Lung cancers with a mutated epidermal growth factor receptor (EGFR) are a major contributor to cancer fatalities globally. Targeted tyrosine kinase inhibitors (TKIs) have been developed against EGFR and show encouraging results for survival rate and quality of life. However, drug resistance may affect treatment plans and treatment efficacy may be lost after about a year. Predicting the response to EGFR-TKIs for EGFR-mutated lung cancer patients is a key research area. In this study, we propose a personalized drug response prediction model (PDRP), based on molecular dynamics simulations and machine learning, to predict the response of first generation FDA-approved small molecule EGFR-TKIs, Gefitinib/Erlotinib, in lung cancer patients. The patient’s mutation status is taken into consideration in molecular dynamics (MD) simulation. Each patient’s unique mutation status was modeled considering MD simulation to extract molecular-level geometric features. Moreover, additional clinical features were incorporated into machine learning model for drug response prediction. The complete feature set includes demographic and clinical information (DCI), geometrical properties of the drug-target binding site, and the binding free energy of the drug-target complex from the MD simulation. PDRP incorporates an XGBoost classifier, which achieves state-of-the-art performance with 97.5% accuracy, 93% recall, 96.5% precision, and 94% F1-score, for a 4-class drug response prediction task. We found that modeling the geometry of the binding pocket combined with binding free energy is a good predictor for drug response. However, we observed that clinical information had a little impact on the performance of the model. The proposed model could be tested on other types of cancers. We believe PDRP will support the planning of effective treatment regimes based on clinical-genomic information. The source code and related files are available on GitHub at: https://github.com/rizwanqureshi123/PDRP/.

Lung cancer is a leading cause of deaths worldwide, and has the lowest survival rate among all cancer types. It is the second most common type of cancer, and often diagnosed at later stages when metastatic spread to other parts of the body may have occurred. In the last decade, rapid progress has been made in the management of non-small cell lung cancer (NSCLC) patients. Molecular targeting has made great advances, and epidermal growth factor receptor (EGFR) and ErbB family members have been identified as useful therapeutic targets. Over-expression of EGFR is found in about 60% of advanced NSCLC patients. The US Food and Drug Administration (FDA) has approved three generations of small molecule tyrosine kinase inhibitors (TKIs) Gefitinib/Erlotinib/Afatinib/Osimertinib as a first line treatment for lung cancer patients harboring EGFR mutations. These TKIs produced encouraging results at the initial stage of therapy and increased the survival rate and quality of life of patients. However, resistance to these drugs has appeared in many cases. A major cause of this resistance is a secondary point mutation in the kinase domain of EGFR.

Several studies have attempted to decode the mechanism of drug resistance in EGFR-mutated lung cancer. These studies found many reasons, such as the secondary point mutation T790M, breaking of the hydrogen bond at site 790, and reactivation of AKT. In silico methods have been extensively applied to study these drug resistance mechanisms. Molecular dynamics (MD) simulation is a computational tool, which has been used to understand dynamics, stability, and structural variations. Recently, a framework has been developed for...

**Rizwan Qureshi¹, Syed Abdullah Basit², Jawwad A. Shamsi², Xinqi Fan³,⁴, Mehmood Nawaz⁵, Hong Yan³,⁴ & Tanvir Alam¹**

¹College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. ²FAST National University of Computer and Emerging Sciences, Karachi, Pakistan. ³Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong. ⁴Center for Intelligent Multidimensional Data Analysis (CIMDA), City University of Hong Kong, Kowloon, Hong Kong. ⁵Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR, China.  *email: talam@hbku.edu.qa*
the visualization of protein-drug interactions in the analysis of drug resistance in lung cancer. There is still much unexplained variation in patients’ responses to these drugs and clinical-genomic features of patients may play a significant role underlying the mechanism of drug resistance and patient stratification.

The completion of the human genome project allowed a shift of paradigm from the traditional medical model of targeting large populations to precision therapies. Information from genomics and electronic health records provides novel opportunities for patient care, prevention, and devising optimal treatment strategies. Predicting a patient’s response to a drug treatment, or identifying their optimal treatment strategy, such as the combinations and doses of drugs, is challenging for computational methods due to limited data sources, disparity among labels and unknown biological evidence. The position of the drug-binding site, binding free energy, geometrical features, and clinical information may be used to model multi-class drug responses. The protein data bank (PDB) contains several high-resolution structures of EGFR bound to different generations of approved drugs, providing opportunities to build structure-based data-driven models. An EGFR-L858R dimer with Gefitinib is shown in Fig. 1.

As drug responses are often mediated by protein-drug interactions, the geometry of the drug-target binding site or pocket can be a useful predictor of drug response. MD simulation of the binding energy of drug-mutant complexes and related personal characteristics of patients, when used as input to an extreme learning machine (ELM), classified drug response levels into two classes, with an accuracy of 95.3%. Combining local geometrical properties with energy related features in an Eigen binding site method achieved an average accuracy of 69.35% in predicting four classes of drug responses. Patients’ demographic and lifestyle patterns and geometric features of drugs were used for a similar predictive model, and protein-drug interaction footprint tensors were used in a three-level drug response predictive model.

While these studies demonstrate the potential of combining dynamic molecular features with patient data to predict responses to a drug, the quality of the predictions needs to be improved if they are to be of clinical use. In this work, we combined the geometric information of the drug-target binding site, the binding energy, and patient’s clinical information to predict four classes of drug response with an imbalanced dataset. The proposed personalized drug response prediction (PDRP) model achieves state of the art performance in predicting drug responses. The main contributions of this work are:

- We have developed a drug response prediction model using molecular dynamics (MD) simulation and machine learning (ML), which achieves state of the art performance.
- We have proposed a set of novel geometric features by modelling the binding site of drug-target complex. Our work demonstrates the contribution of the geometrical features of binding sites in improving the predictive accuracy of our ML model.
- The proposed model, PDRP, is a generic concept, which can be used for other cancers or diseases with minor modifications, with potential applications in clinical decision support systems and patient stratification.

Results

In this work, we have developed a drug-response predictive model for lung cancer patients. Demographic and clinical information, such as age, gender, survival time, smoking history, tumor progression level and type of mutation, were collated from previous studies. We simulated EGFR-Gefitinib complexes in the AMBER software suite for 2 ns and extracted the trajectories. Based on the simulated trajectories, we proposed novel...
geometrical features $x^1_g$ (the matching rate, number of connected atoms, and number of hydrogen bonds) and $x^2_g$ (the number of convex atoms and Euclidean distance) from the EGFR mutant-drug complexes to predict the drug-response level as one of complete response, partial response, stable disease or progressive disease. We also used binding free energy as a feature in the machine learning model. The combination of clinical, geometric and energy related features boosted the performance of the machine learning model.

Baseline statistics. 201 NSCLC patients form the cohort of this study. The patients had a median age of 63 years, 35% (71) were female and 65% (130) male, and about 75% were non-smokers. All patients received EGFR-TKIs as their first line of treatment. Response levels 0 and 1 indicate complete and partial responses to the drug. Response levels 2 and 3 correspond to stable and progressive disease (No response). The dataset used here consisted of 19, 118, 30, 34 patients at response levels 0, 1, 2 and 3, corresponding to complete, partial, stable, and progressive disease (No response), respectively.

Proposed feature set summary. For a patient, his/her demographic and clinical information (DCI), the energy and geometric features of their mutant EGFR-Gefitinib complex were obtained and used to predict their drug response level through machine learning classifiers. A detailed description of the features and their value ranges are presented in Table 1. In total we extracted 4, 4, and 5 features from DCI, energy and geometric types, respectively. Figure 2 shows the box plot for different features and the correlations among them.

A total of 33 different EGFR mutations occurred in these patients. The root mean square deviations (RMSD) of the trajectories from MD simulations of WT and four mutants and the patient disease response classification by mutation type are shown in Fig. 3. The most common mutations were L858R, delE746−750 and L858R−T790M. All mutations were modelled based on the EGFR 3D structure using Rosetta.

The potency of an inhibitor can be measured by the time a patient survives and their drug response level. The relationship between the drug response

| Feature type | Attributes | Description                  | Discrete/Continuous | Range     |
|--------------|------------|------------------------------|---------------------|-----------|
| DCI          | Age        | Patient's personal information | Discrete            | [0–4]     |
|              | Sex        |                               | Discrete            | [0–2]     |
|              | Smoking history |                             | Discrete            | [0–2]     |
|              | Response status |                           | Discrete            | [0–3]     |
| Energy       | VDW        | Van der Walls energy          | Continuous          | [− 60 to − 45] |
|              | EEL        | Electrostatic interactions    | Continuous          | [− 23–11] |
|              | ESURF      | No polar component of the solvation energy | Continuous | [− 45 to − 1] |
|              | EPB        | Polar component of solvation free energy | Continuous | [27–40] |
| Geometric    | Matching rates | Matched atoms               | Discrete            | [0, 17]   |
|              | Convex atoms | Strength of interaction      | Discrete            | [0, 43]   |
|              | Connectivity | Connected atoms             | Discrete            | [0, 23]   |
|              | Euclidean distance | Distance between drug and target | Continuous | [30–39]   |
|              | Hydrogen bonds | Number of hydrogen bonds     | Discrete            | [775–1650] |

Table 1. Clinical information, energy and geometrical features: description and values.

Figure 2. Box plot of normalized values for energy, and geometrical features (left panel), and correlation among features (right panel).
and the personal and energy features were not linearly, or one-to-one, related to drug response or survival time (Fig. 4). Drug response was classified into four levels, based on the response evaluation criteria in solid tumors (RECIST)\(^3^7\). A list of mutations types is provided in the Supplementary File S1.

Performance of the proposed PDRP model. To build the PDRP model, we incorporated clinical information, MD simulation results and novel geometric features from the protein–drug interactions. An ablation study, based on each type of feature and the corresponding model’s performance, showed that the geometrical features were the most powerful predictors, followed by DCI and energy-related features (Fig. 5). Of the models tested, XGboost had the best performance. Demographic and clinical features brought a little improvement to the performance, and the combination of all three types of features boosted the performance in XGboost, Random Forest and the Neural Network.

The confusion matrices for the training and classification reports are shown in Figs. 7 and 8, respectively. The XGboost classifier achieved 97.5% accuracy, 97% recall, 93% precision, and 97% F1-score with only two mis-classifications for an independent testing set of 61 samples. The performance of the random forest and neural network were close to, but not better than, XGboost. The classifiers were trained using a nested-cross validation approach, and the parameters of classifiers were optimized using a grid search approach.
Discussion
Our main focus in this work was on modeling the geometry of the drug-binding pocket and combining this with the demographic and clinical information (DCI) of the patients. The geometrical features (the number of convex atoms at the interaction surface of the complex, the matching rates of surface atoms, the dynamic distances between the center of the drug molecule and binding-site residues throughout the trajectory, and the number of hydrogen bonds) were the most discriminative features. Combining clinical and molecular predictors to identify drug-sensitive patients was most effective. The mutation changes the shape and structure of the drug binding complex, which leads to changes in the values of the geometrical features and thus the drug response. Further investigation of this model may result in geometrical features that better stratify patients based on mutation, age or gender specific therapies and the model might also help in selecting the optimal drug for specific patients.

In addition, we perform another ablation study only on geometric features. We found that geometric feature 2 has more predictive capability than geometric feature 1. However combining the feature further boosts the performance, as shown in Fig. 6. Moreover, it will be interesting to see the performance of modern geometrical based deep learning algorithms, if a large number of samples are available.

Comparison against other methods. Table 3 shows that the classification accuracy of the proposed method achieved state of the art performance relative to related works and achieved this while predicting the highest number of classes (four) of drug responses. Methods predicting fewer response levels performed better.
than earlier four-level predictors, however, our method outperformed all previous work. The combination of geometrical, energy, and personal features seems to be a more optimal strategy for predicting the drug response. The highly accurate pioneering work\textsuperscript{13} used binding free energy, and personal features of 168 patients but only predicted a two-class drug response. Figure 5 shows the contribution of each feature. The demographic and clinical (DCI) feature slightly improved the drug response prediction as shown in Table 2. To further test the performance of the proposed model, we convert our problem into 2 class binary classification. We combined partial response and full response to one class, and stable disease and no response to another class. After that, we re-run the method proposed by\textsuperscript{13} and\textsuperscript{39} on our task. The results are shown in Table 3. Our method also performs well on binary classification problem.

Figure 7. Confusion matrix for testing dataset.
Figure 8. Classification performance on testing dataset.

Table 2. Comparison of PDRP with other methods. The best accuracy is in bold. The 2-class response includes complete response and no-response, while 3-class response includes low, moderate, and full response. The 4-class response includes complete-response, partial-response, no-response, and progressive disease.

| Reference | Patients | Features | Method | Response level | Accuracy |
|-----------|----------|----------|--------|----------------|----------|
| 15        | 167      | Personal energy | Extreme learning machines | 2 | 95.13 |
| 20        | 355      | Personal and genetic | Sequential minimization optimization | 2 | 76.56 |
| 20        | 137      | Geometrical and personal | Softmax regression | 4 | 70.78 |
| 20        | 311      | Energy and geometrical | Support vector machine | 4 | 69.35 |
| 31        | NA       | Protein-drug interactions | Naive Bayes | 3 | 95.50 |
| PDRP      | 201      | DCI, Energy, Geometric | XGboost | 4 | 97.50 |
Performance of PDRP on gender stratified samples. Experiments using PDRP on gender stratified samples showed that the model performed well on both male and female patients (Fig. 9). Drug responses were correctly identified in 21 females patients with no mis-classification. However, there were two mis-classifications, between partial and complete responses, in the male dataset of 39 patients.

Translation research. In this paper, we proposed two novel composite features. Composite feature $x_{1}$ captures the interaction strength of the protein-drug complex, and composite feature $x_{2}$ captures the shape and local surface information of the protein-drug complex. The morphological properties of the binding site, the number of convex atoms and the dynamic distance help us to model the geometrical properties. Length and strength of protein-drug interactions are represented by these features, which helps in predicting the drug-response. The features that play a prominent role in the drug response could be further studied to screen inhibitors for personalized therapy and may become useful biomarkers for drug screening. We also showed that combining DCI, geometrical and energy related features helps to boost the performance of the drug response prediction model.

Limitations. A limitation of this study is that it contained only the 33 most common EGFR mutations from the 594 EGFR mutations available in COSMIC database. However, the mutations studied account for about 90% of all mutations. It is difficult to determine drug sensitivity to rare mutations due to limited patient data. Another limitation is the small dataset of 201 patients. Obtaining clinical data is difficult due to privacy and ethical considerations. Most clinical studies consist of fewer than 400 patients (Table 3) and may have imbalanced numbers of patients at each response level. Despite this, our model achieved a highly accurate prediction rate.

Protein-ligand interactions usually take place in nanoseconds, so our simulation is adequate. Many of the previous studies on EGFR drug response prediction also used 2-ns MD simulation. However, the time scale is highly case-specific and may cause computing costs to explode. Of course, it would be useful to study the interactions for longer time, but MD simulations only provide an approximate solution, and the later part of the simulations can be less accurate. Quantum mechanics (QM) based analysis would be more accurate, but the computational cost is currently beyond the reach of most computers unless we only deal with very small molecules.

Methods
The framework to classify individual patient outcomes is divided into three modules: computational modeling of mutant/drug structures, MD simulations, and classification. Figure 10a shows a method for computationally modeling the structure of the EGFR-Gefitinib drug complex; PDB 2ITY was used as the template structure.
Figure 10b and c show the steps of the MD simulation and the formulation of the classification models. Each mutation was modeled in the 3D structure of the protein-drug complex. Our dataset consisted of 33 different EGFR point mutations. For each mutant, MD simulations were performed and the binding free energy was calculated. Multiple machine learning models were used to predict the response level.

Dataset collection. All the methods under this study were performed in accordance with the relevant guidelines and regulations of City University of Hong Kong (CityU). Informed consent was taken from all participants. The experimental protocols were approved by the Institutional Review Board (IRB) of City University of Hong Kong. The clinical information used in this study was collected from several published sources. A dataset of 201 NSCLC patients was obtained and 140 samples were used to train the model. 61 independent samples were used to test the model, with 3, 5, 36, and 17 samples for stable disease, no response, partial response and complete response respectively. The L858R mutation was the most frequent mutation, followed by secondary point mutation L858R-T790M and delE746-750, as shown in Fig. 3b. The EGFR mutations were selected based on a survey.

Demographic and clinical information (DCI). Demographic information such as age, sex, smoking history as well as clinical information such as survival [0, 1], drug response level and performance status were extracted from the clinical dataset. Age was encoded between 0 to 4 based on the age brackets (0, 40), (41, 50), (51, 60), (61, 70), and (71 and above), due to inherent modes.

Modeling of 3D structures. The 3D mutant structures were predicted based on the crystal structure of wildtype EGFR taken from Protein Data Bank (PDB) with PDB ID 2ITY. The high resolution dgdmonomer (HRDM) protocol in Rosetta was used to predict point mutations, and the comparative modeling protocol was used to predict multi-point mutations. Quality assessment of predicted structures was performed by Verify3D and Q-mean.

Molecular dynamics simulation. MD simulations of the protein-drug complex were performed using the QM/MM method in Amber with a surrounding waterbox neutralized using Na+ and Cl- atoms and the ff9SB and GAFF force fields. The total energy of the system was the summation of bonded (stretch, bend, torsion) and non-bonded (electrostatic, van der Waals) terms.

$$E_{total} = E_{stretch} + E_{bend} + E_{torsion} + E_{electrostatic} + E_{vdw}$$

Energy minimization was used to refine the modeled structure before the MD run, which begins with the system being heated from 0K to 300K, followed by equilibration at constant pressure and density for 500-ps and 50-ps respectively. The SHAKE algorithm was utilized to control temperature and constrain bond stretching. After reaching a stable state, MD simulation runs were carried out for 2-ns at a steady temperature (300K) and pressure (1 Atm). A twelve core 3.47 GHz CPU with eight GB RAM was used to run MD simulations, with
a Tesla C2075 GPU used in production runs. Each simulation took roughly 12 h to complete. The trajectory was extracted using the CPPTRAJ package in Amber, with frames gathered every 10-ps, producing 200 frames per run.

**Binding free energy.** The free energy of binding of a drug to a protein in a solvated environment estimates the binding affinity. The parallel version of MM-GBSA on a twelve core, 3.47 GHz CPU was used for the simulation. The MD trajectory was considered as input to MM-GBSA. Each round of simulation took on average twelve hours for computation. The binding free energy is calculated based on the theory of the thermodynamic cycle in vacuum and solvated environments as:

\[
\Delta G = \Delta G_{\text{Bind,Vacuum}} + \Delta G_{\text{Solv,Complex}} - \left( \Delta G_{\text{Solv,ligand}} + \Delta G_{\text{Solv,Receptor}} \right)
\]

where \(\Delta G\) is the binding free energy difference of the receptor-ligand system in a vacuum. \(\Delta G_{\text{Solv,Complex}}, \Delta G_{\text{Solv,ligand}},\) and \(\Delta G_{\text{Solv,Receptor}}\) represent their energy differences between vacuum and solvent states. The energy component is composed of Van der Waals forces (VDW), electrostatic energy (EEL), the electrostatic contribution to solvation, and non-polar contributions to the solvation free energy (ESURF).

**Geometric features.** Interactions between the binding site residues of a protein and small molecule inhibitors are commonly used in prediction methods. Local geometric surface properties were determined based on the alpha shape using the computational geometry algorithm library (CGAL).

**Convex atoms.** Each atom in a drug-mutant system has a position and mass, represented as \(a = (p, w)\), where \(p\) is the position and \(w\) is the mass of the atom. Two atoms \(a_1 = (p_1, w_1)\) and \(a_2 = (p_2, w_2)\) are defined as orthogonal or sub-orthogonal using the following equation.

\[
\begin{align*}
|p_1p_2| &= w_1 + w_2, \quad a_1 \perp a_2 \\
|p_1p_2| &= w_1 > w_2, \quad a_1 \perp a_2
\end{align*}
\]

From the alpha shape, the solid angle of atoms was determined to characterize the geometric properties of the local surface. If \(A, B, C, \) and \(D\) are the vertices of a tetrahedron, the solid angle, \(\Omega_i\), is:

\[
\Omega_i = \sum_i (\phi_i^{AB} + \phi_i^{BC} + \phi_i^{AC} - \pi),
\]

where \(\phi_i^{AB}, \phi_i^{BC},\) and \(\phi_i^{AC}\) represent the dihedral angles of tetrahedron \(i\).

\[
\Omega' = \frac{\cos(\Omega_i)}{4}
\]

\(\Omega'\) results in a convex shape if positive and a concave shape if negative (Fig. 11). The number of atoms in a convex shape at the local surface of the drug-dimer complex was used.

**Matching rates.** The atoms at the interface of the structures create the interaction between the drug and the target. The surface atoms were collected using the alpha shape algorithm, and named as point set \(A\). After that, the convex shape at the local surface of the drug-dimer complex was used. If \(A, B, C, \) and \(D\) are the vertices of a tetrahedron, the solid angle, \(\Omega_i\), is:

\[
\Omega_i = \sum_i (\phi_i^{AB} + \phi_i^{BC} + \phi_i^{AC} - \pi),
\]

where \(\phi_i^{AB}, \phi_i^{BC},\) and \(\phi_i^{AC}\) represent the dihedral angles of tetrahedron \(i\).

\[
\Omega' = \frac{\cos(\Omega_i)}{4}
\]

\(\Omega'\) results in a convex shape if positive and a concave shape if negative (Fig. 11). The number of atoms in a convex shape at the local surface of the drug-dimer complex was used.

The matching rate was determined by selecting atoms at the drug and the target. If one of the atoms is convex and other is concave, the pair is recorded as matched and there is a strong interaction between them. If both atoms are convex or concave, the pair is unmatched, and the interaction is weak, as shown in Fig. 11. The matched and unmatched atoms are determined as:

\[
f(B, C) = \begin{cases} 
1 & \Omega_B \times \Omega_C < 0 \\
0 & \text{otherwise}
\end{cases}
\]

Matching rates are calculated for each frame of the MD trajectory as:

\[
MR = \frac{\sum_{ij} f(B_i, C_j)}{N}
\]

\(MR\) represents the matching rate, \(f(B_i, C_j)\) is a matched atom pair, and \(N\) is the total number of MD snapshots. The matching rate is used as a feature in this work, and low matching rates were linked to low drug responses.

**Connectivity measure.** Connectivity changes between binding site residues and the drug molecule throughout the MD simulation. We define a local threshold value of 40 based on Euclidean distance and then record the
number of atoms that remain within this threshold throughout the MD Simulation. The consistency of these connections may identify critical atoms and could be used as a predictor of the drug response level:

\[ C_{k,i} = \sum_j A_{k,i,j} \]  

where \( A_{k,i,j} \) represents the connection between the \( i^{th} \) EGFR atom and the \( j^{th} \) drug atom in the \( k^{th} \) MD snapshot, and is 1 if there is a connection and zero otherwise. Let

\[ D_k = \sum_i C_{k,i} > 0 \]  

which represents number of connected atoms in the MD snapshot. The number of connected atoms over the entire trajectory was used as a feature.

\[ E_k = \frac{\sum_{i=1}^N (D_{k,i})}{N} \]  

The positioning is evaluated using the Euclidean distance between the EGFR binding site atoms and the center of the drug-molecule.

\[ D(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} \]  

The binding site residues are represented by their alpha-carbon (CA) atoms. For example, if there are 14 CA atoms at the binding site, and two atoms at the drug molecule center, then a 14 \( \times \) 2 or 28 \( \times \) 1 vector will represent this. The distance over the entire MD simulation of 200 frames can be represented as a 200\( \times \)28 matrix. The binding site position is represented as the average distance between the drug and the target:

**Figure 11.** Atoms in convex and concave shapes at the surface curvature. The figure at the top shows a matched concave-convex pair, thus a strong interaction, while the figure at the bottom shows an unmatched pair and a weak interaction.
\[ D_{\text{avg}} = \frac{\sum_{i=1}^{N} D_i}{N} \]

where \( D_{\text{avg}} \) shows the binding site position, \( D_i \) shows the ith MD snapshot distance, and \( N \) is the number of MD snapshots. All the feature values were normalized to \([0,1]\). Generally, drug sensitive mutants have less distance between the drug and the target.

**Hydrogen bonds.** Hydrogen bonds contribute to the stability of a structure and can provide insights about interactions within the structure. Stable systems tend to have more hydrogen bonds. The number of hydrogen bonds in the EGFR-drug complex were calculated using the hbond command in Amber.

**Composite geometric features.** The geometric features were combined to make two composite features, with the matching rate, number of connected atoms, and number of hydrogen bonds as one feature, \( x_{\text{g1}} \) and the number of convex atoms and Euclidean distance, \( x_{\text{g2}} \), as the other. Similarly, the personal features and energy features were combined as \( x_{\text{p}} \) and \( x_{\text{e1}} \), respectively.

**Feature normalization.** Each feature was normalized to the range \([-1, 1]\) using z-score normalization:

\[ z_i = \frac{x_i - \mu}{\sigma} \]

where \( z_i \) represents the normalized value, \( \mu \) represents the mean value, and \( \sigma \) represents the standard deviation value for each feature. Composite geometric features, energy features, clinical features and their density distributions are presented in Fig. 12.

**Development of the classification model.** For patients with clinical information, geometrical and energy features were obtained from their EGFR mutant drug complex, and classifiers were trained to predict one of the four-classes of drug response level. Five popular classifiers i.e., KNN, SVM, Artificial Neural Network, Random Forest, and XGboost were tested using Python Scikit-learn and Tensorflow library. We also used CARET package of RStudio for feature visualization plots. We used 141 samples for training the model. The parameters were optimized using a GridSearch approach. To ensure that the predictions made by the model are not by chance, we also performed y-scrambling and randomly shuffled the labels for 50 iterations. The mean accuracy drops down to 0.48587 from 0.975. To further avoid over-fitting and careful selection of hyper-parameters and model, we used a nested cross-validation function, with an outer loop of 10 and inner loop of 3. The parameters are selected with a grid search of (10, 100, 500) estimators, (2, 4, 6, 10, 12) features and maximum depth of 3 for XGBoost and Random Forest. The final proposed XGBoost has a mean accuracy of 0.971 with a standard deviation of 0.035. The results vary between 0.8571 and 1 during cross validation. We also trained a neural network with four layers, with sigmoid activation for hidden layers and softmax for the output layer, for 8000 epochs. We used different activation functions, including Relu and dropout with different thresholds during the model selection. The categorical cross-entropy as the loss function, and RMSProp as an optimizer. We also implemented early stopping to monitor the validation loss with a patience value of 100. The source code is available on GitHub https://github.com/rizwanqureshi123/PDRP.

**Performance evaluation metrics.** Classification model evaluation performance metrics used were precision, recall, F1-measure, and balanced accuracy defined as follows:

\[ \text{Precision} = \frac{TP}{TP + FP} \]

\[ \text{Recall} = \frac{TP}{TP + FN} \]

\[ F1 - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ \text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \]

The terms TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively.

**Conclusion**

Computational methods, especially machine learning based techniques are widely used to predict responses to lung cancer drugs. In this work, we developed a machine learning based model (PDRP), that uses clinical and demographic information (DCI), energy, and geometrical features in machine learning classifiers to predict the four levels of drug response. PDRP achieved state-of-the-art performance at 97.5% accuracy with an XGBoost classifier, even though only a small number of patients’ had information available. Our model provides a personalized drug response level prediction with a highly accurate prediction rate that could be tested on other
types of cancer or other diseases. PDRP shows that modeling geometry, even in silico, can provide a powerful biomarker to predict the drug response of lung cancer patients. In future, we will further explore the dynamics and geometry of the binding-sites of protein-drug complexes. More clinical data will be collected to further refine the prediction model, and test it on other diseases.

Data availability
The datasets generated and/or analysed during the current study are not publicly available due non-disclosure agreement but are available from the corresponding author on reasonable request.

Received: 30 April 2022; Accepted: 3 November 2022
Published online: 07 November 2022

References
1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. CA Cancer J. Clin. (2022).
2. Gupta, G. P. & Massagué, J. Cancer metastasis: Building a framework. Cell 127, 679–695 (2006).
3. Qureshi, R. et al. Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).
4. Kawaguchi, T. et al. Randomized phase iii trial of erlotinib versus docetaxel as second-or third-line therapy in patients with advanced non-small-cell lung cancer: Docetaxel and erlotinib lung cancer trial (delta). J. Clin. Oncol. 32, 1902–1908 (2014).
5. Pao, W. et al. Egfr receptor gene mutations are common in lung cancers from ‘never smokers’ and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc. Natl. Acad. Sci. 101, 13306–13311 (2004).
6. Zhang, H. Three generations of epidermal growth factor receptor tyrosine kinase inhibitors developed to revolutionize the therapy of lung cancer. Drug Des. Dev. Ther. 10, 3867 (2016).

Figure 12. Geometric, energy and personal features and their distributions.
7. Singh, D., Attri, B. K., Gill, R. K. & Barwal, J. Review on egfr inhibitors: Critical updates. *Mini Rev. Med. Chem.* **16**, 1134–1166 (2016).
8. Tetsu, O., Hangauer, M. J., Phuchareon, J., Eisele, D. W. & McCormick, F. Drug resistance to egfr inhibitors in lung cancer. *Chemotherapy* **61**, 223–235 (2016).
9. Rho, J. K. et al. Combined treatment with silibinin and epidermal growth factor receptor tyrosine kinase inhibitors overcomes drug resistance caused by L790m mutation. *Mol. Cancer Ther.* **9**, 3233–3243 (2010).
10. Balbus, T. E. & Bizzo, R. C. Quantitative prediction of fold resistance for inhibitors of egfr. *Biochemistry* **48**, 8435–8448 (2009).
11. Tetsu, O., Phuchareon, J., Eisele, D. W., Hangauer, M. J. & McCormick, F. Akt inactivation causes persistent drug tolerance to egfr inhibitors. *Pharmacol. Res.* **102**, 132–137 (2015).
12. Guardiola, S., Varese, M., Sánchez-Navarro, M. & Giralt, E. A third shot at egfr: New opportunities in cancer therapy. *Trends Pharmacol. Sci.* **40**, 941–955 (2019).
13. Wang, D. D., Zhou, W., Yan, H., Wang, M. & Lee, V. Personalized prediction of egfr mutation-induced drug resistance in lung cancer. *Sci. Rep.* **3**, 1–8 (2013).
14. Karplus, M. & McCammon, J. A. Molecular dynamics simulation of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
15. Qureshi, R., Ghosh, A. & Yan, H. Correlated motions and dynamics in different domains of egfr with L858R and 790M mutations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (2020).
16. Wang, S. & Covey, R. V. Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. *J. Comput. Chem.* **32**, 2843–2852 (2011).
17. Qureshi, R., Zhu, M., Ghosh, A. & Yan, H. Computational analysis of structural dynamics of egfr and its mutants. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2784–2791 (IEEE, 2019).
18. Rizwan, Q., Zhu, M. & Yan, H. Visualization of protein-drug interactions for the analysis of drug resistance in lung cancer. *IEEE J. Biomed. Health Inform.* (2020).
19. Peng, Y. et al. Apatinib to combat egfr-tki resistance in an advanced non-small cell lung cancer patient with unknown egfr status: A case report. *Onco Targets Ther.* **10**, 2289 (2017).
20. Mehner, C. et al. Egfr as a prognostic biomarker and therapeutic target in ovarian cancer: Evaluation of patient cohort and literature review. *Gynecol. Cancer* **8**, 589 (2017).
21. Collins, F. S., Morgan, M. & Patrinos, A. The human genome project: Lessons from large-scale biology. *Science* **300**, 286–290 (2003).
22. Ashley, E. A. Towards precision medicine. *Nat. Genet.* **47**, 507–522 (2016).
23. Horer, A. & Ammenwerth, E. Electronic health records. *Methods Inf. Med.* **49**, 320–336 (2010).
24. Mok, T. S. Personalized medicine in lung cancer: What we need to know. *Nat. Rev. Clin. Oncol.* **8**, 661–668 (2011).
25. French, B. et al. Statistical design of personalized medicine interventions: The clarification of optimal anticagulation through genetics (coag) trial. *Trials* **11**, 1–9 (2010).
26. Ma, L., Wang, D. D., Zou, B. & Yan, H. An eigen-binding site based method for the analysis of anti-egfr drug resistance in lung cancer treatment. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **14**, 1187–1194 (2016).
27. Basit, S. A., Qureshi, R., Shahid, A. R. & Khan, S. Survival prediction of lung cancer patients by integration of clinical and molecular features using machine learning. In *2021 15th International Conference on Open Source Systems and Technologies (ICOSST)*, 1–6 (IEEE, 2021).
28. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
29. Wang, R. et al. Taxicore: Recommending road clusters to taxi drivers using ranking-based extreme learning machines. *IEEE Trans. Knowl. Data Eng.* **30**, 585–598 (2018).
30. Duan, B., Zou, B., Wang, D. D., Yan, H. & Han, L. Computational evaluation of egfr dynamic characteristics in mutation-induced drug resistance prediction. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2239–2304 (IEEE, 2015).
31. Zou, B., Lee, V. H. & Yan, H. Prediction of sensitivity to gefitinib/erlotinib for egfr mutations in nsclc based on structural interaction fingerprints and multinlinear principal component analysis. *BMC Bioinform.* **19**, 1–13 (2018).
32. Lee, V. H. et al. Association of exon 19 and 21 egfr mutation patterns with treatment outcome after first-line tyrosine kinase inhibitor in metastatic non-small-cell lung cancer. *J. Thoracic Oncol.* **8**, 1148–1155 (2013).
33. Ma, L. et al. Egfr mutant structural database: computationally predicted 3d structures and the corresponding binding free energies with gefitinib and erlotinib. *BMC Bioinform.* **16**, 1–10 (2015).
34. Zou, B. et al. Deciphering mechanisms of acquired L790M mutation after egfr inhibitors for nsclc by computational simulations. *Sci. Rep.* **7**, 1–13 (2017).
35. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210 (2013).
36. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using rosetta. In *Methods in Enzymology*, vol. 383, 66–93 (Elsevier, 2004).
37. Lencioni, R. & Llvet, J. M. Modified recist (mrecist) assessment for hepatocellular carcinoma. In *Seminars in Liver Disease*, vol. 30, 052–060 (Thieme Medical Publishers, 2010).
38. Covra, T. F. & Pais, A. A. Deep learning for deep chemistry: Optimizing the prediction of chemical patterns. *Front. Chem.* **7**, 809 (2019).
39. Kureshi, N., Abidi, S. S. R. & Blouin, C. A predictive model for personalized therapeutic interventions in non-small cell lung cancer. *IEEE J. Biomed. Health Inform.* **20**, 424–431 (2014).
40. Bamford, S. et al. The cosmic (catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
41. Shen, L. & Yang, W. Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks. *J. Chem. Theory Comput.* **14**, 1442–1455 (2016).
42. Kobayashi, Y. & Mitsudomi, T. Not all epidermal growth factor receptor mutations in lung cancer are created equal: Perspectives for individualized treatment strategy. *Cancer Sci.* **107**, 1179–1186 (2016).
43. Kellogg, E. H., Lever-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinform.* **79**, 830–838 (2011).
44. Martí-Renom, M. A. et al. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
45. Eisenberg, D., Lüthy, R. & Bowie, J. U. verify3d: Assessment of protein models with three-dimensional profiles. In *Methods in enzymology*, vol. 277, 396–404 (Elsevier, 1997).
46. Thakur, Z., Dharra, R., Saini, V., Kumar, A. & Mehta, P. K. Insights from the protein-protein interaction network analysis of mycobacterium tuberculosis toxin-antitoxin systems. *Bioinformatics* **13**, 380 (2017).
47. Case, D. A. et al. Amber 10 (University of California, 2008).
48. Zhou, C.-Y., Jiang, F. & Wu, Y.-D. Residue-specific force field based on protein coil library rlff2: Modification of amber ff99sb. *J. Phys. Chem. B* **119**, 1035–1047 (2015).
49. Kräutler, V., Van Gunsteren, W. F. & Hünenberger, P. H. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–5 (2001).
50. Aktulga, H. M., Fogarty, J. C., Pandit, S. A. & Grama, A. Y. Parallel reactive molecular dynamics: Numerical methods and algorithmic techniques. *Parallel Comput.* **38**, 245–259 (2012).
51. Gotz, A. W. et al. Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born. J. Chem. Theor. Comput. 8, 1542–1552 (2012).
52. Roe, D. R. & Cheatham, T. E. III. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. J. Chem. Theor. Comput. 9, 3084–3095 (2013).
53. Gohlke, H., Kiel, C. & Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the ras-raf and ras-raldgs complexes. J. Mol. Biol. 330, 891–913 (2003).
54. Ozturk, H., Oegir, A. & Ozkirimli, E. Deepdfa: Deep drug-target binding affinity prediction. Bioinformatics 34, i821–i829 (2018).
55. Genheden, S. & Ryde, U. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. Expert Opin. Drug Discov. 10, 449–461 (2015).
56. Reddy, M. R. & Erion, M. D. Calculation of relative binding free energy differences for fructose 1, 6-bisphosphatase inhibitors using the thermodynamic cycle perturbation approach. J. Am. Chem. Soc. 123, 6246–6252 (2001).
57. Naderi, M. et al. Binding site matching in rational drug design: Algorithms and applications. Brief. Bioinform. 20, 2167–2184 (2019).
58. Wilson, J. A., Bender, A., Kaya, T. & Clemons, P. A. Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors. J. Chem. Inf. Model. 49, 2231–2241 (2009).
59. Fabri, A., Grezeman, G.-I., Kettner, L., Schirra, S. & Schönherr, S. On the design of cgal a computational geometry algorithms library. Softw. Pract. Exp. 30, 1167–1202 (2000).
60. Ma, L., Zou, B. & Yan, H. Identifying egfr mutation-induced drug resistance based on alpha shape model analysis of the dynamics. Proteome Sci. 14, 12 (2016).
61. Saranya, C. & Manikandan, G. A study on normalization techniques for privacy preserving data mining. Int. J. Eng. Technol. (IJET) 5, 2701–2704 (2013).
62. Pedregosa, F. et al. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
63. Géron, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (O’Reilly Media Inc, 2019).
64. Kuhn, M. et al. Building predictive models in r using the caret package. J. Stat. Softw. 28, 1–26 (2008).
65. Parvandeh, S., Yeh, H.-W., Paulus, M. P. & McKinney, B. A. Consensus features nested cross-validation. Bioinformatics 36, 3093–3098 (2020).
66. Fruzna, O., Inkpen, D. & Tran, T. A machine learning approach for identifying disease-treatment relations in short texts. IEEE Trans. Knowl. Data Eng. 23, 801–814 (2010).
67. Hao, G.-F., Yang, G.-F. & Zhan, C.-G. Structure-based methods for predicting target mutation-induced drug resistance and rational drug design to overcome the problem. Drug Discov. Today 17, 1121–1126 (2012).

Acknowledgements
This work is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), the Hong Kong Research Grants Council (Project 11204821) and City University of Hong Kong (Project 9610034). The open access publication of this article was funded by College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Qatar.

Author contributions
R.Q. contributed in conceptualization, experiments, writing, and modeling; S.A.B. contributed to developing the classification model and drawing illustrations; X.F. and M.N. assisted in modeling, conceptualization, and writing; H.Y. provided the dataset and assisted in revision of the manuscript; J.A.S. helped with writing the manuscript; T.A. contributed to conceiving, designing experiments, designing machine learning models, analysing results, writing manuscript.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-23649-0.

Correspondence and requests for materials should be addressed to T.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022