung cancer and drug resistance," *J. Theor. Biol.*, vol. 464, pp. 63–71, 2019.

[57] 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 Bioinf.*, vol. 19, no. 1, 2018, Art. no. 88.

[58] C.-H. Yun *et al.*, "The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP," *Proc. Nature Acad. Sci. USA*, vol. 105, no. 6, pp. 2070–2075, 2008.

[59] S. Goyal *et al.*, "Structural investigations of T854A mutation in EGFR and identification of novel inhibitors using structure activity relationships," *BMC Genomic.*, vol. 16, no. 5, 2015, Art. no. S8.

[60] 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, Struct. Function Bioinf.*, vol. 65, no. 2, pp. 331–346, 2006.

[61] R. Qureshi, M. Nawaz, A. Ghosh, and H. Yan, "Parametric models for understanding atomic trajectories in different domains of lung cancer causing protein," *IEEE Access*, vol. 7, pp. 67551–67563, 2019.

[62] S. A. Foster *et al.*, "Activation mechanism of oncogenic deletion mutations in BRAF, EGFR, and HER2," *Cancer Cell*, vol. 29, no. 4, pp. 477–493, 2016.

[63] B. Zou *et al.*, "Deciphering mechanisms of acquired T790M mutation after EGFR inhibitors for NSCLC by computational simulations," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 6595.

[64] D. D. Wang *et al.*, "Personalized prediction of EGFR mutation-induced drug resistance in lung cancer," *Sci. Rep.*, vol. 3, 2013, Art. no. 2855.

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

[66] Q. Shao and W. Zhu, "Ligand binding effects on the activation of the EGFR extracellular domain," *Phys. Chem. Chem. Phys.*, vol. 21, no. 15, pp. 8141–8151, 2019.

[67] H. N. Pham and T. H. Le, "Attention-based multi-input deep learning architecture for biological activity prediction: An application in EGFR inhibitors," in *Proc. 11th Int. Conf. Knowl. Syst. Eng.*, 2019, pp. 1–9.

[68] M. Joo *et al.*, "A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients," *Int. J. Mol. Sci.*, vol. 20, no. 24, 2019, Art. no. 6276.

[69] L. Ou-Yang *et al.*, "Predicting the impacts of mutations on protein-ligand affinity based on molecular dynamics simulations and machine learning methods," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 439–454, 2020.

[70] I. Massova and P. A. Kollman, "Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding," *Perspectives Drug Discov. Des.*, vol. 18, no. 1, pp. 113–135, 2000.

[71] A. R. Brice and B. N. Dominy, "Analyzing the robustness of the MM/PBSA free energy calculation method: Application to DNA conformational transitions," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1431–1440, 2011.

[72] D. D. Wang *et al.*, "Selectivity profile of afatinib for EGFR-mutated non-small-cell lung cancer," *Mol. BioSyst.*, vol. 12, no. 5, pp. 1552–1563, 2016.

[73] C. C. Jaffe, "Measures of response: RECIST, WHO, and new alternatives," *J. Clin. Oncol.*, vol. 24, no. 20, pp. 3245–3251, 2006.

[74] W. J. Gullick, "The type 1 growth factor receptors and their ligands considered as a complex system," *Endocrine-Related Cancer*, vol. 8, no. 2, pp. 75–82, 2001.

[75] A. B. Singh and R. C. Harris, "Autocrine, paracrine and juxtacrine signaling by EGFR ligands," *Cell. Signalling*, vol. 17, no. 10, pp. 1183–1193, 2005.

[76] D. Dhar *et al.*, "Liver cancer initiation requires p53 inhibition by CD44-enhanced growth factor signaling," *Cancer Cell*, vol. 33, no. 6, pp. 1061–1077, 2018.

[77] 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 Res.*, vol. 66, no. 20, pp. 10100–10111, 2006.

[78] 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," *Mol. Biol. Cell*, vol. 13, no. 7, pp. 2547–2557, 2002.

[79] J. Cairns, B. L. Fridley, G. D. Jenkins, Y. Zhuang, J. Yu, and L. Wang, "Differential roles of ERRFI1 in EGFR and AKT pathway regulation affect cancer proliferation," *EMBO Rep.*, vol. 19, no. 3, 2018, Art. no. e44767.

[80] 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," *Int. J. Cancer*, vol. 94, no. 6, pp. 774–782, 2001.

[81] S. J. Klempner, A. P. Myers, and L. C. Cantley, "What a tangled web we weave: Emerging resistance mechanisms to inhibition of the phosphoinositide 3-kinase pathway," *Cancer Discov.*, vol. 3, no. 12, pp. 1345–1354, 2013.

[82] 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.

[83] M. Guix *et al.*, "Acquired resistance to EGFR tyrosine kinase inhibitors in cancer cells is mediated by loss of IGF-binding proteins," *J. Clin. Invest.*, vol. 118, no. 7, pp. 2609–2619, 2008.

[84] A. Maynard *et al.*, "Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing," *Cell*, vol. 182, no. 5, pp. 1232–1251, 2020.

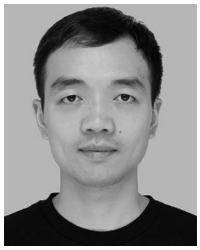
[85] J. Zhang, M. Guan, Q. Wang, J. Zhang, T. Zhou, and X. Sun, "Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas," *Brief. Bioinf.*, vol. 21, no. 3, pp. 1080–1097, 2020.

[86] A. Dixit and G. M. Verkhivker, "Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases," *PLoS Comput. Biol.*, vol. 7, no. 10, 2011, Art. no. e1002179.

[87] R. Qureshi, A. Ghosh, and H. Yan, "Correlated motions and dynamics in different domains of EGFR with L858R and T790M mutations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, May 18, 2020, doi: 10.1109/TCBB.2020.2995569.
[88] N. M. Goodey and S. J. Benkovic, "Allosteric regulation and catalysis emerge via a common route," *Nature Chem. Biol.*, vol. 4, no. 8, 2008, Art. no. 474.
[89] W. Zheng, B. R. Brooks, and D. Thirumalai, "Allosteric transitions in biological nanomachines are described by robust normal modes of elastic networks," *Curr. Protein Peptide Sci.*, vol. 10, no. 2, pp. 128–132, 2009.
[90] S. L. Dixon *et al.*, "PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results," *J. Comput.-Aided Mol. Des.*, vol. 20, no. 10/11, pp. 647–671, 2006.
[91] S. C. Peter *et al.*, "Quantitative structure-activity relationship (QSAR): Modeling approaches to biological applications," in *Encyclopedia Bioinf. Comput. Biol.*, 2019, pp. 661–676.
[92] Y. Jiang, Q.-J. Han, and J. Zhang, "Hepatocellular carcinoma: Mechanisms of progression and immunotherapy," *World J. Gastroenterol.*, vol. 25, no. 25, 2019, Art. no. 3151.
[93] C. Minnelli, E. Laudadio, G. Mobbili, and R. Galeazzi, "Conformational insight on WT-and mutated-EGFR receptor activation and inhibition by epigallocatechin-3-gallate: Over a rational basis for the design of selective non-small-cell lung anticancer agents," *Int. J. Mol. Sci.*, vol. 21, no. 5, 2020, Art. no. 1721.
[94] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Brief. Bioinf.*, vol. 18, no. 5, pp. 851–869, 2017.
[95] N. J. Rollins *et al.*, "Inferring protein 3D structure from deep mutation scans," *Nature Genet.*, vol. 51, no. 7, 2019, Art. no. 1170.
[96] B. Yadav, "Quantitative modeling and analysis of drug screening data for personalized cancer medicine," University of Helsinki, Helsinki, Finland, 2017.
[97] X. Xu, M. C. Farach-Carson , and X. Jia, "Three-dimensional in vitro tumor models for cancer research and drug evaluation," *Biotechnol. Adv.*, vol. 32, no. 7, pp. 1256–1268, 2014.
[98] R. Qureshi, M. Nawaz, and H. Yan, "Personalized drug-response prediction model for lung cancer patients using machine learning," *IEEE Techrxiv*, pp. 1–12, 2020.
[99] 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.
[100] 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*, vol. 9, pp. 5704–5715, 2021.
[101] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Trans. Graph.*, vol. 13, no. 1, pp. 43–72, 1994.
[102] W. Zhou, H. Yan, and Q. Hao, "Analysis of surface structures of hydrogen bonding in protein–ligand interactions using the alpha shape model," *Chem. Phys. Lett.*, vol. 545, pp. 125–131, 2012.
[103] D. A. Case *et al.*, "The amber biomolecular simulation programs," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1668–1688, 2005.
[104] A. Fabri and S. Pion, "CGAL: The computational geometry algorithms library," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2009.
[105] E.-S. Lee *et al.*, "Prediction of recurrence-free survival in postoperative non–small cell lung cancer patients by using an integrated model of clinical information and gene expression," *Clin. Cancer Res.*, vol. 14, no. 22, pp. 7397–7404, 2008.
[106] L. Wang, H. L. McLeod, and R. M. Weinshilboum, "Genomics and drug response," *New England J. Med.*, vol. 364, no. 12, pp. 1144–1153, 2011.
[107] A. Bag, "DFT based computational methodology of IC50 prediction," *Curr. Comput.-Aided Drug Des.*, vol. 17, pp. 244–253, 2020.
[108] J. Sebaugh, "Guidelines for accurate EC50/IC50 estimation," *Pharm. Statist.*, vol. 10, no. 2, pp. 128–134, 2011.
[109] 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 *Proc. Pacific Symp. Biocomput.*, 2014, pp. 63–74.
[110] J. Barretina *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.
[111] 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 *Proc. Adv. Protein Chem. Structural Biol.*, vol. 96, pp. 1–37, 2014.
[112] S.-S. Chiang, C.-I. Chang, and I. W. Ginsberg, "Unsupervised target detection in hyperspectral images using projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1380–1391, Jul. 2001.
[113] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: Challenges and recent progress," *NPJ Precis. Oncol.*, vol. 4, no. 1, pp. 1–10, 2020.
[114] M. A. Wiering and M. Van Otterlo, "Reinforcement learning," *Adaptation Learn. Optim.*, Springer, vol. 12, no. 3, 2012.
[115] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, "Contextual bandits for adapting treatment in a mouse model of de Novo Carcinogenesis," in *Proc. 3rd Mach. Learn. Healthcare Conf.*, 2018, pp. 67–82.
[116] A. Voulodimos *et al.*, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–14, 2018.
[117] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, 2018.
[118] J. Moult, "A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction," *Curr. Opin. Struct. Biol.*, vol. 15, no. 3, pp. 285–289, 2005.
[119] J. Jumper *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[120] S. Wang *et al.*, "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *Eur. Respirat. J.*, vol. 53, no. 3, 2019, Art. no. 1800986.
[121] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," 2014, *arXiv:1404.1869*.
[122] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[123] H. Itakura *et al.*, "Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities," *Sci. Transl. Med.*, vol. 7, no. 303, pp. 303ra138–303ra138, 2015.
[124] S. Kang and K. Cho, "Conditional molecular design with deep generative models," *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 43–52, 2018.
[125] T. Unterthiner *et al.*, "Deep learning as an opportunity in virtual screening," in *Proc. Deep Learn. Workshop NIPS*, 2014, pp. 1–9.
[126] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Con. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[127] 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," *Biol. Direct*, vol. 10, no. 1, 2015, Art. no. 10.
[128] A. A. Toropov and E. Benfenati, "SMILES in QSPR/QSAR modeling: Results and perspectives," *Curr. Drug Discov. Technol.*, vol. 4, no. 2, pp. 77–116, 2007.
[129] L. Fei-Fei , R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[130] A. Drouin, G. Letarte, F. Raymond, M. Marchand, J. Corbeil, and F. Laviolette, "Interpretable genotype-to-phenotype classifiers with performance guarantees," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, 2019.
[131] 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, 2016, Art. no. 51619.
[132] W. Yang *et al.*, "Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D955–D961, 2012.
[133] A. Valavanidis, "Scientific reviews artificial intelligence application with machine-learning algorithm identified a powerful broad-spectrum antibiotic," MIT News. [Online]. Available: https://news.mit.edu/2020/artificial-intelligence-identifies-new-antibiotic-0220

[134] W. Hugo *et al.*, "Non-genomic and immune evolution of melanoma acquiring MAPKi resistance," *Cell*, vol. 162, no. 6, pp. 1271–1285, 2015.
[135] P. Jiang, W. R. Sellers, and X. S. Liu, "Big data approaches for modeling response and resistance to cancer drugs," *Annu. Rev. Biomed. Data Sci.*, vol. 1, pp. 1–27, 2018.
[136] J. M. Jerez *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–115, 2010.
[137] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
[138] A. Kapur, K. Marwah, and G. Alterovitz, "Gene expression prediction using low-rank matrix completion," *BMC Bioinf.*, vol. 17, no. 1, pp. 1–13, 2016.
[139] T. Verbovšek, "A comparison of parameters below the limit of detection in geochemical analyses by substitution methods Primerjava ocenitev parametrov pod mejo določljivosti pri geokemičnih analizah z metodo nadomeščanja," *RMZ–Mater. Geoenviron.*, vol. 58, no. 4, pp. 393–404, 2011.
[140] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: Local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2005.
[141] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell RNA-sequencing data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 376–389, Mar./Apr. 2020.
[142] V. Oleshchuk and R. Fensli, "Remote patient monitoring within a future 5G infrastructure," *Wireless Pers. Commun.*, vol. 57, no. 3, pp. 431–439, 2011.
[143] S. Sengan *et al.*, "Medical information retrieval systems for e-Health care records using fuzzy based machine learning model," *Microprocessors Microsyst.*, Oct. 2020, Art. no. 103344.
[144] R. Qureshi, M. Uzair, and K. Khurshid, "Multistage adaptive filter for ECG signal processing," in *Proc. Int. Conf. Commun. Comput. Digit. Syst.*, 2017, pp. 363–368.
[145] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," *Bioinf. Biol. Insights*, vol. 14, 2020, Art. no. 1177932219899051.
[146] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *New England J. Med.*, vol. 363, no. 2, pp. 166–176, 2010.
[147] S. Leem, I. Huh, and T. Park, "Enhanced permutation tests via multiple pruning," *Front. Genetics*, vol. 11, 2020, Art. no. 509.
[148] E. Venkatraman, "A permutation test to compare receiver operating characteristic curves," *Biometrics*, vol. 56, no. 4, pp. 1134–1138, 2000.
[149] E. Edgington and P. Onghena, *Randomization Tests*. Boca Raton, FL, USA: CRC Press, 2007.
[150] J. Lee *et al.*, "Constant pH molecular dynamics in explicit solvent with enveloping distribution sampling and hamiltonian exchange," *J. Chem. Theory Comput.*, vol. 10, no. 7, pp. 2738–2750, 2014.
[151] M. C. Rendleman *et al.*, "Machine learning with the TCGA-HNSC dataset: Improving usability by addressing inconsistency, sparsity, and high-dimensionality," *BMC Bioinf.*, vol. 20, no. 1, 2019, Art. no. 339.
[152] E. Cerami *et al.*, "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discov.*, vol. 2, pp. 401–404, 2012.
[153] A. A. Kalinin *et al.*, "Deep learning in pharmacogenomics: From gene regulation to patient stratification," *Pharmacogenomics*, vol. 19, no. 7, pp. 629–650, 2018.
[154] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.
[155] W. J. Becktel and J. A. Schellman, "Protein stability curves," *Biopolymers: Original Res. Biomol.*, vol. 26, no. 11, pp. 1859–1877, 1987.
[156] C.-C. Chang, M.-S. Cheng, Y.-C. Su, and L.-S. Kan, "A first-order-like state transition for recombinant protein folding," *J. Biomol. Struct. Dyn.*, vol. 21, no. 2, pp. 247–255, 2003.
[157] P. T. Harrison, S. Vyse, and P. H. Huang, "Rare epidermal growth factor receptor (EGFR) mutations in non-small cell lung cancer," *Seminars Cancer Biol.*, vol. 61, pp. 167–179, 2020.
[158] S. Wang *et al.*, "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *Eur. Respiratory J.*, vol. 53, no. 3, 2019, Art. no. 1800986.
[159] O. Vinyals *et al.*, "Matching networks for one shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3630–3638.
[160] H. Wang *et al.*, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Comput. Phys. Commun.*, vol. 228, pp. 178–184, 2018.
[161] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, "ACEMD: Accelerating biomolecular dynamics in the microsecond time scale," *J. Chem. Theory Comput.*, vol. 5, no. 6, pp. 1632–1639, 2009.
[162] M. J. Harvey and G. De Fabritiis , "AceCloud: Molecular dynamics simulations in the cloud," 2015. [Online]. Available: https://www.acellera.com/cloud-molecular-dynamics-simulation-acecloud/
[163] D. E. Shaw *et al.*, "Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2014, pp. 41–53.
[164] R. J. Gowers *et al.*, "MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations," Los Alamos National Lab. (LANL), Los Alamos, NM, USA, 2019. [Online]. Available: https://www.mdanalysis.org/
[165] R. T. McGibbon *et al.*, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," *Biophys. J.*, vol. 109, no. 8, pp. 1528–1532, 2015.
[166] 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.
[167] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the drugbank database for 2018, " *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018.
[168] J. Li *et al.*, "A survey of current trends in computational drug repositioning," *Brief. Bioinf.*, vol. 17, no. 1, pp. 2–12, 2015.
[169] R. Kneller, "The importance of new companies for drug discovery: Origins of a decade of new drugs," *Nature Rev. Drug Discov.*, vol. 9, no. 11, pp. 867–882, 2010.

Rizwan Qureshi received the BE degree in electronic engineering from the Mehran University of Engineering & Technology, Jamshoro, Pakistan, in 2010, the master of science degree in electrical engineering from the Institute of Space Technology, Islamabad, Pakistan, in 2015, and the PhD degree in electrical engineering from the City University of Hong Kong, Hong Kong, in 2021. He is currently an assistant professor with the Fast School of Computing, National University of Computer and Emerging Sciences, Karachi, Pakistan. Before joining City University, he was a lecturer with Electrical Engineering Department, COMSATS University Islamabad, Wah Campus, Pakistan. His research interests include bioinformatics, signal and image processing, machine learning, and spectral imaging. He also served as a reviewer for a number of international conferences and reputed journals.



Bin Zou received the PhD degree in electrical engineering from the City University of Hong Kong, Hong Kong. He is currently a senior engineer with BGI-shenzhen, China. His research interests include computational biology, signal and image processing, single-cell spatial transcriptomics. Currently, his research focuses on the development and application of innovative computational and analytical approaches and deep learning techniques to solve problems in the field of healthcare & life science.

Tanvir Alam is currently an assistant professor with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. Among his notable research works are on the transcription regulation of non-coding RNAs and their roles in different diseases. His research work also centered around the application of artificial intelligence (AI) on the diagnosis and prognosis of communicable and non-communicable diseases. He is a member of FANTOM Consortium. He also served as a reviewer for a number of international conferences and reputed journals.

Jia Wu received the PhD degree from the University of Pittsburgh, Pittsburgh, Pennsylvania, in 2013. He is currently an assistant professor (tenure-track) with the Department of Imaging Physics, Division of Diagnostic Imaging, University of Texas MD Anderson Cancer Center, Houston, Texas. He received postdoctoral training from Stanford University, Palo Alto, California, and the Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania. His research is focused on the development and application of innovative computational and analytical approaches to improve the diagnosis, treatment, early detection, and prevention of cancer. He received Pathway to Independence Award, NIH/NCI, 2018 and the UT Rising STARs Award, The University of Texas System, 2021. He also won the Research Career Accelerator Program, Stanford Center for Clinical & Translational Research & Education in 2018 and Rexanna's Foundation Award for Fighting Lung Cancer in 2021. He has published in several prestigious journals including the *Nature Machine Intelligence*.

Victor H. F. Lee received the graduate degree from the University of Hong Kong, Hong Kong, in 2002. He is currently clinical associate professor with the Department of Clinical Oncology, University of Hong Kong. After internship, he received post-graduate residency training in clinical oncology with Tuen Mun Hospital, Hong Kong. He joined the Department of Clinical Oncology in 2008. He obtained his fellowship with the Royal College of Radiologists in clinical oncology in 2010. Afterwards, he received further specialist training in interstitial brachytherapy for head and neck cancers and sarcoma in Institut Gustave Roussy in Paris, France and novel radiation techniques like stereotactic radiosurgery and stereotactic ablative radiotherapy in Stanford University USA. In 2013, he received further training on stereotactic body radiation therapy for liver tumors at Princess Margaret Hospital, Toronto, Canada. More recently in 2015, he was awarded HKCR 15A Traveling Fellowship and pursued subspecialty training in image-guided brachytherapy for cervical cancer and pediatric oncology. His current research interests include clinical and genetic studies on nasopharyngeal cancer, head and neck cancers, lung cancers, liver cancers, and gastrointestinal cancers. He has published extensively in these respects. In addition, he has special interest in dosimetric studies on intensity-modulated radiation therapy, stereotactic radiosurgery and selective internal radiation therapy with Yttrium-90 microspheres for liver tumors.

Hong Yan (Fellow, IEEE) received the PhD degree from Yale University, New Haven, Connecticut. He was professor of imaging science with the University of Sydney and is currently Wong Chun Hong professor of data engineering, chair professor of computer engineering with the City University of Hong Kong, and director of Centre for Intelligent Multidimensional Data Analysis Limited (CIMDA). His research interests include image processing, pattern recognition, and bioinformatics, and he has more than 600 journal and conference publications in these areas. He is a fellow of International Association for Pattern Recognition (IAPR), fellow of US National Academy of Inventors (NAI), and a member of the European Academy of Sciences and Arts. He received the 2016 Norbert Wiener Award from the IEEE SMC Society for contributions to image and biomolecular pattern recognition techniques.