communications
chemistry

ARTICLE

https://doi.org/10.1038/s42004-023-00920-7 OPEN

# A mutation-induced drug resistance database (MdrDB)

Ziyi Yang[1,2], Zhaofeng Ye[1,2], Jiezhong Qiu[1], Rongjun Feng[1], Danyu Li[1], Changyu Hsieh[1], Jonathan Allcock[1] & Shengyu Zhang[1✉]

Mutation-induced drug resistance is a significant challenge to the clinical treatment of many diseases, as structural changes in proteins can diminish drug efficacy. Understanding how mutations affect protein-ligand binding affinities is crucial for developing new drugs and therapies. However, the lack of a large-scale and high-quality database has hindered the research progresses in this area. To address this issue, we have developed MdrDB, a database that integrates data from seven publicly available datasets, which is the largest database of its kind. By integrating information on drug sensitivity and cell line mutations from Genomics of Drug Sensitivity in Cancer and DepMap, MdrDB has substantially expanded the existing drug resistance data. MdrDB is comprised of 100,537 samples of 240 proteins (which encompass 5119 total PDB structures), 2503 mutations, and 440 drugs. Each sample brings together 3D structures of wild type and mutant protein-ligand complexes, binding affinity changes upon mutation (ΔΔG), and biochemical features. Experimental results with MdrDB demonstrate its effectiveness in significantly enhancing the performance of commonly used machine learning models when predicting ΔΔG in three standard benchmarking scenarios. In conclusion, MdrDB is a comprehensive database that can advance the understanding of mutation-induced drug resistance, and accelerate the discovery of novel chemicals.

[1] Tencent Quantum Laboratory, Shenzhen 518057 Guangdong, China. [2] These authors contributed equally: Ziyi Yang, Zhaofeng Ye. ✉email: shengyzhang@tencent.com

Structural mutation of proteins can directly affect their folding and stability[1–3], function[4,5], interactions with other proteins[6,7] and binding affinity[8]. In some cases, it can result in significant perturbations or even complete abolishment of protein function, potentially leading to diseases or cancer[9,10]. The evolutionary pressure imposed by small molecule drugs on many quickly evolving systems, including cancer cells, viruses, and bacteria, can lead to the rapid development of resistance[11–14]. While novel and cheap high-throughput sequencing technologies have made it possible to identify mutations in large populations[15,16], the significance and characteristics of any novel polymorphisms currently require time-consuming and expensive experiments to determine[17]. Protein-ligand binding affinity data are of great value for understanding the impact of polymorphisms on disease and identifying mutations that lead to drug resistance[18,19]. Convenient and broad access to such data for wild-type and mutant proteins would aid our understanding of the mechanisms of mutation-induced drug resistance, increase the accuracy of extrapolations to novel mutations and systems, enable more effective computational approaches for drug resistance prediction, and facilitate the development of combination therapies and the discovery of novel chemicals.

Fortunately, a number of databases on the effects of mutation on protein-ligand binding affinity have been compiled and released in recent years. In particular, Platinum[17] was the first database to provide experimental data on changes in protein-ligand affinities upon mutation, along with three-dimensional structures, while tyrosine kinase inhibitor (TKI)[18] contains reliable inhibitor ΔpIC50 data for 144 clinically identified mutants of the human kinase Abl. Together, these two datasets are the current gold standards, both providing well-curated data and three-dimensional structural information. Other resources include Auto In Silico Macromolecular Mutation Scanning (AIMMS)[20], RET[21] and kinase mutations and drug response (KinaseMD)[22], which provide kinase inhibitor resistance information. However, while the value of these datasets is significant, there remain a number of deficiencies. Platinum and TKI are restricted to known cocrystal structures and thus contain relatively few samples (approximately 1000 and 150, respectively), which can limit their use with machine learning models. For AIMMS, RET, and KinaseMD, structural information is not provided. In addition, while all of the aforementioned databases consider single-point and multi-site substitution mutations, they do not include more varied and complex mutation types—such as deletions, insertions, and insertion-deletions (indel)—which play an important role in certain disease progressions.

To tackle these challenges, we have created MdrDB, an all-encompassing database that significantly expands the amount of information on drug resistance. MdrDB combines the five databases mentioned above and achieves considerable augments by incorporating large-scale drug response data, and somatic mutation information from various cancer cell lines, as reported in the Genomics of Drug Sensitivity in Cancer (GDSC) dataset[23], and Cancer Dependency Map (DepMap)[24]. Further supplemented with data from supporting databases such as RCSB Protein Data Bank (PDB), UniProtKB and PubChem, and 3D structures computed with PyMOL and AlphaFold 2.0[25], MdrDB brings together 3D structures of wild-type and mutant protein-ligand complexes, binding affinity changes upon mutation (ΔΔG), and biochemical features calculated from complex structures. MdrDB contains 100,537 samples, generated from 240 proteins (5119 total PDB structures), 2503 mutations, and 440 drugs. A total of 95,971 samples are based on available PDB structures, and 4566 samples are based on AlphaFold 2.0 predicted structures.

MdrDB offers several key advantages over existing publicly available protein mutation databases. Firstly, MdrDB offers massive and diverse data. With over 100,000 samples, MdrDB integrates information from multiple sources, includes various mutation types beyond single site-point mutation and covers mutations across a broad range of protein families, which is the largest mutation-induced drug resistance database to our knowledge. Secondly, MdrDB provides 3D structural information on all wild-type and mutant proteins along with 146 associated biochemical features, which are useful and important in accurate drug resistance modeling. Furthermore, we evaluated the classical drug resistance prediction machine learning models' performance on three standard benchmarking scenarios. By using MdrDB as a training set, nearly all models gain significant performance improvement. In summary, MdrDB has the size, breadth, and complexity to offer new resources for protein mutation studies and drug resistance modeling.

## Results

**Database statistics**. At the time of writing, MdrDB contains 100,537 samples, generated from 240 proteins (5119 total PDB structures), 2503 mutations, and 440 drugs. In all, 95,971 of the total samples are based on available PDB structures, with the remaining 4566 samples predicted with AlphaFold 2.0. In addition to single-point substitution mutations and multiple-point substitution mutations, MdrDB also contains complex mutations including deletion, insertion, and indel (insertion-deletion) mutations, as well as multiple-site mutations containing a number of the aforementioned mutations. Table 1 summarizes the data contained in MdrDB, and an overview of the data sources is shown in Supplementary Table 1. Figure 1A shows MdrDB samples categorized by their mutation types. Approximately 83.6% of the samples are single substitution mutations; 11.9% are multiple substitution mutations; and 4.5% are complex mutations, of which deletion mutations account for the largest proportion.

The distribution of mutation-induced ligand binding affinity changes, measured by Δ$\Delta G$, is shown in Fig. 1B. A standard for defining resistant samples was given in[18], where samples with a larger than 10-fold drop in binding affinity, corresponding to $\Delta\Delta G > 1.36$ kcal mol$^{-1}$, are considered to be resistant. Based on this standard, 8197 (8.2%) samples in MdrDB are mutation resistant.

We analyzed the distribution of amino acid types in substitution mutation samples, before and after mutation, to shed light on mutation preferences. The 20 natural amino acids were divided into five groups: positively charged, negatively charged, polar, hydrophobic, and three amino acids classified as special cases (Fig. 1C, Supplementary Fig. 1A, and Supplementary Tables 2 and 3). Overall, 30.8% of mutations are observed within the same amino acid group, while 69.2% of mutations occur across different amino acid groups. Arginine (R) is the most frequently mutated amino acid, followed by glycine (G) and valine (V), which are reported to contribute to human genetic diseases[26]. Leucine (L), alanine (A), serine (S) and cysteine (C) are the most frequently mutated amino acids. Similar conclusions hold for the MdrDB_CoreSet (Supplementary Figs. 1B and 2 and Supplementary Tables 4 and 5), where non-repetitive samples considering only unique "protein-drug-mutations" are kept.

The mutation frequencies of different amino acids at the mutation sites are highly correlated with the frequencies calculated based on codon frequencies (Pearson $r = 0.945$, Supplementary Figs. 3A–C and Supplementary Table 6)[26]. However, the frequencies of amino acids after the mutation show a lower correlation with the frequencies calculated based on codon frequencies (Pearson $r = 0.633$, Supplementary Figs. 3D–F)[26]. Many complex factors could contribute to this, such as epigenetic

**Table 1 Overview of data represented in MdrDB.**

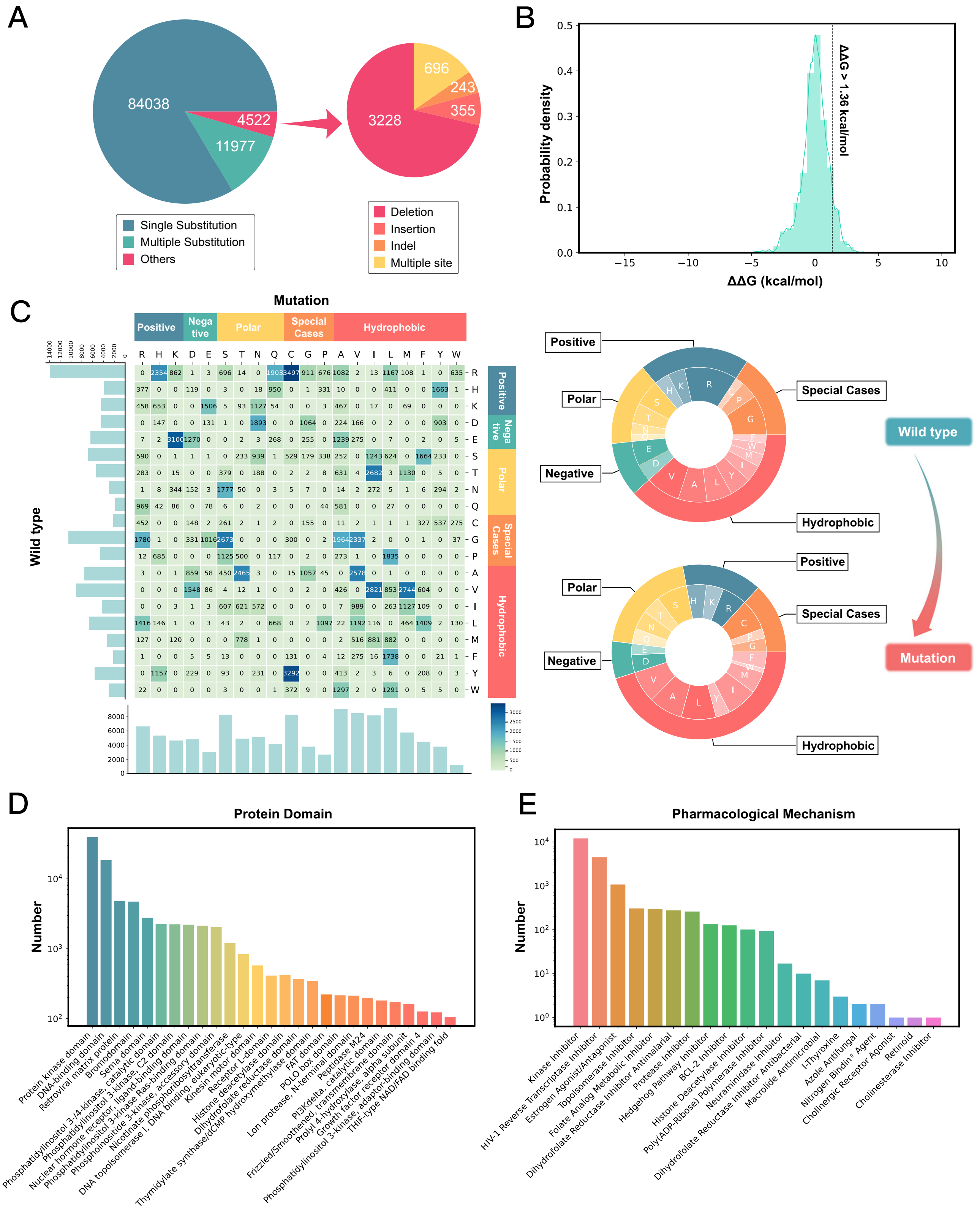
| Property | No. of counts | No. of counts in source datasets | | | | |
|---|---|---|---|---|---|---|
| No. of samples | 100,537 | AIMMS (5048), DepMap (1500), GDSC (91,215), KinaseMD (1334), Platinum (840), RET (456), TKI (144) | | | | |
| No. of unique UniProt IDs | 240 | AIMMS (10), DepMap (5), GDSC (99), KinaseMD (31), Platinum (126), RET (1), TKI (4) | | | | |
| No. of unique ligands | 440 | AIMMS (16), DepMap (5), GDSC (180), KinaseMD (48), Platinum (126), RET (3), TKI (8) | | | | |
| No. of mutations | 2503 | AIMMS (121), DepMap (49), GDSC (1805), KinaseMD (43), Platinum (530), RET (11), TKI (31) | | | | |
| No. of unique UniProt ID-Mutations | 2673 | AIMMS (173), DepMap (49), GDSC (1818), KinaseMD (43), Platinum (578), RET (11), TKI (80) | | | | |
| No. of unique PDB IDs | 5119 | AIMMS (430), DepMap (220), GDSC (4128), KinaseMD (802), Platinum (241), RET (21), TKI (8) | | | | |
| | No. of Platinum processed | No. of TKI processed | No. of PyMOL mutated | No. of AlphaFold 2 folded | | |
| Mutant structure source | 840 | 144 | 94,987 | 4566 | | |
| | No. of Platinum cocrystal | No. of TKI cocrystal | No. of TKI docked | No. of docked | | |
| Drug pose source | 840 | 131 | 13 | 99,553 | | |
| | No. of single substitution | No. of multiple substitutions | No. of deletion | No. of insertion | No. of indel | No. of complex |
| Mutation types | 84,038 | 11,977 | 3228 | 355 | 243 | 696 |

modifications[27], disease preferences[28], co-evolution[29], and experimental assay biases[30]. In addition, we analyzed the mutation spectrums for each type of amino acid, which revealed mutation preferences. Interestingly, as shown in Supplementary Fig. 4, most of the amino acids have a mutation spectrum that is similar to the one based on codon frequencies (Pearson $r > 0.8$). For the amino acids which deviate the most from this expected spectrum, a higher ratio proportion of lysine (K), glutamine (Q), cysteine (C), and tryptophan (W) are mutated to A, which may come from the loss of function mutation experiments, such as alanine scanning mutagenesis[30]. We carried out the same analyses on susceptible/resistant samples (Supplementary Fig. 5), and observed an obvious decrease in the frequency correlation of resistant samples before mutation (Pearson $r = 0.874$), where serine (S), lysine (K), and tyrosine (Y) showed the largest differences. For the frequency of each amino acid after mutation in the susceptible /resistant samples, although the correlations are similar, their distributions vary greatly.

Furthermore, we analyzed the secondary structure elements where mutations occurred across all samples (Supplementary Fig. 6) as well as in resistant samples (Supplementary Fig. 7). As shown in Supplementary Figs. 6 and 7, the $\Delta\Delta G$ value distributions were not significantly different across samples whether mutations are located in $\alpha$-helices, $\beta$-sheets or loops.

Annotations for the proteins and drugs are provided for each sample in MdrDB. For proteins, a total of 160 kinds of protein domains, 154 protein families and 143 protein superfamilies can be found in MdrDB. The protein domains with at least 100 samples are shown in Fig. 1D and Supplementary Table 7. Protein kinase domains account for the largest proportion (39.5%) of all samples, followed by DNA-binding domains (18.5%), retroviral matrix proteins (4.8%) and bromodomains (4.7%). For drug mechanism annotations, a total of 20 FDA-documented pharmacological classes, 60 Medical Subject Headings (MeSH) pharmacological classes and 9 PubChem drug classes can be found in MdrDB. 67 (15.2%) out of all 440 drugs have known FDA pharmacological classes, and the corresponding samples were counted and shown in Fig. 1E and Supplementary Table 8. The largest proportion of samples takes kinase inhibitors as drugs (12.0%), followed by HIV-1 reverse transcriptase inhibitors (4.5%).

**Model performance evaluation**. Rapid and accurate computational methods could impact clinical decision making by providing oncologists with an initial indication of whether an observed protein mutation may lead to drug resistance to certain inhibitors. Molecular dynamics (MD)-based free-energy calculations and physics- and knowledge-based potential scoring functions (e.g., Rosetta) are two common types of computational methods for estimating affinity changes in proteins with mutations[31–36]. Although these two types of methods can achieve remarkable performance and serve as a gold standard, they suffer from high computational overheads. Data-driven machine learning methods have recently been developed to predict the impact of ligand binding affinity changes upon protein mutations and identify resistance-causing mutations[19,37], and can make predictions within seconds once the features associated with protein mutation affinity changes are input. Current machine learning methods, however, are prone to overfitting, and limited training data—expensive to acquire—reduces their performance stability in real applications. For instance, Aldeghi et al.[19] employed extremely randomized regression trees to predict TKI affinity change values. Their model—trained on a subset of Platinum database (no tyrosine kinase) with 484 training samples and then tested on the TKI dataset—achieved low performance in



terms of correlation and classification ability. To alleviate the overfitting problem, our previous work[38] incorporated additional physics-based structural features and protein family information as inputs to learn more comprehensive knowledge from limited training data. The empirical results show that the method we previously proposed is slightly superior to MD simulations with the Amber force fields. In this section, we investigated three scenarios to test whether using the MdrDB_CoreSet as the training dataset (using 146 calculated biochemical features) can improve the model generalization performance in predicting mutation-induced drug resistance in the TKI dataset.

We conducted experiments on the TKI dataset under three scenarios. The first scenario was to train the machine learning methods on the subset of the Platinum dataset, which had no

**Fig. 1 MdrDB mutation statistics, ΔΔG distribution, and protein and drug annotations. A** Number of samples of each mutation type, where a sample corresponds to a (Uniprot, PDB, mutation, drug) combination. "Others" includes four mutation types: deletion, insertion, indel, and complex. **B** Histogram of protein mutation-induced ligand binding affinity changes measured by ΔΔG (kcal $mol^{-1}$). Line at ΔΔG = 1.36 kcal $mol^{-1}$ separates mutations defined as resistant from susceptible. **C** Number of amino acid changes from substitution mutations. Left: heatmap shows the number of amino acid changes from substitution mutations, with different colors corresponding to the number of changes in different amino acids. The number of samples for each amino acid in the wild-type (mutation) is displayed as a bar chart along the left axis (bottom) of the plot. Right: donut charts show the number of amino acids in the samples corresponding to substitution mutations. The proportion of each amino acid in wild-type (mutated) samples is displayed top (bottom). **D** Number of samples annotated by protein domains. *X*-axis gives protein domain, with *Y*-axis the corresponding sample number on logarithmic scale. **E** Number of samples annotated by pharmacological mechanisms.

tyrosine kinase samples, and then test them on the TKI dataset. This setting is consistent with the setting in work[19,38]. The second scenario was to train the machine learning methods on the Platinum dataset, and test them on the TKI dataset. The third scenario was to train the machine learning methods on a subset of MdrDB_CoreSet, removing samples derived from the TKI dataset and selecting samples corresponding to single substitutions, and then test them on the TKI dataset. Since the TKI dataset contains only the single substitution type, we select samples belonging to a single substitution in MdrDB_Coreset. Compared with Scenario 1, the latter two scenarios evaluate whether adding a small amount of tyrosine kinase information or increasing samples and protein family information in the training dataset can improve the capability of the model in predicting affinity changes in Abl kinase mutants. Detailed information on the training datasets for the three scenarios is given in Supplementary Table 9. According to the feature calculation section, 146 features were calculated as model inputs, carrying potentially useful information for predicting affinity changes introduced by protein mutations.

Four families of methods were used to evaluate the drug resistance prediction performance under three scenarios. The first family is tree-based methods, including decision tree (DecisionTree)[39], random forest (RandomForest)[40], and extremely randomized regression trees (ExtraTrees)[41]. The second family is linear-based methods containing support vector regression (SVR)[42], elastic net linear regression (Elastic Net)[43], and lasso regression (Lasso)[44]. The third family of baselines is ensemble-based methods including bagging regressor (Bagging)[45], AdaBoost[46], and gradient boosting (GradienBoost)[47]. The fourth family is neural network-based methods such as multi-layer perceptron[48].

Root mean square error (RMSE), Pearson correlation coefficient (Pears), and the area under the precision-recall curve (AUPRC) were used to evaluate the model performance under three training scenarios[19]. Consistent with the previous work[19,38], resistant mutations are defined as the affinity changes for mutants by least 10-fold, i.e., $\Delta\Delta G_{exp}$ >1.36 kcal $mol^{-1}$.

We train the machine learning methods on different training datasets (i.e., Platinum (no tyrosine kinase), Platinum, and MdrDB_Coreset (single substitution)) and report the test performance averaged over five repetitions for each method on the TKI dataset in Fig. 2A. We can clearly see that the machine learning methods obtained poor performance in terms of correlation and classification ability when training on Platinum (no tyrosine kinase). For instance, ExtraTrees achieved low accuracy on TKI dataset with weak correlation (Pearson = 0.075), and poor classification performance (AUPRC = 0.198). When the training dataset provided a small amount of tyrosine kinase information, i.e., training on the Platinum dataset, the test prediction performance of most machine learning methods improved slightly. ExtraTrees obtained slightly better estimates (RMSE = 0.907, Pearson = 0.094, and AUPRC = 0.243) compared with its training on Platinum (no tyrosine kinase). When the machine learning methods were trained on the subset of MdrDB_CoreSet, we found that the prediction accuracy of most machine learning methods was improved on the TKI dataset compared to the previous two scenarios, except for AdaBoost, DecisionTree, and SVR. In particular, ExtraTrees, GradienBoost, Bagging, and RandomForest have significantly improved prediction performance. For instance, ExtraTrees achieved highly accurate on the TKI dataset when training on the subset of MdrDB_CoreSet with low absolute errors (RMSE = 0.656 kcal $mol^{-1}$), strong correlation (Pearson = 0.607), and good classification performance (AUPRC = 0.538). This result outperformed MD simulations with the Amber force fields (e.g., A99 and A99l) reported in the work[19] by a considerable margin, demonstrating that MdrDB_CoreSet could improve the generalization of the machine learning method in predicting affinity changes in Abl kinase mutants. Figure 2B shows the scatter plots of the experimental versus calculated Δ$\Delta G$ values of each machine learning method obtained at a certain time by training on the subset of MdrDB_CoreSet and testing on the TKI dataset. The scatter plots of the experimental versus calculated Δ$\Delta G$ values in the first and second scenarios are shown in Supplementary Figs. 8 and 9.

In addition, we also provide a comprehensive evaluation of 10 common machine learning models in several different scenarios (Supplementary Figs. 10–23 and Tables 10–23), and provide baseline prediction results on the MdrDB database. It is our hope that this will help further the development of new machine learning algorithms using the MdrDB database and facilitate drug resistance research. Please refer to Supplementary Note 1 for details.

## Conclusion

Here we have introduced MdrDB, the largest drug resistance database to provide all information highly relevant to protein mutation-induced drug binding affinity changes. In addition to basic information regarding the proteins, drugs, mutations and changed affinity, and structural data for wild-type and mutant complexes, 146 calculated biochemical features and extra annotations are also provided. These features and annotations were chosen to be convenient to use for model training and sample splitting in machine learning algorithms for drug resistance prediction[38]. In addition, MdrDB is also the first database that includes drug resistance-related mutation types beyond substitution mutations, and the availability of wider and more complex mutation types can be used to test the generalization of machine learning models.

A comprehensive resource providing a variety of information for studying protein mutation, predicting drug resistance and discovering novel chemical compounds, MdrDB will be updated regularly and include more public data in the future.

## Methods

The MdrDB database construction pipeline is shown in Fig. 3. Data preparation includes four steps: (1) data collection from the seven publicly available datasets. (2) Data preprocessing and integration of information from other databases, so that each sample contains information on the protein, drug, mutation and ΔΔG. (3) 3D structure generation for the wild-type and mutant protein-drug complexes. (4) Calculation of biochemical features based on the complex structures.

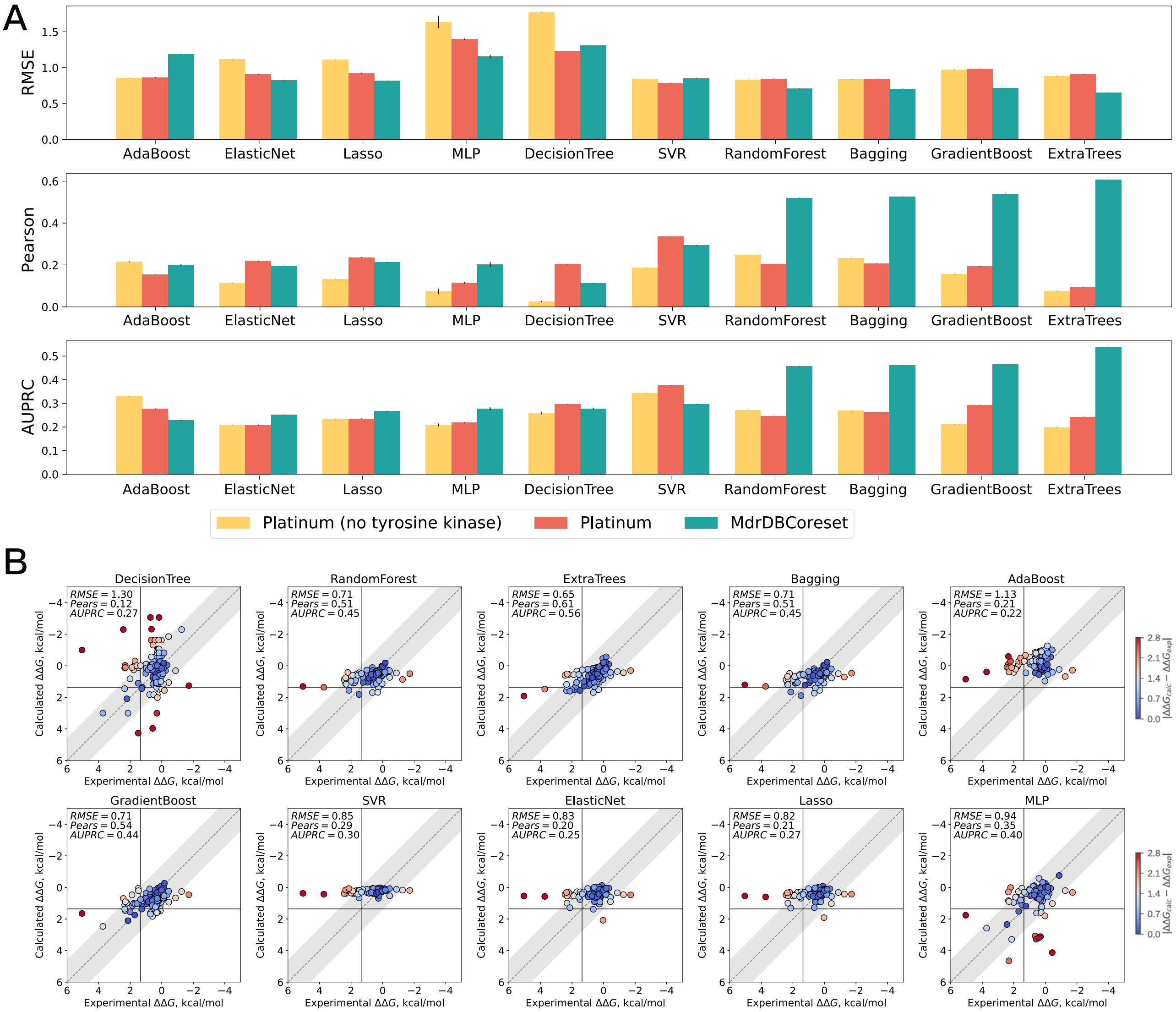


**Fig. 2 Machine learning methods performance evaluation. A** Summary of the test performance of the ΔΔG prediction across machine learning methods under three training scenarios in terms of RMSE, Pearson correlation, and AUPRC. Means with error bars (standard deviation) for all competing methods. **B** Scatter plots of the experimental versus calculated ΔΔG values in Scenario 3. *X*-axis denotes the experimental ΔΔG values (kcal $mol^{-1}$). *Y*-axis denotes the calculated ΔΔG value (kcal $mol^{-1}$). Each Δ$\Delta G$ estimate is color-coded according to its absolute error *w. r. t.* the experimental Δ$\Delta G$ value; at 300 K, the 1.4 kcal $mol^{-1}$ error corresponds to a 10-fold error in the $K_d$ change and 2.8 kcal $mol^{-1}$ error corresponds to a 100-fold error in the $K_d$ change.

**Data collection**. MdrDB contains data from seven publicly available datasets: Platinum[17], AIMMS[20], TKI[18], RET[21], KinaseMD[22], GDSC[23], and DepMap[24]. These datasets differ slightly in terms of the information they contain.

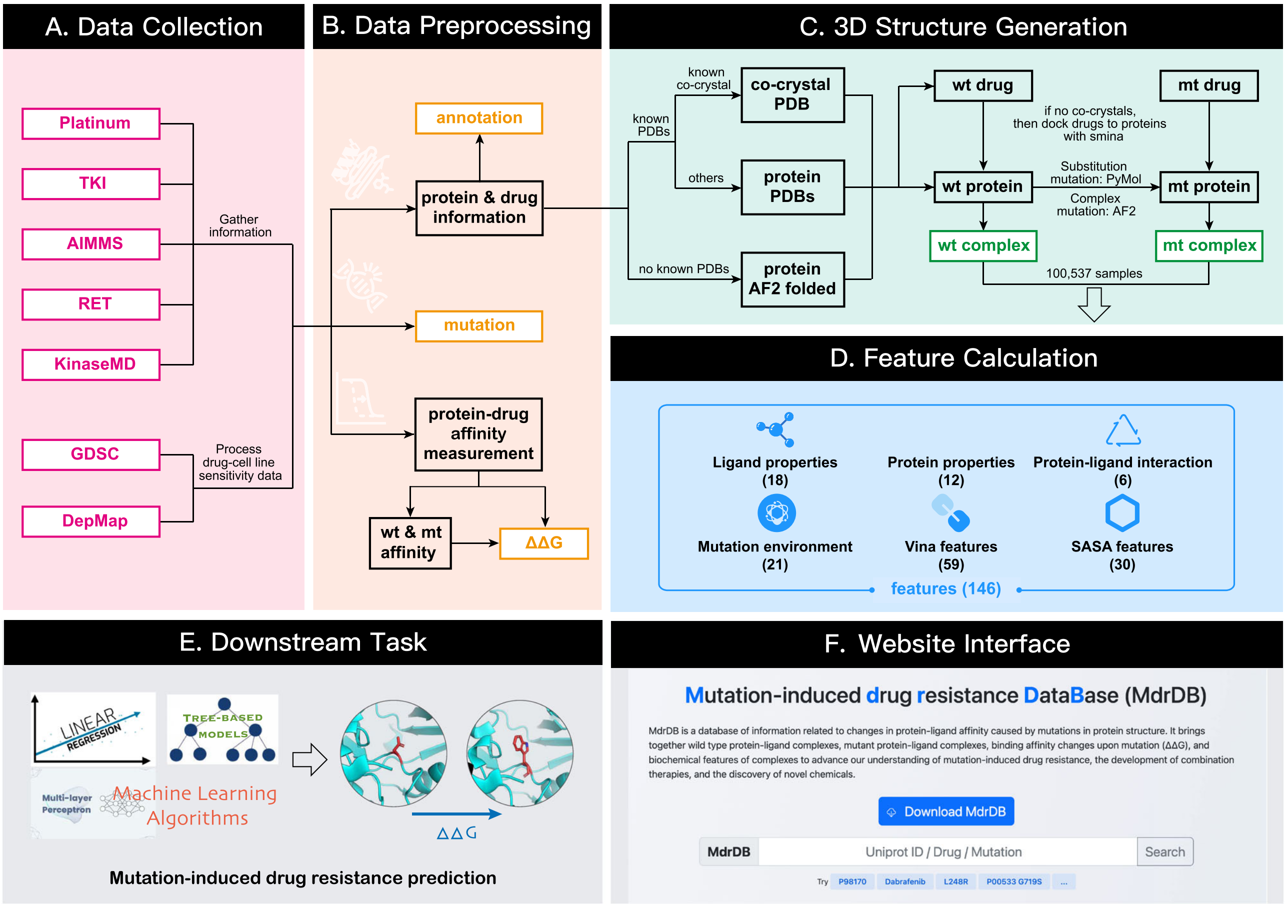
- Platinum: provides data on affinity changes upon site mutation (ΔΔG) and experimental cocrystal structures of protein-drug complexes from the RCSB PDB. It contains more than 1000 manually curated mutations[17].
- AIMMS: a web server for predicting site mutation-induced drug resistance, and contains a dataset of 17 protein-ligand complexes involving 311 ΔΔG values and mutations[20].
- TKI: a database of TKIs resistances in ABL tyrosine kinase site mutations. A total of 144 ΔΔG values are included, along with ABL-TKI complex structures from PDB[18].
- RET is a dataset specifically related to drug resistances between three drugs (cabozantinib, lenvatinib, vandetanib) and the RET kinase mutants. 56 IC50 measurements and protein-drug complex structures were reported in the research[21].
- KinaseMD: focuses on kinase mutations, and integrates information from the Cancer Cell Line Encyclopedia (CCLE)[49] and GDSC[23] databases, and provides various annotations of drug responses on kinase mutants. In all, 79 ΔΔG values were collected from this database[22].
- GDSC: the largest public resource for information on drug sensitivity in cancer cells. The database contains 4.4 million drug sensitivity (IC50) values across 518 drugs and 1000 cancer cell lines. It also records information related to basal expression, mutation, copy number variation, and gene methylation in the cell lines.
- DepMap: similar to GDSC, and contains drug sensitivity data on cell lines from the Achilles[50] and CCLE projects[51].

For Platinum, AIMMS, TKI, RET and KinaseMD, mutation information and mutation-induced binding affinity changes are directly obtainable from the datasets. However, for GDSC and DepMap, extra steps are required to obtain the mutations and ΔΔG values (see next section).

**GDSC/DepMap raw data processing**. Mutation and ΔΔG values were obtained from these two datasets via a process similar to that used by KinaseMD: (i) mutation information acquisition; (ii) drug-cell line response calculation; and (iii) ΔΔG calculation.

Cell lines were first grouped into wild-type and mutated samples for a specific protein, and mutation information for the protein was gathered. Specifically, for a particular protein, cell lines that did not have mutations on them were considered to be control (wild-type samples), while cell lines with mutations were considered to be mutant. Mutations were then filtered to retain only six types of mutation in terms of amino acid changes in protein sequences:

- Single substitution mutation: replacement of one amino acid in a protein with a different amino acid.



Mutation-induced drug resistance DataBase (MdrDB)

MdrDB is a database of information related to changes in protein-ligand affinity caused by mutations in protein structure. It brings together wild type protein-ligand complexes, mutant protein-ligand complexes, binding affinity changes upon mutation (ΔΔG), and biochemical features of complexes to advance our understanding of mutation-induced drug resistance, the development of combination therapies, and the discovery of novel chemicals.

**Fig. 3 The MdrDB database construction pipeline. A** Data collection from public datasets. **B** Data preprocessing to extract sample information. **C** 3D structure generation of protein-drug complexes for both wild-type and mutant proteins. **D** Feature calculation. **E** Downstream task: mutation-induced drug resistance prediction. **F** Website construction for browsing, search and downloading of the data. Color-coded boxes: red—the collected publicly available datasets, the number of original samples is shown; yellow—annotations, mutation information and ΔΔ*G* values, provided in MdrDB in .tsv file format; green—structural information provided in MdrDB.

- Multiple substitution mutation: more than one amino acid substitution in a protein.
- Deletion mutation: removal of one amino acid or an amino acid sequence in a protein.
- Insertion mutation: addition of one amino acid or an amino acid sequence in a protein.
- Indel mutation: a deletion mutation followed by an insertion mutation, i.e., replacement of an amino acid sequence in a protein with another amino acid sequence.
- Complex mutation: combinations of the above five amino acid mutation types.

Then, we filtered the drugs with known protein targets and queried drug sensitivity (IC50 values) data on all cell lines. For drugs with multiple known protein targets, each protein was considered individually. If, in the cell line, only one of these proteins was mutated, the cell line was kept. Otherwise, the cell line was discarded. After this, for each (protein, cell line, drug) mutant sample, the mutation string was generated by merging all mutations of the mutant protein to obtain a (protein, cell line, drug, mutation) sample. Then IC50 values were averaged over all cell lines for the (protein, cell line, drug, mutation) sample following the sensitivity data processing in KinaseMD[22]. This average value was taken as the mutant protein-drug affinity value IC50(mut). A corresponding wild-type sample was similarly assigned and IC50s again averaged over cell lines, and this average value was taken as the wild-type protein-drug affinity IC50(wt).

Finally, these affinity pairs were used to calculate ΔΔG according to the formula[18,37]:

$$\Delta\Delta G_{\mathrm{exp}} = RT\ln\frac{K_{\mathrm{i,mut}}}{K_{\mathrm{i,wt}}} \approx RT\ln\frac{\mathrm{IC}_{50,\mathrm{mut}}}{\mathrm{IC}_{50,\mathrm{wt}}} \qquad (1)$$

**Sample data consolidation**. From all seven datasets, five basic pieces of information were collected and consolidated: protein name, UniProt ID, mutation string, drug name, and ΔΔG value. UniProt IDs were identified by querying UniProtKB[52] with protein names. The consolidated data was then divided into separate tables, with samples of the same mutation type kept together. For samples from Platinum and TKI, wild-type and mutant protein-drug complex structures were also obtained from the original datasets. For protein annotations, we used the Interpro API[53] to query the protein family, homologous superfamily and domain information. For drug annotations, we used the PubChem PUG REST API[54] to query the CID, Depositor-Supplied Synonyms, FDA mechanism, MeSH and Drug Classes information.

**Three-dimensional structure generation**. With the exception of samples from Platinum and TKI, which already directly include 3D structure information, we prepared 3D structures of wild-type proteins, mutant proteins and drug binding poses for all samples from the other datasets.

For the preparation of wild-type protein structure, first for each protein (UniProt ID), all associated PDBs were identified with the RCSB REST API[55]. The .pdb or .cif for 3D structures and .fasta for sequences were downloaded from RCSB PDB. Then, all water, solvent, and ions were removed from the PDB files. Next, the protein chains and ligands were split into separate files. Each chain was annotated and only the chains corresponding to the UniProt ID were kept. If multiple chains were found to exist for protein, the longest one was kept. Meanwhile, the largest ligand was kept as the docking box generation reference. Finally, each mutation for a protein was checked against all available PDBs. If the mutation sites could be found in the PDB, the mutation and drug would be assigned to the PDB.

For the mutant protein structure generation, we generated mutant structures from wild-type structures using PyMOL[56] or from wild-type sequences using AlphaFold 2.0[25]. Specifically, the Mutagenesis Wizard module of pymol-open-source v2.5.0 was used to generate a mutation by replacing a residue with a new amino acid type, sample several rotamers from the rotamer library and generate several non-clashing conformations. Then, the most likely rotamer was chosen as the mutated residue. For AlphaFold 2.0, we used the protein amino acid sequence as the input to predict the structures. A length threshold of 2000 was set for computing resource considerations. For Multiple Sequence Alignment searching, “reduced_db” was used. For inference, the “monomer_ptm” models were used. Five models were generated and the one with the highest averaged pLDDT (predicted lDDT-Cα) value was chosen as the predicted structure for further procedures. Several rules were followed to decide which tool was used for mutant structure generation. For proteins with known PDBs containing the mutation sites:

- For single substitution and multiple substitution mutations, PyMOL was used for mutant protein generation.
- For deletion, insertion, indel and complex mutations, AlphaFold 2.0 was used for mutant protein structure prediction. For fair comparison and feature calculation, post-processing was carried out to keep the residue numbers the same except at the mutated sites.

For proteins with no known PDBs containing the mutation sites, AlphaFold 2.0 was used for both wild-type protein and mutant protein structure prediction. Structures with an average pLDDT larger than 70 for the whole structure were kept, which was a confidence threshold for the predicted structures in AlphaFold 2.0. In addition, if a mutated site was located in a region that was poorly predicted (pLDDT < 50), the sample was also discarded. After structure generation, the mutant protein was aligned with the wild-type protein. This alignment was carried out by only taking the backbone atoms of both wild-type and mutant residues whose pLDDTs were larger than 70 into consideration.

For drug structure generation, SMILES strings for all drugs were first obtained using the PubChem PUG REST API[54] (several drugs that could not be directly identified via PubChem were manually checked and assigned). Ions and salts were then removed from the SMILES, the structures neutralized, and the resulting SMILES strings rewritten in canonical format. 3D structures were first generated using Open Babel 3.1.1[57], using the “–gen3D” flag, and non-polar hydrogens were added to the generated structures with the “–addpolarH” flag. After structure generation, molecular docking was carried out with smina[58], using default docking parameters:

- If the wild-type PDB contained a known in-pocket ligand, then “–autobox_ligand” was selected.
- If no ligand was present, the whole protein was used to generate the docking box.

After docking, the conformation with the best smina score was kept.

**Feature calculations**. Based on the structures of the wild-type and mutant protein-ligand complexes, we calculated a total of 146 biochemical features relevant to machine learning prediction of mutation-induced affinity changes. These features were first used in the work of Aldeghi et al.[19], and we follow the procedures in their original paper for their calculation:

- Eighteen ligand properties such as logP, molecular weight, number of hydrogen acceptors and donors were calculated using RDKit.
- Twelve features describing the changes for the mutated amino acids were calculated (again with RDKit), such as hydrophobicity, number of heavy atoms, and the change in side-chain volume.
- Twenty-one features describing the mutation environment were calculated with Biopython, including the distribution of ligand and protein atoms around the mutation site and the number of residues in different property groups.
- Six features related to the protein-ligand interactions were calculated using the Protein-Ligand Interaction Profiler[59]: hydrogen bonds, hydrophobic contacts, salt bridges, $\pi$-stacking, cation-$\pi$ interactions, and halogen bonds.
- The Vina binding score as well as 58 scoring function terms were calculated using Delta-Vina XGBoost[60,61].
- Thirty pharmacophore-based solvent accessible surface area features are calculated using Delta-Vina XGBoost[61].

With the exception of the ligand property features, the numerical values $x$ of all other features given in MdrDB correspond to the difference between the mutant ($x_{mt}$) and wild-type values ($x_{wt}$), i.e., $x = x_{mt} - x_{wt}$.

**Website interface**. The web interface of MdrDB was implemented using caddy2 and Bootstrap v5.0. All processed data, structure files, biochemical features and annotations were stored in Tencent Cloud Object Storage. Interactive charts were implemented using Apache ECharts. A full tutorial for MdrDB is available at https://quantum.tencent.com/mdrdb/tutorial. A user-friendly website provides access to the curated data and structure information on wild-type and mutant complexes, and provides functions for browsing, searching, displaying and downloading the data. For detailed information on the constructed website, please see Supplementary Note 2: web design and interface. Supplementary Fig. 24 shows the search page, browse page and sample display page of the MdrDB website.

## Data availability

All data are available to browse and download on the MdrDB website: https://quantum.tencent.com/mdrdb/.

## Code availability

The code is available at https://github.com/tencent-quantum-lab/MdrDB/.

Received: 12 March 2023; Accepted: 2 June 2023;
Published online: 14 June 2023

## References

1. Ode, H. et al. Computational characterization of structural role of the non-active site mutation m36i of human immunodeficiency virus type 1 protease. *J. Mol. Biol.* **370**, 598–607 (2007).
2. Alfalah, M., Keiser, M., Leeb, T., Zimmer, K.-P. & Naim, H. Y. Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency. *Gastroenterology* **136**, 883–892 (2009).
3. Koukouritaki, S. B. et al. Identification and functional analysis of common human flavin-containing monooxygenase 3 genetic variants. *J. Pharmacol. Exp. Ther.* **320**, 266–273 (2007).
4. Teng, S., Madej, T., Panchenko, A. & Alexov, E. Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys. J.* **96**, 2178–2188 (2009).
5. Yamada, Y. et al. Catalytic inactivation of human phospholipase d2 by a naturally occurring gly901asp mutation. *Arch. Med. Res.* **37**, 696–699 (2006).
6. Hashimoto, K. & Panchenko, A. R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl Acad. Sci. USA* **107**, 20352–20357 (2010).
7. Jones, R. et al. A cdkn2a mutation in familial melanoma that abrogates binding of p16ink4a to cdk4 but not cdk6. *Cancer Res.* **67**, 9134–9141 (2007).
8. Nishi, H. et al. Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS One* **8**, e66273 (2013).
9. Li, M., Petukh, M., Alexov, E. & Panchenko, A. R. Predicting the impact of missense mutations on protein–protein binding affinity. *J. Chem. Theory Comput.* **10**, 1770–1780 (2014).
10. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
11. Cohen, M. L. Epidemiology of drug resistance: implications for a post-antimicrobial era. *Science* **257**, 1050–1055 (1992).
12. Martinez, J. & Baquero, F. Mutation frequencies and antibiotic resistance. *Antimicrob. Agents Chemother.* **44**, 1771–1777 (2000).
13. Friedman, R. Drug resistance missense mutations in cancer are subject to evolutionary constraints. *PLoS One* **8**, e82059 (2013).
14. Kelso, A. & Hurt, A. C. The ongoing battle against influenza: drug-resistant influenza viruses: why fitness matters. *Nat. Med.* **18**, 1470–1471 (2012).
15. Consortium, I. C. G. et al. International network of cancer genome projects. *Nature* **464**, 993 (2010).
16. MacLean, D., Jones, J. D. & Studholme, D. J. Application of ‘next-generation’ sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* **7**, 96–97 (2009).
17. Pires, D. E., Blundell, T. L. & Ascher, D. B. Platinum: a database of experimentally measured effects of mutations on structurally defined protein–ligand complexes. *Nucleic Acids Res.* **43**, D387–D391 (2015).
18. Hauser, K. et al. Predicting resistance of clinical abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Commun. Biol.* **1**, 1–14 (2018).
19. Aldeghi, M., Gapsys, V. & de Groot, B. L. Predicting kinase inhibitor resistance: physics-based and data-driven approaches. *ACS Cent. Sci.* **5**, 1468–1474 (2019).
20. Wu, F.-X. et al. AIMMS suite: a web server dedicated for prediction of drug resistance on protein mutation. *Brief. Bioinforma.* **21**, 318–328 (2020).
21. Liu, X., Shen, T., Mooers, B. H., Hilberg, F. & Wu, J. Drug resistance profiles of mutations in the ret kinase domain. *Br. J. Pharmacol.* **175**, 3504–3515 (2018).
22. Hu, R., Xu, H., Jia, P. & Zhao, Z. KinaseMD: kinase mutations and drug response database. *Nucleic Acids Res.* **49**, D552–D561 (2021).

23. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2012).
24. DepMap, Broad. DepMap 21Q2 Public. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14541774.v2 (2021).
25. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
26. Vitkup, D., Sander, C. & Church, G. M. The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* **4**, 1–10 (2003).
27. Habig, M., Lorrain, C., Feurtey, A., Komluski, J. & Stukenbrock, E. H. Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. *Nat. Commun.* **12**, 5869 (2021).
28. Zeitz, C. et al. Chm mutation spectrum and disease: an update at the time of human therapeutic trials. *Hum. Mutat.* **42**, 323–341 (2021).
29. Rong, S. et al. Mutational bias and the protein code shape the evolution of splicing enhancers. *Nat. Commun.* **11**, 2845 (2020).
30. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5**, 302–307 (2001).
31. Hao, G.-F., Yang, G.-F. & Zhan, C.-G. Computational mutation scanning and drug resistance mechanisms of HIV-1 protease inhibitors. *J. Phys. Chem. B* **114**, 9663–9676 (2010).
32. Aldeghi, M., Gapsys, V. & de Groot, B. L. Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent. Sci.* **4**, 1708–1718 (2018).
33. Steinbrecher, T. B. et al. Accurate binding free energy predictions in fragment optimization. *J. Chem. Inf. Model.* **55**, 2411–2420 (2015).
34. Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703 (2015).
35. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
36. Barlow, K. A. et al. Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).
37. Sun, T., Chen, Y., Wen, Y., Zhu, Z. & Li, M. PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. *Commun. Biol.* **4**, 1311 (2021).
38. Yang, Z.-Y., Ye, Z.-F., Xiao, Y.-J., Hsieh, C.-Y. & Zhang, S.-Y. Spldextratrees: robust machine learning approach for predicting kinase inhibitor resistance. *Brief. Bioinform.* **23**, bbac050 (2022).
39. Breiman, L., Friedman, J., Stone, C., Olshen, R. & Stone, C. Classification and regression trees (Wadsworth, Belmont, CA, 1984). In *Proceedings of the Thirteenth International Conference, Bari, Italy*, 148 (1996).
40. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
41. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
42. Awad, M., Khanna, R. Support Vector Regression. In *Efficient Learning Machines* https://doi.org/10.1007/978-1-4302-5990-9_4 (Apress, Berkeley, CA, 2015).
43. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).
44. Zou, H. The adaptive Lasso and its Oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006).
45. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
46. Drucker, H. Improving regressors using boosting techniques. In *ICML '97 Proc. Fourteenth International Conference on Machine Learning*, (ed Kaufmann, M.) 107–115 (ICML, Lille, 1997).
47. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
48. Gardner, M. W. & Dorling, S. Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).
49. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
50. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
51. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
52. Boutet, E. et al. Uniprotkb/swiss-prot, the manually annotated section of the UniProt knowledgebase: how to use the entry view. In *Plant Bioinformatics: Methods and Protocols* (ed. Edwards, D.) 23–54 (Springer, 2016).
53. Hunter, S. et al. Interpro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
54. Kim, S. et al. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
55. Rose, Y. et al. RCSB Protein Data Bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *J. Mol. Biol.* **433**, 166704 (2021).
56. DeLano, W. L. et al. Pymol: an open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).
57. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
58. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
59. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated Protein–Ligand Interaction Profiler. *Nucleic Acids Res.* **43**, W443–W447 (2015).
60. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
61. Wang, C. & Zhang, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.* **38**, 169–177 (2017).

## Acknowledgements

We are grateful to our colleagues in the Tencent Quantum Laboratory for helpful feedback and suggestions on the construction of the MdrDB website.

## Author contributions

Z.Y.Y. and Z.F.Y. collected, preprocessed, and analyzed all the data. J.Z.Q. used Alpha-Fold 2.0 to predict the structures of proteins. Z.Y.Y. calculated a total of 146 biochemical features and evaluated the performance of the current machine learning methods. R.J.F., D.Y.L., Z.Y.Y., and Z.F.Y. designed and developed the MdrDB website. Z.Y.Y., Z.F.Y., and J.A. wrote and reviewed the manuscript. C.Y.H., and S.Y.Z. read the manuscript. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42004-023-00920-7.

**Correspondence** and requests for materials should be addressed to Shengyu Zhang.

**Peer review information** *Communications Chemistry* thanks Matteo Aldeghi and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.