D3EGFR: a webserver for deep learning-guided drug sensitivity prediction and drug response information retrieval for EGFR mutation-driven lung cancer (Postprint)

Authors: Yulong Shi, Chongwu Li, Xinben Zhang, Cheng Peng, Peng Sun, Qian Zhang, Leilei Wu, Ying Ding, Dong Xie, Zhijian Xu, Weiliang Zhu, Ying Ding, Dong Xie, Zhijian Xu, Weiliang Zhu

Date: 2024-05-13T00:00:00+00:00

Abstract

As key oncogenic drivers in non-small-cell lung cancer (NSCLC), various mutations in the epidermal growth factor receptor (EGFR) with variable drug sensitivities have been a major obstacle for precision medicine. To achieve clinical-level drug recommendations, a platform for clinical patient case retrieval and reliable drug sensitivity prediction is highly expected. Therefore, we built a database, D3EGFRdb, with the clinicopathologic characteristics and drug responses of 1,339 patients with EGFR mutations via literature mining. On the basis of D3EGFRdb, we developed a deep learning-based prediction model, D3EGFRAI, for drug sensitivity prediction of new EGFR mutation-driven NSCLC. Model validations of D3EGFRAI showed a prediction accuracy of 0.81 and 0.85 for patients from D3EGFRdb and our hospitals, respectively. Furthermore, mutation scanning of the crucial residues inside drug-binding pockets, which may occur in the future, was performed to explore their drug sensitivity changes. D3EGFR is the first platform to achieve clinical-level drug response prediction of all approved small molecule drugs for EGFR mutation-driven lung cancer and is freely accessible at https://www.d3pharma.com/D3EGFR/index.php.

Full Text

Preamble

D3EGFR: A Webserver for Deep Learning-Guided Drug Sensitivity Prediction and Drug Response Information Retrieval for EGFR Mutation-Driven Lung Cancer

Yulong Shi[1,2,†], Chongwu Li[3,†], Xinben Zhang[1,†], Cheng Peng[1,2], Peng Sun[4], Qian Zhang[5], Leilei Wu[3], Ying Ding[6,], *Dong Xie*[3,], Zhijian Xu[1,2,], *Weiliang Zhu*[1,2,]

[1]State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China
[2]School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200433, China
[4]Key Laboratory of Human Functional Genomics of Jiangsu Province, Department of Biochemistry and Molecular Biology, Nanjing Medical University, Nanjing 211166, China
[5]School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
[6]Department of Pathology, the First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China

*Corresponding authors: Ying Ding (dingying@njmu.edu.cn), Dong Xie (kongduxd@163.com), Zhijian Xu (zjxu@simm.ac.cn), and Weiliang Zhu (wlzhu@simm.ac.cn)
†These authors contributed equally to this study.

Key Points

1. D3EGFR efficiently retrieves drug responses based on a searchable database of previously reported patient cases with EGFR mutations.

2. D3EGFR provides reliable drug response predictions through a deep learning-guided prediction model.

3. Mutation scanning of crucial residues was performed to characterize potential future mutations and their impact on sensitivity to approved drugs.

Abstract

As key oncogenic drivers in non-small-cell lung cancer (NSCLC), various mutations in the epidermal growth factor receptor (EGFR) exhibit variable drug sensitivities, representing a major obstacle for precision medicine. To achieve clinical-level drug recommendations, a platform integrating clinical patient case retrieval with reliable drug sensitivity prediction is urgently needed.

We constructed D3EGFRdb, a comprehensive database containing clinicopathologic characteristics and drug responses for 1,339 patients with EGFR mutations through systematic literature mining. Building upon this database, we developed D3EGFRAI, a deep learning-based prediction model for drug sensitivity in EGFR mutation-driven NSCLC. Model validation demonstrated

prediction accuracies of 0.81 and 0.85 for patients from D3EGFRdb and our hospitals, respectively. Furthermore, we performed mutation scanning of crucial residues within drug-binding pockets to explore potential future mutations and their impact on drug sensitivity. D3EGFR represents the first platform to achieve clinical-level drug response prediction for all approved small-molecule drugs in EGFR mutation-driven lung cancer and is freely accessible at https://www.d3pharma.com/D3EGFR/index.php.

Keywords: lung cancer, EGFR mutation, drug sensitivity prediction, patient case database, deep learning

Introduction

Lung cancer is the most common malignant disease and the leading cause of cancer mortality worldwide, with approximately 2.2 million new cases and 1.8 million deaths reported in 2020 [1]. Non-small-cell lung cancer (NSCLC) accounts for 85% of all lung malignancies [2, 3], predominantly comprising adenocarcinoma (ADC), squamous cell carcinoma, and large cell carcinoma. Epidermal growth factor receptor (EGFR) mutations are closely associated with carcinogenesis [4] and have been identified in approximately 32.3% of NSCLC cases [5]. Mutations in the kinase domain of EGFR promote ligand-independent dimerization and receptor activation, resulting in constitutive activation of downstream signaling pathways that drive tumorigenesis [6, 7].

EGFR-tyrosine kinase inhibitors (EGFR-TKIs) represent the standard treatment for advanced EGFR mutation-driven lung cancer [8, 9]. In patients with EGFR-sensitive mutations, EGFR-TKIs significantly improve objective response rates and prolong progression-free survival (PFS) and overall survival (OS) compared to platinum-based chemotherapy [10-12]. However, patients with different EGFR mutations exhibit variable responses to EGFR-TKIs, primarily due to intrinsic or acquired resistance [13]. Advances in DNA sequencing technologies have enabled identification of numerous novel and uncharacterized EGFR variants [14], further complicating precision medicine for patients with rare mutations [15, 16].

To date, only nine small-molecule drugs have been approved worldwide for metastatic EGFR mutation-positive NSCLC (Table S1). First-generation EGFR-TKIs, including gefitinib and erlotinib, serve as first-line therapy for common EGFR mutations such as exon 19 deletion (19del) or L858R point mutation [17, 18], while the third-generation agent osimertinib benefits patients with T790M resistance mutation [19, 20]. However, the efficacy of these EGFR-TKIs against uncommon or novel EGFR mutations remains inadequately characterized.

Benefiting from cumulative clinical trial experience over the past two decades, the risks of adverse effects and poor therapeutic efficacy in patients with common mutations have remained low throughout treatment. Notably, individual patient characteristics—including gender, age, and smoking status—also corre-

late with the incidence of EGFR mutation-driven lung cancer [5, 21]. Although considerable progress has been made in integrating information on EGFR mutants and targeted drugs [22-27], systematic retrospective clinical analysis has been limited by the lack of credible resources profiling clinical characteristics and outcomes of patients with EGFR mutations. Therefore, a comprehensive, searchable database detailing patient cases—including EGFR mutation status, clinicopathological characteristics, and therapeutic responses to approved drugs —is urgently needed to inform treatment decisions.

For rare or newly emerged variants, EGFR mutation status has been explored as a predictive and prognostic marker for targeted therapy efficacy [28-30]. For instance, Ikemura et al. [31] successfully predicted diverse in vitro and in vivo sensitivities of exon 20 insertion mutants using molecular dynamics (MD) simulations, obtaining ΔGbind values for mutant-inhibitor complexes in approximately one week. Wang et al. [32] combined MD simulations with extreme learning machines to construct a personalized drug resistance prediction model. However, the high computational costs and time requirements of MD simulations limit their widespread application. Moreover, previous studies have typically predicted drug responses for only two or fewer agents (Table S2). Recently, artificial intelligence has demonstrated enhanced capability in identifying, processing, and extrapolating drug-target interactions from existing biological activity data [33, 34], offering an effective approach for developing rapid and accurate drug sensitivity prediction models for rare and newly emerged mutations.

In this study, we aimed to investigate the impact of EGFR mutations on drug sensitivity and provide optimal treatment guidance through a real patient case database combined with a drug sensitivity prediction tool. First, we introduce D3EGFRdb, a clinical patient database detailing literature sources, case numbers, distribution of patient characteristics, and statistical analyses. Second, we evaluate the feasibility of molecular docking and deep learning approaches for drug sensitivity prediction and employ the selected deep learning model to explore potential drug sensitivity changes caused by amino acid mutations around the EGFR drug-binding pocket. Finally, we describe the construction and usage of the D3EGFR website to assist users in effectively utilizing both the D3EGFRdb patient database and D3EGFRAI prediction model.

Materials and Methods

Construction of a Clinical Medication Database for Patients with EGFR Mutations

We performed a literature search in PubMed [35] for relevant studies published before February 16, 2023, using the following strategy: (1) titles or abstracts must contain "EGFR mutation" and "non-small cell lung cancer" ; (2) titles or abstracts must include at least one approved EGFR-TKI agent, including "tyrosine kinase inhibitors," "gefitinib," "erlotinib," "icotinib," "afatinib," "osimertinib," "olmutinib," "dacomitinib," "almonertinib," and "furmonertinib" ; and (3)

full texts must contain keywords related to drug responses. Drug response was evaluated according to World Health Organization criteria [36] and Response Evaluation Criteria in Solid Tumors (RECIST) V1.0 or V1.1 guidelines [37, 38], categorized as complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). Therefore, full texts had to contain at least one of these four response keywords.

Prediction of Drug Sensitivity for EGFR Mutants Based on Molecular Docking

Molecular docking is a structure-based strategy for predicting potential drug-protein binding [39]. We constructed various docking models for different EGFR mutants and calculated correlations between docking scores and bioactivity to assess the feasibility of drug sensitivity prediction. The bioactivity dataset originated from Robichaux et al., encompassing 1,349 experimentally measured biological activities (log(mutant/wild type) of IC50 values) for 18 EGFR-TKIs and 77 EGFR mutants [40]. After excluding mutants reporting only exon-level information (e.g., 19del), we performed homology modeling to construct 3D structures for 64 mutants with clear mutation sites using the X-ray structure (PDB ID: 3POZ, resolution: 1.50 Å) as a template in MODELLER (version 9.24) [41]. Generated mutant protein structures were protonated at pH 7.4 using pdb2pqr software [42]. Molecular docking was conducted using smina [43], a fork of AutoDock Vina [44] with improved performance. Docking boxes for all mutants were generated by extending 4 Å in each dimension based on reference ligand coordinates in crystal complexes, with docking performed using random seed 0.

Deep Learning Model for Predicting Drug Sensitivity of EGFR Mutations

Deep learning enables feature detection from large-scale bioactivity data and has achieved remarkable success in drug-target interaction prediction through flexible neural network architectures [45]. Additionally, deep learning operates independently of protein 3D structures, thereby avoiding biases from structural modeling. We explored deep learning models with different encoder combinations for drugs and protein mutants to identify the optimal architecture for drug sensitivity prediction. Drug and protein encoders were provided by DeepPurpose [46], offering 80 encoder combinations (Table 1).

For dataset preparation, one-tenth of the 1,349 experimentally determined bioactivity data points were designated as the test set, with remaining data randomly divided into 10 different training and validation sets at a 9:1 ratio for 10-fold cross-validation. Models achieving average Pearson correlation > 0.8 across the 10 folds were retained. Evaluation metrics were calculated as follows:

Pearson correlation =
Mean Square Error (MSE) =

where N represents the number of samples, while and represent the labels and predicted values, respectively.

The test set was then used to evaluate retrained models after merging training and validation sets. We predicted binding affinities for mutations collected in D3EGFRdb and mapped predicted values to drug responses using multinomial logistic regression from the sklearn machine learning library. The logistic regression analysis used the 'newton-cg' solver, L2 penalty, C = 1.0, and balanced mode to automatically adjust weights inversely proportional to class frequencies. Figure 1 [Figure 1: see original paper] illustrates the D3EGFRAI drug sensitivity prediction framework.

Average Clinical Drug Response (ACR) for Quantitative Representation of Drug Response

Due to individual differences and complex factors, patients with identical mutation types and drug treatments may exhibit different responses. For instance, in D3EGFRdb, five patients with D770insSVD mutation received erlotinib, with three showing PD and two showing SD responses, making direct use of individual labels for model evaluation unreasonable. Therefore, we defined ACR to represent overall efficacy for patients sharing the same mutation type and drug treatment. Drug responses were converted to numerical values: CR/PR = -1, SD = 0, and PD = 1. Drug-mutant pairs with more than three patient cases in D3EGFRdb were screened, and their average clinical response value (ACRV) was calculated using Equation 3, then converted to ACR using Equation 4. This yielded a representative D3EGFRdb subset containing 43 drug-mutant pairs for model evaluation.

$$\text{ACRV} = (-1 \times \text{NCR/PR} + 0 \times \text{NSD} + 1 \times \text{NPD}) \,/\, (\text{NCR} + \text{NPR} + \text{NSD} + \text{NPD})$$

$$\text{ACR} = \text{- CR/PR if ACRV} > 0.5 \text{ - SD if } -0.5 \leq \text{ACRV} \leq 0.5 \text{ - PD if ACRV} < -0.5$$

where NCR/PR, NSD, and NPD represent the numbers of CR/PR, SD, and PD patients with the same mutation type and drug treatment, respectively.

Results

D3EGFRdb Overview and Statistical Analysis

Through systematic literature search and manual curation, we identified 141 studies on clinical medication and drug responses in EGFR-mutant patients, including 108 retrospective case reports/series, 26 prospective clinical trials, and 7 prospective cohort studies. All patients with EGFR mutations were collected and annotated with clinical information (mutation site, gender, age, smoking status, pathology, EGFR-TKI treatment), clinical outcomes (drug response, time to progression, PFS, OS), study type, and original literature references, constructing the D3EGFRdb clinical medication database.

D3EGFRdb contains 1,339 patients with 257 distinct mutation types, including 1,032 patients in the response group (CR/PR/SD) and 307 in the non-response group (PD). Reported mutation sites were predominantly located in exons 18-21 (Figure 2A [Figure 2: see original paper]), which encode the tyrosine kinase domain and represent drug binding sites. Exon 19 deletion and exon 21 L858R are the most common mutations, while less frequent variants include G719X and E709X in exon 18, S768I and T790M in exon 20, and L861Q and K860I in exon 21. From a clinical application perspective, the first-generation inhibitor gefitinib from AstraZeneca is the most extensively used and studied EGFR-TKI (951 cases, 71.0%), followed by another first-generation inhibitor erlotinib (256 cases, 19.1%). Gefitinib demonstrated slightly better clinical response rates than erlotinib (gefitinib: CR/PR vs. SD/PD = 51.2% vs. 48.8%; erlotinib: CR/PR vs. SD/PD = 44.1% vs. 55.9%) (Figure 2B). The relatively low usage of second-generation inhibitors afatinib and dacomitinib correlates with increased toxicity from non-specific targeting of wild-type EGFR [47, 48]. The third-generation EGFR-TKI osimertinib is the first FDA- and EMA-approved agent for metastatic NSCLC patients with T790M resistance mutation [49]. Icotinib, a potent and specific EGFR-TKI approved in China in 2011 [50], is also included. Together with gender, age, smoking status, pathology, time to progression, PFS, OS, study type, and original literature, D3EGFRdb serves as a comprehensive resource for retrospective medical record searches.

Multivariate analysis of D3EGFRdb (Figure 3 [Figure 3: see original paper]) revealed that females (47.8% vs. 31.6% males), individuals aged 60–79 years (34.1%), and non-smokers (39.1% vs. 23.8% smokers) were the most prevalent patient groups with EGFR mutations, consistent with previous reports [5, 21]. Adenocarcinoma predominated among pathologies (ADC vs. non-ADC = 68.1% vs. 7.9%), and point mutations were most common (48.6%), followed by deletion mutations (16.3%), primarily comprising the L858R substitution in exon 21 and deletions in exon 19.

External Clinical Dataset for Assessment

To validate the D3EGFRAI prediction model, we utilized clinical information and outcomes from 102 EGFR-TKI-treated patients at Shanghai Pulmonary Hospital between March 2015 and October 2020 as an external clinical dataset (Table 2 and Table S3). The Ethics Committee of Shanghai Pulmonary Hospital approved this retrospective study, with informed consent waived. Patient ages ranged from 33–85 years (median 61 years), with adenocarcinoma as the predominant histology. Objective responses were evaluated according to RECIST V1.1 guidelines [38], yielding 13 distinct drug-mutant pairs for which ACR was re-evaluated.

No Correlation Between Molecular Docking and Drug Response

Drug sensitivity prediction via molecular docking focuses on somatic mutations in exons 18–21 of the EGFR tyrosine kinase domain, based on the hypothe-

sis that docking scores correlate with drug sensitivity. We calculated docking scores for six approved drugs against 64 mutants and correlated them with experimental values. However, no significant correlation was observed (maximal $R^2 = 0.143$; Figure S1), suggesting molecular docking is unreliable for drug sensitivity prediction, likely due to low accuracy of homology modeling that fails to capture protein structural changes induced by residue mutations.

Deep Learning Models with High Prediction Accuracy

We calculated correlations between scores predicted by 80 deep learning models and experimental values. Seventeen models achieved average correlation > 0.8, demonstrating deep learning effectiveness in predicting binding affinity between protein mutants and EGFR-TKIs (Figure 4A–B [Figure 4: see original paper]). After retraining these 17 models by merging training and validation sets, 14 models maintained correlation > 0.8 on the test set (Figure 4C [Figure 4: see original paper]). A multinomial logistic regression model mapped predicted values to drug responses using the representative D3EGFRdb subset. The Morgan + CNN model performed best, achieving correlation coefficients of 0.81 on the biological activity validation dataset and 0.86 on the test dataset, with 0.81 prediction accuracy on the D3EGFRdb subset (Table S4). This model was selected as the final D3EGFRAI implementation.

As drug-mutant pairs may exhibit one or two predominant responses, Figure 5 [Figure 5: see original paper] shows predicted probabilities for each drug response in the representative D3EGFRdb subset. For example, afatinib-A767dupACS showed predicted probabilities of 47.0% for CR/PR, 47.75% for SD, and 5.3% for PD, indicating both CR/PR and SD as likely outcomes. Reporting only the highest-probability response would be insufficient. Therefore, D3EGFRAI displays both the most likely drug response and associated probabilities for each category. Evaluating the top two most likely responses improved prediction accuracy from 0.81 to 0.95 on the D3EGFRdb subset. Application to the external clinical dataset yielded accuracies of 0.85 (top response) and 0.92 (top two responses) (Table 3). Notably, 61.5% of drug-mutant pairs in the external dataset were absent from D3EGFRdb, demonstrating D3EGFRAI' s excellent generalization ability.

Mutation Scanning of Key Residues in the EGFR Drug-Binding Pocket Using D3EGFRAI

Amino acid mutations in the EGFR drug-binding site directly affect protein-drug binding affinity. We used D3EGFRAI to predict effects of potential mutations on drug sensitivity. Nineteen crystal complex structures of approved drugs were collected from the RCSB PDB database [51] (PDB IDs: 1M17, 2ITO, 2ITY, 2ITZ, 3UG2, 4G5J, 4G5P, 4HJO, 4I22, 4I23, 4I24, 4WKQ, 4ZAU, 6JWL, 6JX0, 6JX4, 6JXT, 6LUD). Twenty-six residues within 4 Å of the protein pocket were identified based on reference ligands: L718, G719, S720, F723, V726, K728, A743, I744, K745, E762, M766, L788, T790, Q791, L792, M793, P794, F795,

G796, C797, D800, E804, R841, L844, T854, and D855. Mutating these 26 residues into the other 19 standard amino acids generated 520 EGFR sequences for mutation affinity scanning, of which only 14 mutations (L718P, G719A, G719R, G719D, G719C, G719S, S720P, F723L, V726M, A743T, I744M, I744V, T790M, and G796S) were previously reported. Figure S2 shows that mutations at G796, T790, L718, L792, G719, and M766 significantly reduce first-generation EGFR-TKI efficacy (e.g., erlotinib, gefitinib, icotinib), indicating high risk for future mutations to compromise currently used drugs. Second-generation drugs (afatinib, dacomitinib) showed stronger binding affinity to most point mutations than first- and third-generation drugs, consistent with reported biological activity [31, 40] and suggesting that severe adverse reactions may relate to excessive binding affinity [47, 52]. Third-generation drugs (olmutinib, osimertinib, almonertinib, furmonertinib) have limited effects on C797, G796, and L718 mutations, with affinities generally stronger than first-generation but weaker than second-generation drugs. Beyond these point mutations, more complex variants warrant further investigation.

D3EGFR Input and Output

The D3EGFR server integrates D3EGFRdb and D3EGFRAI, enabling users to retrieve drug response information and predict responses for rare and novel mutations. Users can combine both approaches to determine optimal treatment. The webserver supports English and Chinese (Simplified), is free for all users, and requires no login.

Figure 6 [Figure 6: see original paper] illustrates the D3EGFR input and output interfaces. Notably, olmutinib has been prohibited for new prescriptions by the Ministry of Food and Drug Safety and was therefore removed from the approved drug list. D3EGFRdb retrieval provides statistical drug response ratios for mutants and drugs, plus detailed clinical characteristics and original literature for each case. For example, the T790M+L858R mutation had 29 cases in D3EGFRdb, with osimertinib achieving a 78.5% CP/PR response rate, superior to gefitinib (0%), erlotinib (0%), and afatinib (14.3%), confirming osimertinib as the effective treatment for T790M+L858R. D3EGFRAI predictions similarly show T790M+L858R sensitivity to osimertinib and resistance to gefitinib, erlotinib, and afatinib, consistent with D3EGFRdb and literature reports [53]. D3EGFRAI delivers prediction results within 10 seconds for novel mutation submissions.

Discussion

Drug sensitivity changes induced by protein mutations severely impact therapeutic benefits of targeted drugs. Hundreds of clinically reported EGFR mutations show inconsistent drug responses, primarily due to mutation-induced alterations in protein-drug binding affinity. At the atomic level, mutated residues may increase steric hindrance or influence protein-ligand interactions, thereby modulating drug binding capacity and treatment effectiveness.

As previously described [54-56], accumulating evidence supports EGFR mutants as predictive biomarkers for drug response in NSCLC. This study selected EGFR mutation and drug response as variables to build a prediction model. Unlike previous studies [31, 32, 57-60] that typically predicted responses for two or fewer drugs, we manually collected large-scale patient cases from two decades of literature to perform data-driven prediction for all approved EGFR-TKIs. However, we observed that patients with identical mutations sometimes showed different responses to the same drug, potentially related to individual characteristics and other unclear factors. We are collecting additional data to incorporate more variables in future model development to enhance predictive performance.

Conclusion

We developed the D3EGFR platform as a clinical-level drug recommendation tool to advance precision medicine. D3EGFRdb provides real patient cases with specific clinical information and treatment outcomes for convenient querying, while D3EGFRAI offers satisfactory prediction performance in clinical cases. Both components will be valuable for future clinical applications and research. Clinicians can combine D3EGFR' s real cases and predictions with their clinical experience and medical tests to make more informed treatment decisions. Additional reported cases and internal clinical trial results will further improve D3EGFR' s prediction accuracy and reliability.

Data Availability

The D3EGFR server is freely accessible at https://www.d3pharma.com/D3EGFR/index.php. The source dataset is available on GitHub (https://github.com/Zhijian-Xu/D3EGFR) and Zenodo (https://zenodo.org/records/10613332).

Supplementary Data

Supplementary data can be found in the Appendix.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFA0502301), the National Natural Science Foundation of China (31870717, 82172991), the Natural Science Research Program for Higher Education in Jiangsu Province (21KJB320015), the Shanghai Health Commission (2019SY072), and the Shanghai Pulmonary Hospital Research Fund (FK18001, FKGG1805).

References

1. Sung H, Ferlay J, Siegel RL et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 2021;71:209-249.

2. Lu T, Yang X, Huang Y et al. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades, Cancer Manag Res 2019;11:943-953.

3. Kosaka T, Yatabe Y, Endoh H et al. Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications, Cancer Res 2004;64:8919-8923.

4. Zhang YL, Yuan JQ, Wang KF et al. The prevalence of EGFR mutation in patients with non-small lung cancer: systematic review and meta-analysis, Oncotarget 2016;7:78985-78993.

5. Sharma SV, Bell DW, Settleman J et al. Epidermal growth factor receptor mutations in lung cancer, Nat Rev Cancer 2007;7:169-181.

6. Red Brewer M, Yun CH, Lai D et al. Mechanism for activation of mutated epidermal growth factor receptors in lung cancer, Proc Natl Acad Sci U S A 2013;110:E3595-E3604.

7. Vestergaard HH, Christensen MR, Lassen UN. A systematic review of targeted agents for non-small cell lung cancer, Acta Oncol 2018;57:176-186.

8. Mok TS, Wu YL, Thongprasert S et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma, N Engl J Med 2009;361:947-957.

9. Zhou C, Wu YL, Chen G et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study, Lancet Oncol 2011;12:735-742.

10. Maemondo M, Inoue A, Kobayashi K et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR, N Engl J Med 2010;362:2380-2388.

11. Mitsudomi T, Morita S, Yatabe Y et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial, Lancet Oncol 2010;11:121-128.

12. Yu HA, Arcila ME, Rekhtman N et al. Analysis of tumor specimens at

the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers, Clin Cancer Res 2013;19:2240-2247.

13. Kohsaka S, Nagano M, Ueno T et al. A method of high-throughput functional evaluation of EGFR gene variants of unknown significance in cancer, Sci Transl Med 2017;9:eaan6566.

14. Kris MG, Johnson BE, Berry LD et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs, JAMA 2014;311:1998-2006.

15. Tu HY, Ke EE, Yang JJ et al. A comprehensive review of uncommon EGFR mutations in patients with non-small cell lung cancer, Lung Cancer 2017;114:96-102.

16. Sutiman N, Tan SW, Tan EH et al. EGFR mutation subtypes influence survival outcomes following first-line gefitinib therapy in advanced Asian NSCLC patients, J Thorac Oncol 2017;12:529-538.

17. Park K, Yu CJ, Kim SW et al. First-line erlotinib therapy until and beyond response evaluation criteria in solid tumors progression in Asian patients with epidermal growth factor receptor mutation-positive non-small-cell lung cancer: The ASPIRATION study, JAMA Oncol 2016;2:305-312.

18. Remon J, Caramella C, Jovelet C et al. Osimertinib benefit in EGFR-mutant NSCLC patients with T790M-mutation detected by circulating tumour DNA, Ann Oncol 2017;28:784-790.

19. Lee J, Choi Y, Han J et al. Osimertinib improves overall survival in patients with EGFR-mutated NSCLC with leptomeningeal metastases regardless of T790M mutational status, J Thorac Oncol 2020;15:1758-1766.

20. Shigematsu H, Lin L, Takahashi T et al. Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers, J Natl Cancer Inst 2005;97:339-346.

21. Patterson S, Statz C, Yin T et al. The JAX clinical knowledgebase: A valuable resource for identifying evidence related to complex molecular signatures in different types of cancer, Cancer Genet 2017;214-215:33.

22. Griffith M, Spies NC, Krysiak K et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer, Nat Genet 2017;49:170-174.

23. Huang L, Fernandes H, Zia H et al. The cancer precision medicine knowl-

edge base for structured clinical-grade mutations and interpretations, J Am Med Inform Assoc 2017;24:513-519.

24. Bamford S, Dawson E, Forbes S et al. The COSMIC (Catalogue of somatic mutations in cancer) database and website, Br J Cancer 2004;91:355-358.

25. Chakravarty D, Gao J, Phillips SM et al. OncoKB: A precision oncology knowledge base, JCO Precis Oncol 2017;2017.

26. Swanton C. My cancer genome: a unified genomics and clinical trial portal, Lancet Oncol 2012;13:668-669.

27. Chou TY, Chiu CH, Li LH et al. Mutation in the tyrosine kinase domain of epidermal growth factor receptor is a predictive and prognostic factor for gefitinib treatment in patients with non-small cell lung cancer, Clin Cancer Res 2005;11:3750-3757.

28. Fang S, Wang Z. EGFR mutations as a prognostic and predictive marker in non-small-cell lung cancer, Drug Des Dev Ther 2014;8:1595-1611.

29. Han SW, Kim TY, Hwang PG et al. Predictive and prognostic impact of epidermal growth factor receptor mutation in non-small-cell lung cancer patients treated with gefitinib, J Clin Oncol 2005;23:2493-2501.

30. Ikemura S, Yasuda H, Matsumoto S et al. Molecular dynamics simulation-guided drug sensitivity prediction for lung cancer with rare EGFR mutations, Proc Natl Acad Sci U S A 2019;116:10025-10030.

31. Wang DD, Zhou W, Yan H et al. Personalized prediction of EGFR mutation-induced drug resistance in lung cancer, Sci Rep 2013;3:2855.

32. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction, Bioinformatics 2018;34:i821-i829.

33. Yang Y, Zhou D, Zhang X et al. D3AI-CoV: a deep learning platform for predicting drug targets and for virtual screening against COVID-19, Brief Bioinform 2022;23.

34. Wheeler DL, Church DM, Edgar R et al. Database resources of the National Center for Biotechnology Information: update, Nucleic Acids Res 2004;32:D35-40.

35. Miller AB, Hoogstraten B, Staquet M et al. Reporting results of cancer treatment, Cancer 1981;47:207-214.

36. Therasse P, Arbuck SG, Eisenhauer EA et al. New guidelines to evaluate

the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada, J Natl Cancer Inst 2000;92:205-216.

37. Eisenhauer EA, Therasse P, Bogaerts J et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), European Journal of Cancer 2009;45:228-247.

38. Kitchen DB, Decornez H, Furr JR et al. Docking and scoring in virtual screening for drug discovery: methods and applications, Nat Rev Drug Discov 2004;3:935-949.

39. Robichaux JP, Le X, Vijayan RSK et al. Structure-based classification predicts drug response in EGFR-mutant NSCLC, Nature 2021;597:732-737.

40. Webb B, Sali A. Comparative protein structure modeling using MODELLER, Curr Protoc Bioinformatics 2016;54:1-5.

41. Dolinsky TJ, Nielsen JE, McCammon JA et al. PDB2PQR: an automated pipeline for the setup Poisson-Boltzmann electrostatics calculations, Nucleic Acids 2004;32:W665-667.

42. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise, J Chem Inf Model 2013;53:1893-1904.

43. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J Comput Chem 2010;31:455-461.

44. Huang K, Fu T, Glass LM et al. DeepPurpose: a deep learning library for drug-target interaction prediction, Bioinformatics 2021;36:5545-5547.

45. Takeda M, Okamoto I, Nakagawa K. Pooled safety analysis of EGFR-TKI treatment for EGFR mutation-positive non-small cell lung cancer, Lung Cancer 2015;88:74-79.

46. Ramalingam SS, O' Byrne K, Boyer M et al. Dacomitinib versus erlotinib in patients with EGFR-mutated advanced nonsmall-cell lung cancer (NSCLC): pooled subset analyses from two randomized trials, Ann Oncol 2016;27:423-429.

47. Remon J, Steuer CE, Ramalingam SS et al. Osimertinib and other third-generation EGFR TKI in EGFR-mutant NSCLC patients, Ann

Oncol 2018;29:i20-i27.

48. Chen X, Zhu Q, Liu Y et al. Icotinib is an active treatment of non-small-cell lung cancer: a retrospective study, PLOS ONE 2014;9:e95897.

49. Burley SK, Berman HM, Bhikadiya C et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic Acids Res 2019;47:D464-D474.

50. Wu YL, Cheng Y, Zhou X et al. Dacomitinib versus gefitinib as first-line treatment for patients with EGFR-mutation-positive non-small-cell lung cancer (ARCHER 1050): a randomised, open-label, phase 3 trial, Lancet Oncol 2017;18:1454-1466.

51. Hsu WH, Yang JC, Mok TS et al. Overview of current systemic management of EGFR-mutant NSCLC, Ann Oncol 2018;29:i3-i9.

52. Del Re M, Rofi E, Cappelli C et al. The increase in activating EGFR mutation in plasma is an early biomarker to monitor response to osimertinib: a case report, BMC Cancer 2019;19:410.

53. Kaneko K, Kumekawa Y, Makino R et al. EGFR gene alterations as a prognostic biomarker in advanced esophageal squamous cell carcinoma, Front Biosci (Landmark Ed) 2010;15:65-72.

54. Dahabreh IJ, Linardou H, Siannis F et al. Somatic EGFR mutation and gene copy gain as predictive biomarkers for response to tyrosine kinase inhibitors in non–small cell lung cancer, Clin Cancer Res 2010;16:291-303.

55. Zou B, Lee V H F, Yan H. Prediction of sensitivity to gefitinib/erlotinib for EGFR mutations in NSCLC based on structural interaction fingerprints and multilinear principal component analysis, BMC bioinformatics 2018;19:1-13.

56. Wang D D, Lee V H F, Zhu G, et al. Selectivity profile of afatinib for EGFR-mutated non-small-cell lung cancer, Mol Biosyst 2016;12(5):1552-1558.

57. Chiu Y C, Chen H I H, Zhang T, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks, BMC Med Genomics 2019;12(1):143-155.

58. Ma L, Wang D D, Zou B, et al. An eigen-binding site based method for the analysis of anti-EGFR drug resistance in lung cancer treatment, IEEE/ACM Trans Comput Biol Bioinform 2016;14(5):1187-1194.

Biographical Notes

Yulong Shi is a Ph.D. student at the Shanghai Institute of Materia Medica. His research interests include computational biology and artificial intelligence.

Chongwu Li is a Ph.D. student at Shanghai Pulmonary Hospital. His research interests include artificial intelligence-based precision medicine.

Xinben Zhang received his master' s degree from East China University of Science and Technology. His research interest is software development.

Cheng Peng received his Ph.D. from the Shanghai Institute of Materia Medica. His research interests include computer-aided drug design and molecular dynamics simulation.

Peng Sun is a master' s supervisor at Nanjing Medical University. His research interests include biochemistry and molecular biology.

Qian Zhang is a master' s supervisor at East China Normal University. Her research interests include machine learning and artificial intelligence.

Leilei Wu is a Ph.D. student at Shanghai Pulmonary Hospital. His research interests include artificial intelligence-based precision medicine.

Ying Ding is a clinician at the First Affiliated Hospital of Nanjing Medical University, primarily engaged in precision medicine for lung cancer.

Dong Xie is a clinician at Shanghai Pulmonary Hospital, primarily engaged in precision surgical treatment of early-stage lung cancer.

Zhijian Xu received his Ph.D. from the Shanghai Institute of Materia Medica in 2012. His research interests include drug-target interaction and virtual screening.

Weiliang Zhu received his Ph.D. from the Shanghai Institute of Materia Medica in 1998. His main research fields include computational biology, computational chemistry, and pharmaceutical chemistry.

Note: Figure translations are in progress. See original paper for figures.