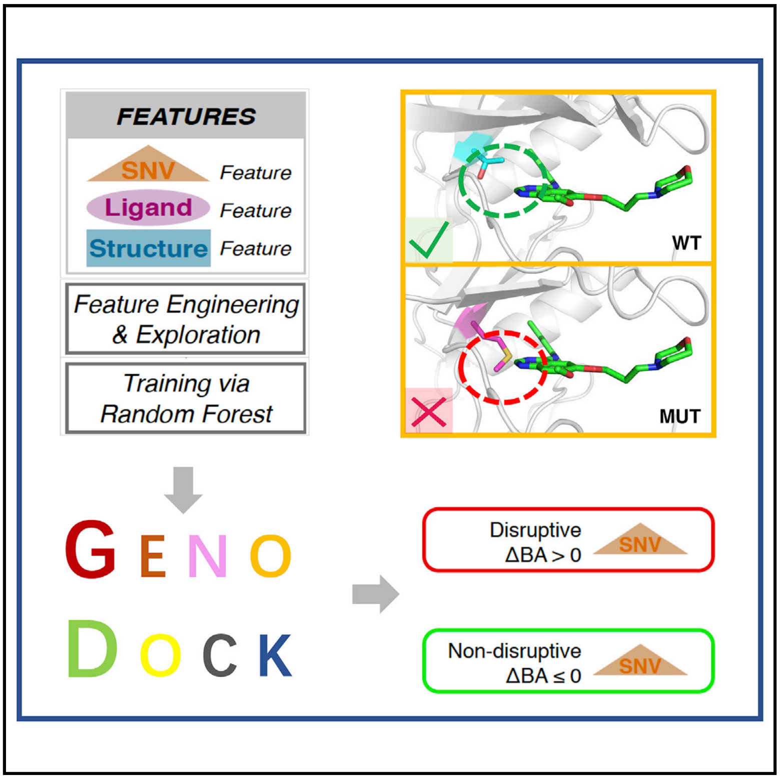
Structure

Resource

Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions

Graphical Abstract



Highlights

- Integrating genetic variations with protein-drug complex structures on a large scale
- Using physics-based calculation to bootstrap the parameterization of a predictive model
- Efficiently predicting the impacts of SNVs on protein-drug interactions

Authors

Bo Wang, Chengfei Yan,
Shaoke Lou, ..., Yucheng T. Yang,
Donghoon Lee, Mark Gerstein

Correspondence

mark@gersteinlab.org

In Brief

Genetic variation may affect drug efficacy by altering its binding affinity to the protein target. GenoDock, developed by Wang et al., is a statistical model to predict the impacts of SNVs on protein-drug interactions by combining genomic, structural and physicochemical features.

Wang et al., 2019, Structure *27*, 1469–1481
September 3, 2019 © 2019 Elsevier Ltd.
https://doi.org/10.1016/j.str.2019.06.001

Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions

Bo Wang,[1,6] Chengfei Yan,[2,3,6] Shaoke Lou,[2,3] Prashant Emani,[2,3] Bian Li,[2,3] Min Xu,[2,3] Xiangmeng Kong,[2,3] William Meyerson,[2,5] Yucheng T. Yang,[2,3] Donghoon Lee,[2] and Mark Gerstein[2,3,4,7,*]
[1]Department of Chemistry, Yale University, New Haven, CT 06520, USA
[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA
[5]Yale School of Medicine, Yale University, New Haven, CT 06520, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: mark@gersteinlab.org
https://doi.org/10.1016/j.str.2019.06.001

SUMMARY

A key issue in drug design is how population variation affects drug efficacy by altering binding affinity (BA) in different individuals, an essential consideration for government regulators. Ideally, we would like to evaluate the BA perturbations of millions of single-nucleotide variants (SNVs). However, only hundreds of protein-drug complexes with SNVs have experimentally characterized BAs, constituting too small a gold standard for straightforward statistical model training. Thus, we take a hybrid approach: using physically based calculations to bootstrap the parameterization of a full model. In particular, we do 3D structure-based docking on ∼10,000 SNVs modifying known protein-drug complexes to construct a pseudo gold standard. Then we use this augmented set of BAs to train a statistical model combining structure, ligand and sequence features and illustrate how it can be applied to millions of SNVs. Finally, we show that our model has good cross-validated performance (97% AUROC) and can also be validated by orthogonal ligand-binding data.

INTRODUCTION

In recent years, the immense growth of both genetic variation (Muir et al., 2016; Zuk et al., 2014) and protein structure datasets (Rose et al., 2015), which benefit from significant advancement in related techniques, has enabled us to study in depth the impact of genomic variants on protein structure and function (Sethi et al., 2015). Great efforts have been made to understand how genetic variants associate with various diseases at a population level, and to exploit those associations in order to potentially enhance drug effectiveness in the era of personalized medicine (Collins and Varmus, 2015; Ginsburg and McCarthy, 2001; Laing et al., 2011). Studies have shown that many drugs are effective toward only a limited fraction of individuals due to different responses from patients (Meyer et al., 2013; Spear et al., 2001; Wilkinson, 2005). One of the reasons for this loss of efficacy is drug-resistant genetic variants carried by patients (Madian et al., 2012; Wilkinson, 2005). A patient's genetic-centric prescription may be a reasonable approach to address the problem of drug ineffectiveness, especially because recent advances of sequencing technologies make it practical and affordable to carry out high-throughput personal genomic analysis.

There are many tools to evaluate the effect of sequence variants. Some widely used tools include SIFT (Kumar et al., 2009; Ng and Henikoff, 2003), PolyPhen-2 (Adzhubei et al., 2013), and combined annotation-dependent depletion (CADD) (Kircher et al., 2014). These mainly focus on sequence conservation within and across species to quantify the general impact of a non-synonymous single-nucleotide variants (SNVs). Studies for this purpose are usually limited to the available experimentally measured SNV-implication characterizations on native and mutant protein samples. Conceptually, we can map SNV data onto associated protein structures to quantitatively investigate how related physical properties are altered upon point mutation. In practice, experimentally measured data are limited to only a small pool of mutations. Specifically, while consortia projects such as The Cancer Genome Atlas (TCGA) and the Exome Aggregation Consortium (ExAC) have identified more than ten million exonic SNVs, available experimental measurements characterizing variant impacts such as protein-ligand binding affinity change are scarce (Pires et al., 2015). When we expand the scope to structural bioinformatics (i.e., just having a structure but not requiring direct experimental measurement of binding affinity), the scale expands. The number of available high-quality 3D protein structures deposited in the RSCB Protein Data Bank (PDB) is significant: there are about 41,000 protein structures available from *Homo sapiens* in the RSCB PDB (Berman et al., 2000). By one estimate, nearly 175,000 exonic variants can be mapped to at least one protein PDB file (with a resolution of at least 2.8 Å) (Kumar et al., 2016). Thus, leveraging knowledge gained from mapping variants onto the considerable databases of 3D protein structures (Glusman et al., 2017; Kumar et al., 2016) provides a promising way for understanding SNV functional impacts, such as influence on drug effectiveness. Although the quantification of functional impact depends on the quality of available computational methods, recent advances directed toward reliable physical property calculations can aid significantly in such an endeavor.

Computational simulations of proteins have been extensively validated as a method to study protein dynamics and conformations (De Vivo et al., 2016). In addition, computational simulations can be used to calculate associated physical properties, such as free energy change, especially when large-scale experimental throughput is limited by available resources. The field has seen rapid and consistent development from the pioneering molecular dynamics work by Levitt and Warshel (1975) and McCammon et al. (1977) decades ago, to more recent structure modeling and docking tools such as UCSF DOCK (Kuntz et al., 1982), Rosetta (Rohl et al., 2004), AutoDock4 (AD4) (Morris et al., 2009), and MODELLER (Webb and Sali, 2016). At every stage, there have been concerted efforts to establish the correspondence with experimental data, such as through the Critical Assessment of Structure Prediction (Moult et al., 2016). With the concurrent growth of genetic variation data at the population level, linking 3D protein structures and genomics using computational models has proven to be a powerful and innovative approach for precision medicine (Meyer et al., 2018). Such integrated approach combining different data type (i.e., structural, physical modeling, and genetic information) can enhance the accuracy and completeness of a model, especially when experimental measurement is limited (Ward et al., 2013). Here, we primarily focused on protein-drug interactions and aimed to determine the likelihood that an SNV perturbs the interaction between a protein and its associated drug ligands. Once the set of genetic variants carried by an individual are identified, the focus can then be shifted to how single-locus alteration of protein residues caused by SNVs would influence drug efficacy. Thus, robust methods to accurately predict if an SNV of interest would disrupt the binding of a drug to its protein target would help determine how an individual's carried variants would potentially affect drug efficacy.

To implement this idea, we developed a supervised learning method, GenoDock, to bridge SNVs on a large population scale and high-resolution protein-drug co-crystal structures. Our primary goal was to investigate how a given variant affects protein-drug binding affinity. We first constructed our database by mapping germline and somatic variants onto their associated protein residues, within co-crystal structures that include drug molecules. We subsequently generated a computational model of the mutated protein based on its sequence and the structure of the native protein using MODELLER (Webb and Sali, 2016). We then examined the binding affinity change (ΔBA) for the bound drug between the native and mutated protein structures associated with each SNV in our database through molecular docking. We grouped the variants based on whether they would lead to a positive shift in binding affinity ($\Delta BA > 0$) or not ($\Delta BA \leq 0$), with a positive shift in binding affinity indicating that the corresponding SNV is a disruptive one. The disruptive SNVs were our main focus in this study due to their high potential to be associated with drug resistance. Due to the fact that available experimentally measured ligand binding affinity change data are highly limited (Benore, 2010), it is not practical to train a supervised learning model based on experimental data. Therefore, in our method we circumvented this issue by constructing a calculated binding affinity change set as our pseudo gold standard using molecular docking programs. This enabled us to train a supervised learning model using a Random Forest algorithm to predict the probability of a given SNV to destabilize protein-drug binding by integrating genomic, structural, and physicochemical features from SNV annotations, protein structures, and drug ligands.

RESULTS

Building the GenoDock Dataset and with Pseudo Gold Standard Set of ΔBA

Figure 1 shows the overall framework of this study. We first build the GenoDock dataset containing SNV, structure, and ligand features for selected SNV-Structure-Ligand entries. Then we leverage physical calculations to construct the pseudo gold standard set of binding affinity changes for the variants in the GenoDock dataset. In the end, we train a full statistical learning model after a feature exploration and engineering step, to score the likelihood of an SNV of interest to disrupt protein-drug ligand binding.

Figure 1A highlights our strategy to construct the dataset (publicly available from the GenoDock website, http://genodock.molmovdb.org/). The database contains 10,283 non-synonymous SNVs from 228 proteins in *Homo sapiens*, and 113 Food and Drug Administration (FDA)-approved drug ligands, which have co-crystal structures with at least one of the 228 proteins. We screened over 30,000 human proteins with high-resolution (better than 3.0 Å) X-ray-solved PDB structures (https://www.rcsb.org/), and kept those with at least one FDA-approved drug ligand in the co-crystal structures. After removing any redundancy in the structures based on sequence alignment, we mapped germline SNVs from the ExAC (Lek et al., 2016) and somatic SNVs from TCGA (Cancer Genome Atlas Research Network, 2008, 2012; Cancer Genome Atlas Research Network et al., 2013) datasets to these 228 protein structures, utilizing the BioMart-derived correspondence between human gene, transcript, and PDB ID (Kasprzyk, 2011). In total, we identified 8,565 SNVs in 166 PDB structures for ExAC germline variants, and 1,718 SNVs in 135 PDB structures for TCGA somatic mutations (see Data S1). The SNVs, protein structures, and drug ligands formed SNV-Structure-Ligand 3-tuple entries in our database. For each SNV-Structure-Ligand entry, as visualized in Figure 1B, we used the MODELLER program suite (Webb and Sali, 2016) to generate a putative structural model of the point mutation through homology modeling. We then used AutoDock Vina (Trott and Olson, 2010) to calculate the binding affinity score for the binding of the ligand with the wild-type protein (ΔG_{WT}) and with the mutated protein (ΔG_{MUT}) in order to determine the score change (ΔBA) in kcal/mol ($\Delta BA = \Delta G_{MUT} - \Delta G_{WT}$). The set of ΔBA values served as the reference dataset, or pseudo gold standard for the GenoDock program suite.

GenoDock aims to predict the change in binding affinity of the drug ligand after the protein target is mutated, based on a Random Forest classifier. $\Delta BA > 0$ indicates that less energy is

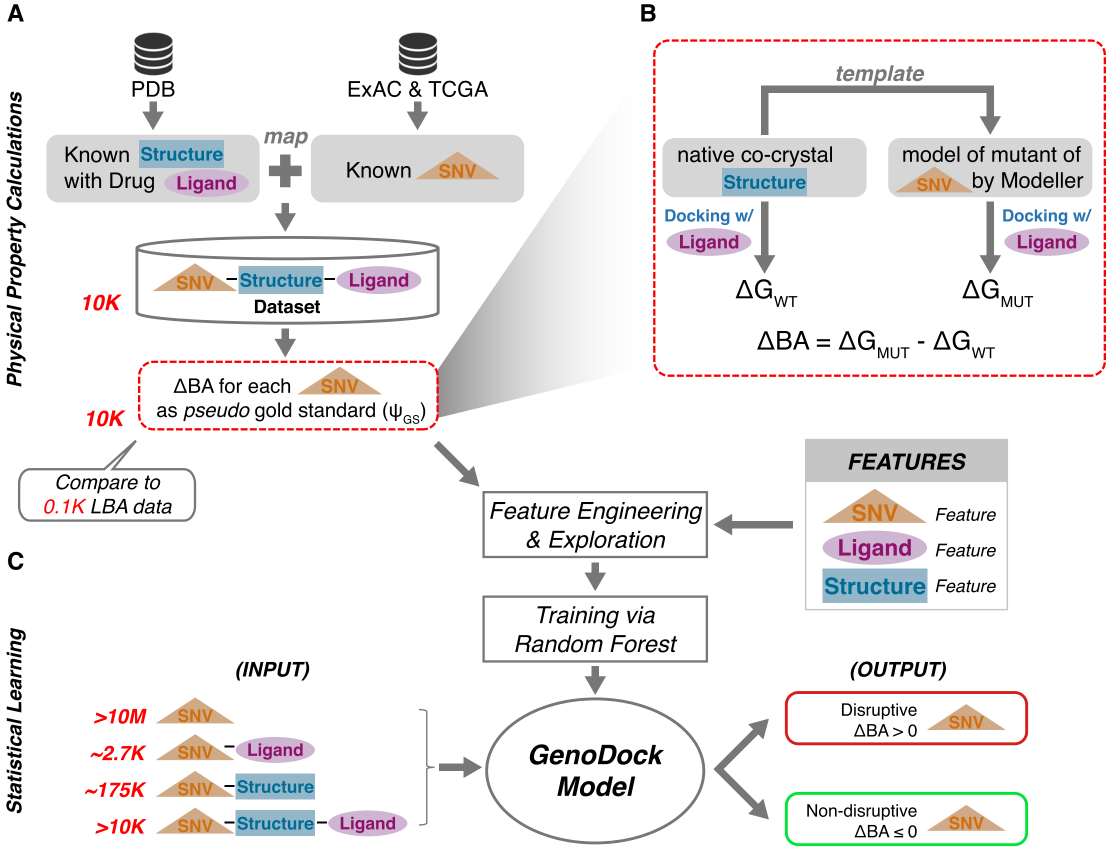


Figure 1. Framework of the GenoDock Project from Dataset Preparation to Model Construction

(A) A flowchart for collecting and processing raw data to construct the GenoDock database from the protein structure data source (RCSB PDB), SNV data sources (ExAC and TCGA), and drug ligand data source (PubChem Compound). SNVs were mapped with protein-drug co-crystal structures to form each SNV-Structure-Ligand entry in our database. We then calculated the binding affinity change for each mutation to construct the pseudo gold standard for the machine-learning model.

(B) Illustration of protein-ligand binding affinity change calculations. For each native co-crystal structure in our dataset, we generated a mutant structure using Modeller. For each native and mutated structure pair, we calculated the binding affinity using Vina, in order to obtain the binding affinity change (ΔBA) upon the point mutation. ΔBA for each SNV served as the pseudo gold standard set for the subsequent classification model.

(C) Construction of the Random Forest model to predict the direction of protein-ligand binding affinity change ($\Delta BA > 0$ or $\Delta BA \leq 0$). With feature engineering and exploration, several SNV features, drug ligand features, and structure features were combined to predict the direction of protein-ligand binding affinity change. The GenoDock program suite was trained based on Random Forest, with four application models available. We employed the Platinum dataset as our real gold standard providing about 100 LBA records for human protein mutations. By constructing docking calculations for binding affinity changes, we calculated the binding affinity change for each of the ~10,000 mutations in the GenoDock database, making it possible to train a supervised learning model with confidence. With the trained GenoDock model, it is then possible to screen large-scale datasets for drug ligands (e.g., ~2.6k drug ligands associated with ~60k SNVs), human protein structures (~30,000 from the RCSB PDB website with resolution higher than 3.0 Å), and exonic SNVs (~1,000,000 sequenced exonic SNVs; ~175,000 SNVs mapped with at least one human protein structure with 2.8 Å or higher resolution from the RCSB PDB database).

required to break the binding between the protein and the ligand, and thus that the point mutation plays a disruptive role that could potentially cause drug resistance. As shown in Figure 1C, we categorized ΔBA values for each SNV-Structure-Ligand entry into two classes: if ΔBA was positive, we tagged it as "disruptive"; if ΔBA was non-positive, we tagged it as "non-disruptive." We integrated selected genomic, structural, and physicochemical features of SNVs, protein structures, and ligands to train the classifier: SNV annotation features included allele frequency, SIFT (Kumar et al., 2009), PolyPhen-2 (Adzhubei et al., 2010), and GERP (Davydov et al., 2010) scores; ligand features included molecular weight, hydrogen bond donor and acceptor count, rotatable bond count, and polar surface area; and protein structure features included binding site, side-chain hydropathy and volume change, and distance of the mutated residue from ligand (see the STAR Methods for details of Random Forest model construction and feature selection; Figures 1, 4, S1, and S2).

An ideal approach is to use experimentally measured ligand binding affinity (LBA) data to characterize the impact of SNVs on protein-drug binding. However, the quantity of available LBA data are far from enough for the training task. For example, the Platinum database (Pires et al., 2015) is a recent effort to collect experimentally measured LBA data for over 1,000 mutations, which could potentially serve as the real gold standard of binding affinity change. However, only around 100 mutations of the Platinum dataset are associated with human proteins.

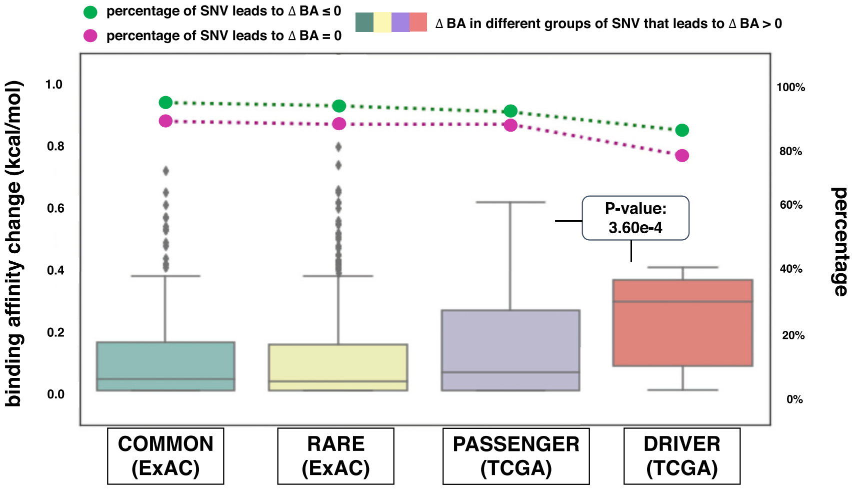


Figure 2. Boxplot of Ligand Binding Affinity Changes for Different Types of SNVs in GenoDock

An overall comparison of common, rare, passenger, and driver SNVs in terms of binding affinity change from the GenoDock data source. SNVs with $\Delta BA > 0$ were plotted in order to compare the extent of destabilization of ligand binding by each SNV group. The mean values for those SNVs leading to ligand-binding disruption for common, rare, passenger, and driver SNVs from the ExAC and TCGA dataset were 0.117, 0.129, 0.159, and 0.236 kcal/mol, respectively. The difference between common and rare SNVs from the ExAC dataset was not significant; the difference between passenger and driver SNVs from TCGA was significant, with a p value of 3.60×10^{-4}, where driver SNVs were more likely to disrupt ligand binding compared with other groups. The green-dotted and pink-dotted lines in the figure show the percentage of SNVs from each group that lead to non-positive shift of binding affinity ($\Delta BA < 0$ or $\Delta BA = 0$; 94%, 93%, 91%, and 85%, respectively), and those that do not change the binding affinity ($\Delta BA = 0$; 88%, 87%, 87%, and 77%, respectively). It is clear that cancer driver SNVs have a greater probability to result in a positive binding affinity change compared with the other three groups.

By constructing the pseudo gold standard set for each of the SNV-Structure-Ligand entries, our training dataset can be expanded to ~10k, which is a 10-fold increase relative to available entries in Platinum. We then validate that the accuracy of GenoDock predictions is acceptable using these Platinum experimentally measured records. Thus, the GenoDock model enables us to prioritize SNVs that may potentially disrupt protein-drug binding by leveraging large-scale database drug ligand, protein structure, and exonic SNV datasets. For instance, there are more than ten million exonic variants sequenced from consortium projects such as ExAC and TCGA; in one study, more than 175k exonic variants were mapped onto at least one protein structure with a resolution of at least 2.8 Å from the PDB (Kumar et al., 2016). The DrugBank database (Wishart et al., 2018) contains ~2.6k approved small-molecule drugs, and ~60k SNVs could be associated with these via ~800 drug-related genes (Schärfe et al., 2017). All of these datasets could potentially be screened with the GenoDock program suite to prioritize the disruptive SNVs.

Sequence Characterization and Distributions of Δ*BA* for SNVs

After constructing the GenoDock dataset, we first analyzed the mutation landscape of TCGA somatic and ExAC germline variants in our dataset (see Figures S3A and S3B). This allowed us to measure known amino acid changes and mutation trends that are under high selective constraint or potentially lead to disease. Next, using the ExAC germline SNVs in our dataset, we then examine whether there is a significant difference between the rare and the common SNV groups in terms of destabilization of the protein-drug complex. Rare and ultra-rare SNVs are generally interpreted as having higher impact than common ones. The allele frequency values in population-level studies also indicate varying degrees of constraint under natural selection. Similarly, we divided the TCGA somatic SNVs into highly deleterious driver SNVs and neutral passenger SNVs to investigate differences in the impact of the two groups on drug binding (Stefl et al., 2013) (see the STAR Methods for details regarding common, rare, passenger, and driver SNV tagging).

Figure 2 visualizes the distributions of binding affinity change for each group, especially the positive shift in Δ*BA* by disruptive SNVs, which contribute to 6.0% and 8.9% of all SNVs in the ExAC and TCGA datasets, respectively (Figure S3C). Although we did not observe a significant difference in Δ*BA* distributions between common and rare SNVs, when we compared the top common and rare germline SNVs with positive Δ*BA* (the "outlier" region in the boxplot), the top rare SNVs had a significantly higher Δ*BA* than the top common ones. This finding implies that the rare SNV pool contains more extremely deleterious samples in terms of disrupting drug-protein binding than those from the common SNV pool (e.g., the difference between the top 50 rare and common SNV groups has $p = 3.5 \times 10^{-7}$; Figure S3D). This observation is intuitively consistent with our expectations, as the rarity of variants stems from higher selective constraints due to greater impacts on protein stability.

Efforts made to characterize cancer genomes (Cancer Genome Atlas Research Network, 2008, 2012; Forbes et al., 2011) have validated important roles of driver SNVs in driving cancer progression (Hong et al., 2015; Raphael et al., 2014). These findings motivated us to probe the impact of SNVs in driver genes on perturbing interactions between associated protein residues and drug ligands. Indeed, our analysis revealed a significant difference between SNVs in passenger and driver genes (Figure 2): cancer-associated driver SNVs tend to destabilize protein-drug binding to a greater extent compared with neutral passenger SNVs ($p = 3.60 \times 10^{-4}$). In Figure 2, we also plotted the percentage of SNVs that led to a non-positive Δ*BA* together with the percentage of SNVs that did not change the binding affinity upon point mutation (Δ*BA* = 0). We found that the portion of SNVs that would cause a non-positive Δ*BA* decrease from common (94%), rare (93%), passenger (91%), to driver (85%) groups. This indicates that, based on our metric, the driver SNV group has a larger proportion of variants that impair drug binding compared with the other groups.

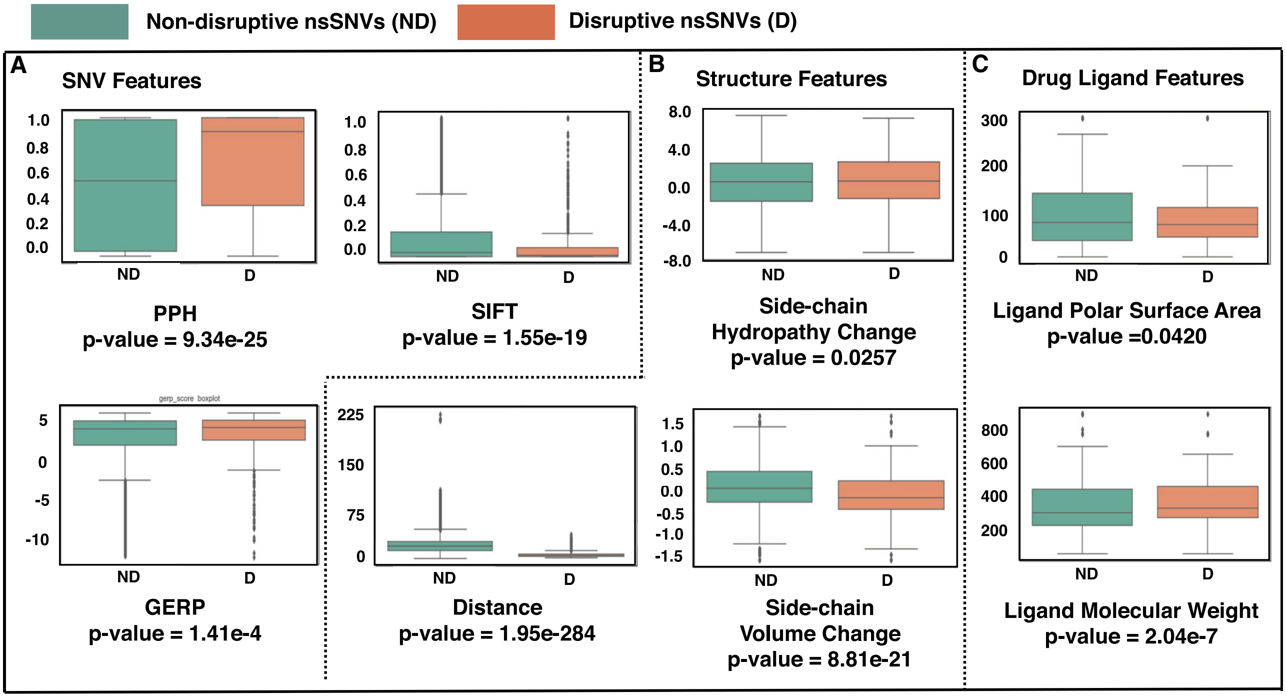


Figure 3. Boxplot Distribution between Disruptive SNVs (Positive Binding Affinity Shift) and Non-disruptive SNVs (Non-positive Binding Affinity Shift) for Different Features Groups

(A) PolyPhen-2, SIFT, and GERP scores as SNV features. We observed that PolyPhen-2, SIFT, and GERP scores for the two groups of SNVs were all significantly different with p values smaller than 0.05 from two-sample Wilcoxon tests. SNVs that disrupt ligand-protein binding had a higher mean PolyPhen-2 score (median PolyPhen-2 value: 0.916 and 0.549 for disruptive and non-disruptive SNVs, respectively) and a lower SIFT score (median SIFT value: 0.010 and 0.030 for disruptive and non-disruptive SNVs, respectively), both indicating a more deleterious role of disruptive SNVs on protein function. In terms of GERP score, SNVs leading to positive binding affinity change were more likely to be associated with protein residues from more conserved regions, indicated by a higher mean GERP score (median GERP value: 4.37 and 4.15 for disruptive and non-disruptive SNVs, respectively).

(B) Side-chain volume and hydropathy change as protein structure features; distance between ligand and mutated residue when co-crystal structure is present. Amino acid side-chain volume and hydropathy change before and after mutation directly affect interaction of the protein residue with the ligand. We observed that the mean value of both side-chain volume and hydropathy were statistically significant. On average, SNVs that destabilized ligand binding had decreased side-chain volumes compared with the other class of non-disruptive SNVs (median volume change index: −0.240 and 0 for disruptive and non-disruptive SNVs, respectively). For side-chain hydropathy change, there was also a significant difference between the two classes of SNVs (median hydropathy change: 0.40 and 0.30 for disruptive and non-disruptive SNVs, respectively). When protein-drug co-crystal structures were present, we directly calculated the distance of the mutated protein residue from the drug ligand. As expected, the SNVs that positively shifted binding affinity were more likely to be mapped onto residues within the binding pocket (median distance from ligand: 5.42 and 17.4 Å for disruptive and non-disruptive SNVs, respectively).

(C) Polar surface area and molecular weight as ligand features. Within the context of protein-drug ligand interaction, physicochemical features of drug molecules play vital roles in interpreting SNV implications. We observed that for SNVs that disrupted binding affinity, the associated drug ligands tended to have a significant smaller average polar surface area that those corresponding to SNVs in the other class (median ligand polar surface area: 77.2 and 77.8 Å^2 for disruptive and non-disruptive SNVs, respectively). We also observed that the average molecular weight of drug ligands interacting with disruptive SNVs was significantly higher than those corresponding to the non-disruptive SNVs (median molecular weight of ligand: 357.8 and 300.4 g/mol for disruptive and non-disruptive SNVs, respectively).

Feature Exploration and Engineering to Classify Disruptive and Non-disruptive SNVs

Genomic, structural, and physicochemical properties (features) of variants, proteins, and ligands play an important role in discerning between disruptive and non-disruptive classes of variants. Thus we extracted and defined a list of features that discriminate the disruptive SNVs from non-disruptive ones, and that thus serve as training parameters in our classifier (see the STAR Methods for details on feature selection and construction). For each SNV-Structure-Ligand entry in the GenoDock database, we constructed three groups of features: SNV annotation features (Figure 3A), protein structure features (Figure 3B), and drug ligand features (Figure 3C). We observed that disruptive compared with non-disruptive SNVs have a significantly lower median SIFT score (mean = 0.010 and mean = 0.030, respectively; p = 1.55 × 10^{-19}) and a significantly higher median PolyPhen-2 score (median = 0.916 and median = 0.549, respectively; p = 9.34 × 10^{-25}) (Figure 3A). This result indicates that the more deleterious SNVs, indicated by a lower SIFT or a higher PolyPhen-2 score (Adzhubei et al., 2010, 2013; Gonzalez-Perez and Lopez-Bigas, 2011; Kumar et al., 2009; Tennessen et al., 2012), are more likely to cause a positive shift in Δ*BA*. The median GERP scores for the two classes also differed significantly (p = 0.000141). SNVs that cause positive Δ*BA* are likely to be mapped onto more conserved regions, indicated by a higher GERP score (Genomes Project et al., 2012; Khurana et al., 2013; Tennessen et al., 2012), on the protein structure than the other group (median = 4.37 versus median = 4.15).

In Figure 3B, we show the boxplot distributions of the two classes of SNVs with respect to protein structure features. The distance between the mutated amino acid residue and the drug molecule is perhaps the most important feature to determine whether a point mutation would be likely to affect ligand binding. We observed that a greater proportion of SNVs that impair binding activity are in the binding pocket (median distance = 5.42 Å) than in the non-disruptive class (median distance = 17.4 Å, $p = 1.95 \times 10^{-284}$). If the distance is greater than our threshold for being within the binding pocket (8 Å), the mutation is less likely to affect the protein and drug ligand binding due to a weaker putative interaction. Another important physical property affecting drug binding is the side-chain volume change between the wild-type and mutated residue. Given our definition of the volume change index between native and mutated amino acid residues, we observed that SNVs that disrupt ligand binding are more likely to result in a decreased side-chain volume (median = −2.40, see the STAR Methods for the definition of the volume change index), whereas on average the SNVs that lead to a non-positive Δ*BA* minimally change the side-chain volume (median = 0; $p = 8.81 \times 10^{-21}$). Side-chain hydropathy change is another feature to consider in ligand-protein interactions. For example, the change in the side-chain hydropathy score (Kyte and Doolittle, 1982), which increases from a hydrophilic residue to a hydrophobic one, may result from breaking a hydrogen bond network or salt bridge between the wild-type residue and drug ligand (see Discussion for detailed case analysis) (Boccuto et al., 2014; Doss and Nagasundaram, 2012; Kumar et al., 2013; Zhang et al., 2013). We observed this trend from the SNVs in our database; the SNVs with a positive Δ*BA* had a higher hydropathy change (median = 0.40) than the non-positive Δ*BA* class (median = 0.30), indicating that disruptive SNVs tend to lead to a more hydrophobic character for the mutant residue ($p = 0.0257$).

To study an SNV's impact on protein-ligand binding, ligand properties are also an important aspect to consider (Figure 3C). We extracted five features among various physicochemical properties for each drug molecule in our database (Figures 3A and S3E). We observed that SNVs with a positive Δ*BA* residue were significantly more likely ($p = 2.04 \times 10^{-3}$) to be in a protein structure with a heavier drug ligand (median = 357 g/mol) than the non-positive Δ*BA* group (median = 300 g/mol). In addition, we found that the polar surface area of the drug ligands with an SNV that leads to positive Δ*BA* tends to be smaller (median = 77.2 Å^2) compared with the non-positive group (median = 77.8 Å^2; $p = 0.042$). One reason for this may be that the sensitivity of a heavier ligand and of a ligand with smaller polar surface area is higher in response to the side-chain volume or hydropathy change upon point mutation.

After a feature exploration and engineering process based on the differential effects of each feature on disruptive and non-disruptive SNVs, we selected the training feature candidates shown in Figure 3 for our machine-learning method. Based on this, the GenoDock classification model includes the following input features: SNV annotation scores including PolyPhen-2, SIFT, and GERP, and ligand molecule properties such as polar surface area, protein structural alterations, and side-chain volume change.

Construction and Evaluation of the GenoDock Toolkit in Classifying Binding Affinity Change

In this study, we present our GenoDock classifier to predict binding affinity score change upon point mutations using docking calculations as the gold standard for Δ*BA*. We designed this tool to identify potential SNVs that cause ligand-binding disruption and drug resistance. We implemented a machine-learning approach to achieve this purpose, with additional steps integrated into our pipeline to evaluate our predictions. To ensure that the GenoDock classifier is unbiased, we designed a method that involves a cross-validation step to pick the best-performing model among a set of chosen learning methods; a grid-search-based model selection step to optimize the parameters for learning model construction; and an evaluation step using an independent test set isolated from the training set (Figure S4A; see the STAR Methods for details). We also recognize that, when applying GenoDock to a new test case, it is possible that some of the features identified above will not be available to a user. For example, a user may only have an SNV and a drug ligand to investigate whether the SNV would be disruptive in ligand binding, with no protein structure available. Thus, we provide four independent models depending on information availability: SNV only; SNV + Structure; SNV + Ligand; SNV + Structure + Ligand. We apply the above procedure to each model separately to allow our pipeline to cover all such instances that may arise. In model selection for different learning methods, we found that the Random Forest classifier performed the best (Figure S4B; see the STAR Methods for model selection).

During our preparation of the training data, we tuned the number of samples of disruptive SNVs and non-disruptive SNVs to be 1:1 in our training set to avoid potential bias from an imbalance in samples from the two classes, while keeping the original sample ratio of two classes unchanged in the test set. For the models in which either the PDB structure or ligand molecule is present, we evaluated the classification performance with the "Binding Site" feature included and excluded during the training process, separately. As depicted in Figure 4A, we tested the classifier trained with the SNV's Binding Site feature (Binding Site is "known") to get the probability of SNVs to disrupt binding. The area under the receiver operator curve (AUROC) for predictions of four models were 0.73 (SNV only), 0.91 (SNV + Structure), 0.96 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. If it was unknown whether target SNVs were in the binding pocket or not, we trained our classifier with the Binding Site feature excluded (Binding Site is unknown) during the training and test process for the "SNV + Structure" and "SNV + Ligand" models. As shown in Figure 4B, the AUROC values for these two models were 0.74 and 0.79, respectively (see Table S1 for AUROC list of different GenoDock models). As a general rule, when we fed the GenoDock classifier with more features, the performance of predictions improved: when the input integrates all of the three feature groups, our method is able to identify most of the SNVs that lead to a positive shift toward binding affinity with an AUROC of 0.97. If we only consider the top 100 predictions ranked according to the predicted "disruptive" probability, the precision reaches 0.92. Using the same learning pipeline, we back-tested the performance of GenoDock with the performance of SIFT, PolyPhen-2, GERP, and CADD (Kircher et al., 2014), independently. GenoDock gave the highest AUROC value

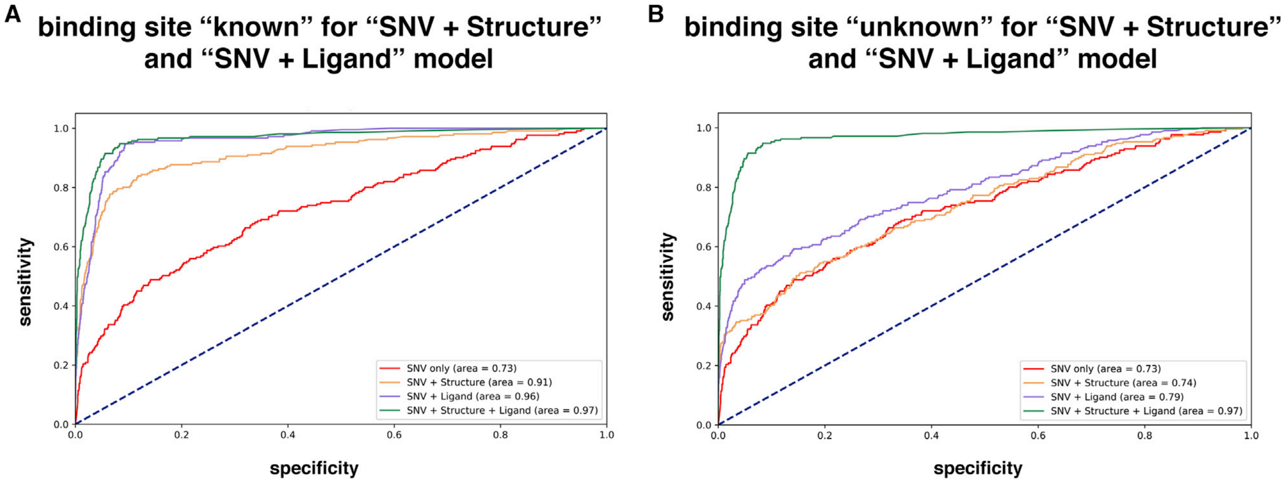


Figure 4. Performance and Implementation of GenoDock for Binding Affinity Change Prediction

(A) ROC plots for four models with different input feature groups (with "Binding Site" feature included during training process in "SNV + Structure" and "SNV + Ligand" models). Our classifier achieved an AUC of 0.73 (SNV only), 0.91 (SNV + Structure), 0.96 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. For SNV + Structure and SNV + Ligand models, we trained the model including binding site information, and we tested the data with original binding site information of each single SNV.

(B) ROC plots for four GenoDock models with different input feature groups (with Binding Site feature excluded during training process in SNV + Structure and SNV + Ligand models). Our classifier achieved AUC of 0.73 (SNV only), 0.74 (SNV + Structure), 0.79 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. For "SNV + PDB" and SNV + Ligand models, we train and test the model without the Binding Site feature to predict the influence of SNVs onto binding affinity change in case we cannot tell whether the associated protein residue is on the binding site or not. In the GenoDock web interface, users can switch Binding Site to be known or unknown for predictions of interest.

among these tools, as it was specifically developed to address the impact of SNVs on ligand-binding affinity change instead of a general annotation of potential benign or deleterious influences on protein function (Table S1 and Figure S4C).

We then applied Gini importance scoring to identify the relative importance of different features during the decision-making process (Menze et al., 2009). We observed that the relative importance of features such as the SNV annotations and binding site remained stable across our different models, revealing the robustness of our method. Identifying the relative importance of a range of genomic and structural features under a uniform learning pipeline provides us with a reasonable way to draw insights into how an SNV would impact ligand binding (Figure S4D).

Independent Validation of GenoDock Predictions Using Experimentally Measured Data

To further evaluate the performance of GenoDock, we applied the program suite on an independent test set parsed from the Platinum database, containing experimentally measured binding affinity changes upon point mutations (see the STAR Methods for details on dataset preparation; see Data S2 for parsed Platinum dataset). For the 86 data entries drawn from Platinum, the area under the curve (AUC) of the ROC reached 0.62, which shows reasonable accuracy of GenoDock benchmarked with experimentally measured results. We then evaluated the precision of GenoDock predictions on the Platinum dataset by tuning the cutoff between disruptive and "non-disruptive" based on a predicted probability of $\Delta BA > 0$. For example, when we set a cutoff of 0.5, SNVs with a probability of positive ΔBA greater than 0.5 were assigned disruptive, and all other SNVs were assigned non-disruptive. We counted the number of true-positive and false-positive entries benchmarked with the gold standard set and calculated the precision. With a cutoff of 0.5, the precision reached 0.84 (Figure S5A).

Based on our performance evaluation results, we have shown that, by integrating features from SNV annotations, protein structures, and drug ligand properties, GenoDock can clearly identify SNVs that lead to a positive ΔBA shift with high accuracy. The performance on the independent test set drawn from experimentally measured LBA data further validates the prediction reliability of GenoDock.

GenoDock Helps Identify Known and Unknown SNVs that Disrupt Protein-Ligand Binding

We present two examples of the implicit decision-making process of GenoDock in Figure 5, based on the overall importance score rankings of different features for the "SNV + Structure + Ligand" model. The T790M mutation example in Figure 5A shows how GenoDock correctly predicts a known drug-binding-disruption variant; in contrast, the R112H mutation example in Figure 5B shows that GenoDock could help to identify a drug-binding-disruptive variant that has not yet been characterized by LBA or clinical studies.

As shown in Figure 5A, GenoDock successfully predicted that the somatic T790M mutation (rs55181378) on human epidermal growth factor receptor (EGFR) (PDB: 2ITY) is very likely to impair the binding between one of its tyrosine kinase inhibitors (TKIs), gefitinib, and the EGFR kinase domain (probability of $\Delta BA > 0$ was 64%). Through molecular and clinical studies, researchers have shown that the resistance toward gefitinib arises from the substitution of a bulkier methionine residue for threonine at

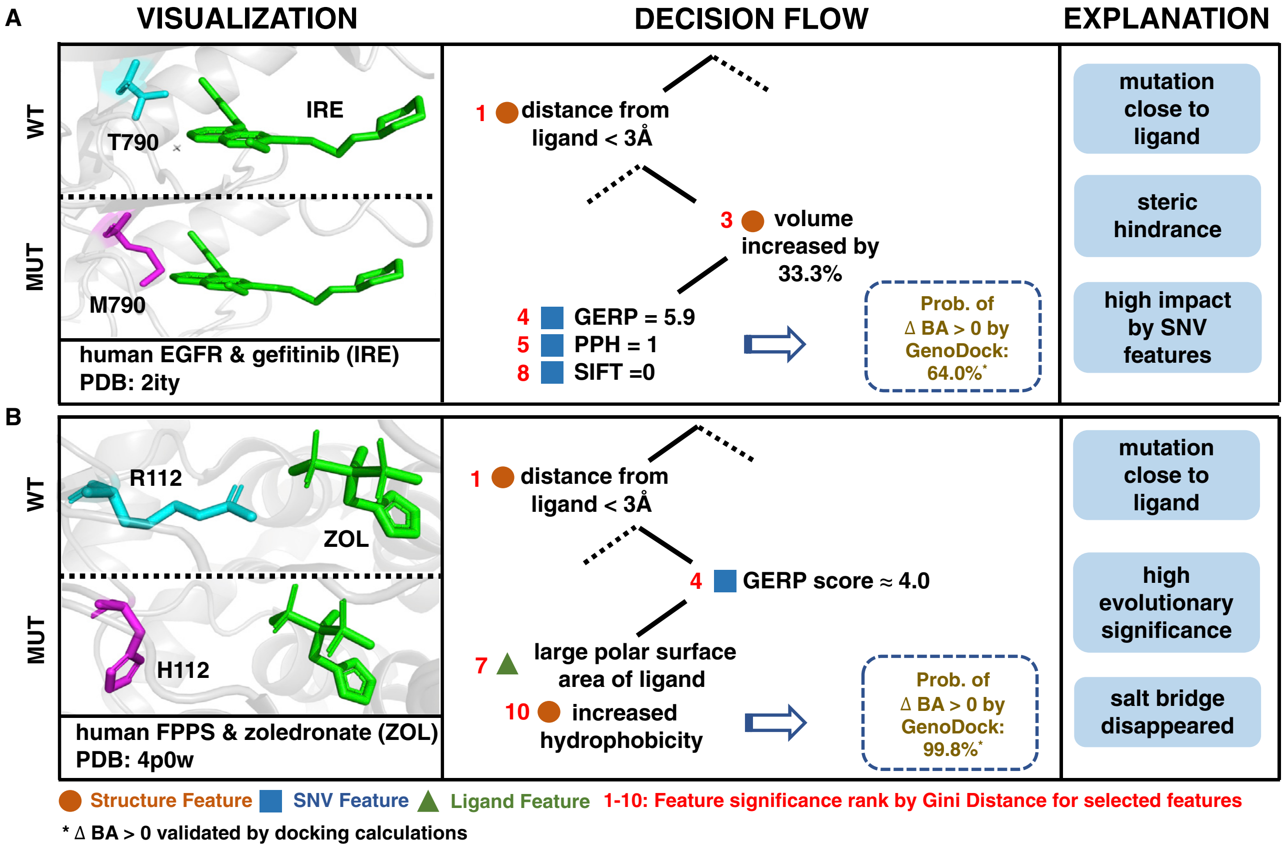

A
VISUALIZATION
DECISION FLOW
EXPLANATION
WT
MUT
T790
IRE
M790
human EGFR & gefitinib (IRE)
PDB: 2ity
1 ● distance from ligand < 3Å
3 ● volume increased by 33.3%
4 ■ GERP = 5.9
5 ■ PPH = 1
8 ■ SIFT =0
Prob. of Δ BA > 0 by GenoDock: 64.0%*
mutation close to ligand
steric hindrance
high impact by SNV features
B
R112
ZOL
H112
human FPPS & zoledronate (ZOL)
PDB: 4p0w
1 ● distance from ligand < 3Å
4 ■ GERP score ≈ 4.0
7 ▲ large polar surface area of ligand
10 ● increased hydrophobicity
Prob. of Δ BA > 0 by GenoDock: 99.8%*
mutation close to ligand
high evolutionary significance
salt bridge disappeared
● Structure Feature ■ SNV Feature ▲ Ligand Feature 1-10: Feature significance rank by Gini Distance for selected features
* Δ BA > 0 validated by docking calculations

Figure 5. Case Study: GenoDock Identifies Known and Unknown Drug-Resistance Mutations

(A) Identification of the T790M mutation on EGFR with the gefitinib-resistant effect. The threonine on chain A in human EGFR protein (PDB: 2ITY) is mutated to methionine by a somatic SNV (rs55181378). T790M is a well-studied mutation in clinical research. Patients with somatic-activating mutations in the EGFR gene would develop resistance to TKIs such as gefitinib (Ligand ID: IRE). With the T790M mutation, drug resistance arises from the steric hindrance of gefitinib binding due to the increased side-chain volume of methionine, leading to a positive shift to binding affinity. GenoDock correctly predicted this shift step-by-step during its decision-making process.

(B) Identification of an unknown mutation potentially leading to drug resistance: resistance effect toward zoledronate acid by R112H mutation on human ASH1L. The arginine on chain A in ASH1L protein (PDB: 4P0W) is mutated to histidine by a somatic SNV (rs155317993). Due to the breaking of the salt bridge between the ARG side chain and the drug ligand zoledronic acid (Ligand ID: ZOL), the resulting uncharged HIS binds to the ligand much more weakly, indicated by a positive shift of binding affinity change, which was correctly predicted by GenoDock.

position 790 (Balak et al., 2006; Janne, 2008; Kobayashi et al., 2005; Kosaka et al., 2006; Pao et al., 2005). Further studies on the EGFR-gefitinib co-crystal structure showed that the larger methionine residue leads to steric hindrance of the aromatic moieties of the gefitinib molecule, preventing the accessibility of gefitinib to the binding pocket of the EGFR kinase domain (Balak et al., 2006; Daub et al., 2004; Janne, 2008; Kobayashi et al., 2005). This biophysical rationale is traced in the classification process of GenoDock. From the decision flow in Figure 5A, the mutated residue is mapped in the binding pocket of the kinase domain, and the side-chain volume is increased by one-third from threonine to methionine, which may potentially block the interaction of the ligand to the binding pocket. Furthermore, the functional annotations of the SNV associated with the T790M mutation indicate that this variant is of high impact, which strengthens our confidence that this variant would impair protein-ligand binding. Together with the fact that the side-chain hydropathy changes from a hydrophilic threonine to a hydrophobic methionine, GenoDock classifies this SNV as very likely to cause a positive shift in binding affinity.

In Figure 5B, we present an example representing the method by which GenoDock helps identify new variant candidates that could potentially lead to drug resistance, using the SNV + Structure + Ligand model. Farnesyl diphosphate synthase (FPPS) is an important target for the bisphosphonate class of drugs such as zoledronate (ZOL). ZOL targets FPPS as an immunomodulator, which alters macrophages from a tumor-promoting to a tumor-killing phenotype (Coscia et al., 2010; Kunzmann et al., 1999; Martin et al., 2001; Russell, 2011; Shipman et al., 1998; Wood et al., 2002). ZOL is a highly hydrophilic binder to FPPS via electrostatic and hydrogen bond interactions (Liu et al., 2014). We visualized the interaction between ZOL and FPPS (PDB: 4P0W) in Figure 5B, in which the ZOL ligand binds to ARG112A via a salt bridge between the positively charged guanidium with the negatively charged sulfate group of ZOL. However, with the mutation R112H (rs155317993), this binding

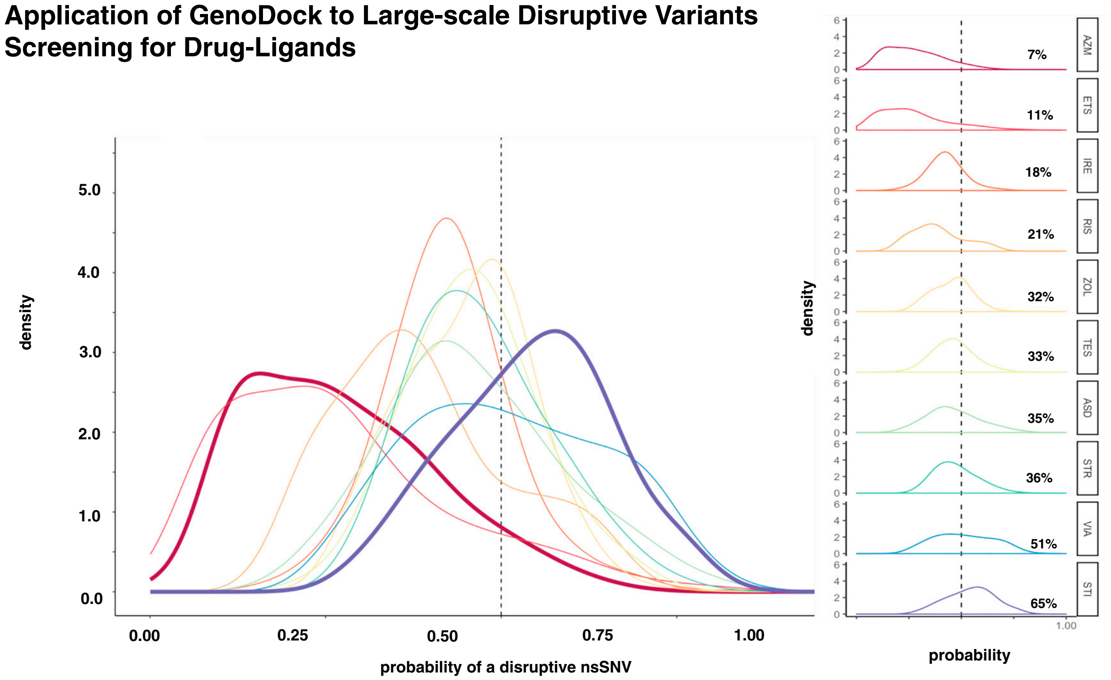


Figure 6. An Example of GenoDock Application on Large-Scale Dataset

We applied GenoDock (''SNV + Ligand'' model) on a pool of 10 drug ligands on a set of 290,515 somatic exonic variants from the original TCGA dataset to estimate how vulnerable each drug ligand is to be disrupted by individual carried variants. We assumed that each variant is associated with amino acids locating in the binding pocket in order to estimate the maximum probability of this SNV to disrupt protein-ligand binding. Each curve depicts the density distribution of the probabilities of each of ~0.3 million variants of being disruptive to the binding of a given drug ligand. Variants with a probability higher than 0.5 are highly likely to impair the binding. The fewer disruptive SNVs a drug is associated with the more likely it is for the drug to retain its efficacy for individuals carrying a variety of variants. Within the ten drug ligands selected, imatinib had the highest proportion of SNVs (65%) that are likely to disrupt its binding with proteins; relatively, acetazolamide had the lowest proportion of disruptive SNVs (7%) compared with other drug ligands.

network no longer exists. GenoDock classifies this SNV as a disruptive one with a probability of 99.8%, following a similar decision-making pipeline as discussed in the previous example. Although, the disruptive role of R112H in ZOL binding to FPPS has not yet been characterized by experimental assays, GenoDock provides evidence that this variant is highly likely to impair the inhibitor effectiveness, warranting further investigation. We validated the predictions of both examples using AD4, which resulted in $\Delta BA > 0$. More biological functional assays can be performed in the future in addition to and motivated by computational prediction. Overall, and as explained below, we have used GenoDock to process large numbers of SNV candidates for disruptive variant screening.

Application of GenoDock to Large-Scale Disruptive Variant Screening for Drug Ligands

We designed the GenoDock program suite for a range of applications. Previous studies revealed that variants have direct impact upon protein structures, which could have significant consequences for drug binding (Collins and Varmus, 2015; Ginsburg and McCarthy, 2001; Laing et al., 2011). However, no computational tool yet provides a large-scale analysis for the implications of variants on drug efficacy. Here, we applied the program to evaluate how likely a drug ligand can be disrupted by the significant numbers of somatic SNVs carried by individuals. We retrieved 290,515 somatic exonic variants from the TCGA dataset, with SNV annotation features (SIFT, PolyPhen-2, and GERP) assigned to each SNV. We then selected the two drug ligands from our previous case study, gefitinib(IRE) and ZOL, together with eight other randomly picked drug ligands from the GenoDock database: risedronate (RIS), sildenafil (VIA), acetazolamide (AZM), imatinib (STI), progesterone (STR), testosterone (TES), androstenedione (ASD), and dorzolamide (ETS). We ran the GenoDock (''SNV + Ligand'' model) for each of the ten drug ligands with every SNV in the pool to calculate the cumulative probability of a certain drug being associated with SNV-disrupted protein binding. We assumed that each variant is associated with amino acid residues located within the binding pocket for that drug in order to evaluate the maximum probability of this variant to be disruptive. For each drug ligand, we plotted the density distribution curve. Each curve represents ~0.3 million calculated probabilities of variant disruption (corresponding to the number of tested SNVs) (Figure 6). Based on the optimized cutoff derived from the Platinum dataset test results, an SNV with a probability higher than 0.5 is likely to impair the binding of an associated drug ligand. This gives us a rough sense of how easily a drug ligand is affected when interacting with protein residues associated with various somatic variants. The higher the proportion of SNVs with a disruption probability higher than 0.5 for a drug ligand, the more likely that drug is to be affected in its efficacy. Of the ten drug ligands, imatinib has the highest number of SNVs that could potentially disrupt its binding with protein residues, whereas only 7% of the somatic SNVs could potentially impair binding activity for acetazolamide. This approach provides a reasonable method to evaluate drug ligand candidates with similar functionality, and we suggest to conduct further computational and experimental efforts to add more confidence and accuracy to the predicted results. A drug ligand with less SNVs that could disrupt its binding may be a relatively better choice compared with other candidates with a higher proportion of disruptive SNVs within a given variant pool.

DISCUSSION

In this study, we constructed a dataset to map SNVs with their annotations from different datasets onto high-resolution protein structures for downstream analysis. Based on this dataset, we parameterized a highly sensitive classification model to prioritize SNV candidates that could potentially cause protein-drug binding disruption based on the integration of genomic annotations and structural properties. We also built a user-friendly web interface, the GenoDock server, which rapidly provides predictions of binding affinity change for SNVs of interest. The GenoDock method is a "hybrid model" that leverages physical calculations as a pseudo gold standard to train a statistical learning model when the availability of experimentally measured gold standards is highly limited.

For the construction of the GenoDock dataset, we employed the ExAC Consortium and the TCGA project databases as the sources of germline and somatic variants, respectively. From a pool of ~2.5 million ExAC germline variants and ~1 million Pan-Cancer somatic mutations, we successfully mapped ~10,000 SNVs onto ~300 human proteins, each of which has a high-resolution co-crystal structure with an FDA-approved drug ligand available. Considering that the original dataset is not large enough for effective training, we separated it into a training set and a test set and applied a "training + cross-validation" strategy for model assessment and selection (Figure S4A), instead of splitting it into a training set, validation set, and test set. We identified 735 SNVs with a predicted positive shift in binding affinity from 123 proteins of that ~300 protein pool, covering 85 drug ligands (see Data S3). For the prioritization of SNVs that would cause binding disruption, we demonstrated that GenoDock is an efficient classifier with an AUC of 0.97 when all features are available. The independent test on the experimentally measured binding affinity change data from Platinum further shows reasonable prediction sensitivity and precision with an AUC of 0.62 and a precision of 0.84 (Figure S5A).

The major challenge of this study was to construct a gold standard set for binding affinity change between mutated and wild-type protein-ligand co-crystal structures, given the lack of associated experimentally measured ligand-binding assay results. For example, Platinum contains about 1,000 mutations, and less than 10% of the mutations (86) are associated with SNVs mapped onto human proteins. These 86 experimental results of binding affinity changes could have been served as the real gold standard set in our study. However, this dataset is far from enough to construct a supervised learning model. To fill this gap, we constructed a pseudo gold standard set of binding affinity change for each of the ~10,000 SNVs in our GenoDock dataset via docking calculations (we built a ligand-binding model using Modeller for each SNV). The prediction results of GenoDock based on the pseudo gold standard is reasonable when benchmarked with the independent test set of 86 experimentally measured entries. This test set (86 entries) is independent from our training set (~10k entries), serving as the direct validation of GenoDock statistical learning model. In terms of validating the rigorousness of the pseudo gold standard set, it is hard to directly benchmark with experiment results due to a severe lack of laboratory measurements. Instead, we achieve this by accessing the rigorousness of the ligand-binding model by testing the accuracy of docking calculations on an independent experiment dataset that is available, and the reasonable rigorousness of the model indicates a good reliability of our pseudo gold standard set used for our core statistical model (see Table S2 for clarification of datasets and models used for validation purposes). By conducting computational docking calculations to construct a relatively large pseudo gold standard dataset to train our statistical learning model, we are then able to process large numbers of variants and structure datasets.

Although our approach can identify SNV candidates that potentially impair protein-drug binding in a rapid yet accurate manner, the method is still limited in two aspects. First, the lack of high-resolution co-crystal structures of protein-drug complexes limits the size of our pseudo gold standard set. As structural data are sparse, only 1% of exome SNVs are mapped onto protein-drug co-crystal structures. Fortunately, with the development of protein structure determination techniques such as nuclear magnetic resonance, electron microscopy, and cryoelectron microscopy (Bai et al., 2015), we foresee that the amount of highly reliable protein-drug structural data will increase rapidly. In addition, remarkable progress in putative 3D protein-drug interaction models based on homology modeling techniques may also potentially expand the structure pool (Marks et al., 2011; Zhan and Guo, 2015). Combined with tremendous progress in revealing the mutational landscape of human genomes via large-scale sequencing projects such as the UK 10000 Project and the International Cancer Genomics Consortium, we will periodically update the GenoDock dataset with new SNV-Structure-Ligand entries for better prediction results.

Second, our binding affinity change data (pseudo gold standard) is calculated based on docking calculations, which limits the upper boundary of our prediction accuracy. Calculation or prediction of binding affinity change between protein and ligand molecules is still a challenging task (Ballester et al., 2014; Smith et al., 2016; Yan et al., 2016). To obtain a relatively reliable binding affinity change calculation set, we constructed our pseudo gold standard from Vina and then validated calculated results by two additional docking programs with different score functions, AD4 and MDock. Vina is state-of-the-art and one of the most reliable docking programs widely used in pharmaceutical research projects (Castro-Alvarez et al., 2017; Wang et al., 2016). We checked the consistency of the ΔBA results for each SNV-Structure-Ligand entry with AD4 and MDock, with results confirming that Vina is consistent with the other two methods. To further access the scientific rigorousness of our docking calculations, we also conducted the Vina calculation on binding affinity change on a different Platinum experimentally measured dataset, and the docking calculations and experimental measurements reached a good consistency (see the STAR Methods for details of consistency evaluation; see Table S2 for clarification of datasets for validation of GenoDock model and of docking calculations). Thus, we have endeavored to make our physical calculation set a reliable one. Admittedly, at the current stage, the pseudo gold standard set based on docking calculations is still far from perfect, and there is huge potential to improve the accuracy of our physical calculations with more powerful docking methods available in the future. If we have enough experimentally measured LBA data for mutations recoded in GenoDock database in the future, we plan to update the Δ*BA* values with experimental results under the same pipeline to further enhance the reliability of GenoDock predictions. In addition, we fixed the protein backbone while conducting docking calculations to avoid concerns and problems raised from protein flexibilities, which makes it hard to probe influence toward binding activities by protein motions or conformational changes. We could address this in the future by more complex computational optimizations involving protein and ligand motional degrees of freedom.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHODS DETAILS
    - ○ GenoDock Database Preparation
    - ○ Mutant Structure and Binding Affinity Change Calculation
    - ○ Features Extraction and Construction for the Machine-Learning Method
    - ○ Protein-Ligand Complex Visualization
- QUANTIFICATION AND STATISTICAL ANALYSIS
    - ○ Training, Testing, and Evaluating the Performance of a Machine-Learning Method
    - ○ Curation of the Independent Test Set Based on Experimental Measurements
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.str.2019.06.001.

ACKNOWLEDGMENTS

This work is supported by NSF (DBI 1660648). We thank Sushant Kumar for sharing the original source code for mapping variants onto protein structures, as well as Jing Zhang and Declan Clarke for helpful discussions and feedback.

AUTHOR CONTRIBUTIONS

B.W., C.Y., and M.G. conceived and designed the study. B.W. carried out the study, developed scripts, constructed the web interface, produced the figures, and wrote the paper. C.Y. and M.X. ran docking studies. S.L., P.E., X.K., W.M., Y.T.Y., and D.L. prepared and processed datasets. B.L. prepared the independent validation set. All authors edited the manuscript. M.G. and C.Y. oversaw the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 30, 2018
Revised: February 14, 2019
Accepted: June 3, 2019
Published: July 3, 2019

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Hamosh et al., 2005; Landrum et al., 2014; Peterson et al., 2013; Reichold et al., 2010; Stenson et al., 2014; Szpiech et al., 2017; Zhang et al., 2012.

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet., Chapter 7, Unit7 20.

Bai, X.C., McMullan, G., and Scheres, S.H. (2015). How cryo-EM is revolutionizing structural biology. Trends Biochem. Sci. *40*, 49–57.

Balak, M.N., Gong, Y., Riely, G.J., Somwar, R., Li, A.R., Zakowski, M.F., Chiang, A., Yang, G., Ouerfelli, O., Kris, M.G., et al. (2006). Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. Clin. Cancer Res. *12*, 6494–6501.

Ballester, P.J., Schreyer, A., and Blundell, T.L. (2014). Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? J. Chem. Inf. Model. *54*, 944–955.

Benore, M. (2010). Response to review of fundamental laboratory approaches for biochemistry and biotechnology. Biochem. Mol. Biol. Educ. *38*, 64.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Boccuto, L., Aoki, K., Flanagan-Steet, H., Chen, C.F., Fan, X., Bartel, F., Petukh, M., Pittman, A., Saul, R., Chaubey, A., et al. (2014). A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. Hum. Mol. Genet. *23*, 418–433.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525.

Castro-Alvarez, A., Costa, A.M., and Vilarrasa, J. (2017). The performance of several docking programs at reproducing protein-macrolide-like crystal structures. Molecules *22*, 136–149.

Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N. Engl. J. Med. *372*, 793–795.

Coscia, M., Quaglino, E., Iezzi, M., Curcio, C., Pantaleoni, F., Riganti, C., Holen, I., Monkkonen, H., Boccadoro, M., Forni, G., et al. (2010). Zoledronic acid repolarizes tumour-associated macrophages and inhibits mammary carcinogenesis by targeting the mevalonate pathway. J. Cell. Mol. Med. *14*, 2803–2815.

Darby, N.J., and Creighton, T.E. (1993). Dissecting the disulphide-coupled folding pathway of bovine pancreatic trypsin inhibitor. Forming the first disulphide bonds in analogues of the reduced protein. J. Mol. Biol. *232*, 873–896.

Daub, H., Specht, K., and Ullrich, A. (2004). Strategies to overcome resistance to targeted protein kinase inhibitors. Nat. Rev. Drug Discov. *3*, 1001–1010.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

Doss, C.G., and Nagasundaram, N. (2012). Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. PLoS One *7*, e31677.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. *39*, D945–D950.

Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Ginsburg, G.S., and McCarthy, J.J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol. *19*, 491–496.

Glusman, G., Rose, P.W., Prlic, A., Dougherty, J., Duarte, J.M., Hoffman, A.S., Barton, G.J., Bendixen, E., Bergquist, T., Bock, C., et al. (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med. *9*, 113.

Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. *88*, 440–449.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517.

Hong, M.K., Macintyre, G., Wedge, D.C., Van Loo, P., Patel, K., Lunke, S., Alexandrov, L.B., Sloggett, C., Cmero, M., Marass, F., et al. (2015). Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nat. Commun. *6*, 6605.

Janne, P.A. (2008). Challenges of detecting EGFR T790M in gefitinib/erlotinib-resistant tumours. Lung Cancer *60* (*Suppl 2*), S3–S9.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database (Oxford) *2011*, bar049.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. (2016). PubChem substance and compound databases. Nucleic Acids Res. *44*, D1202–D1213.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kobayashi, S., Boggon, T.J., Dayaram, T., Janne, P.A., Kocher, O., Meyerson, M., Johnson, B.E., Eck, M.J., Tenen, D.G., and Halmos, B. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N. Engl. J. Med. *352*, 786–792.

Kosaka, T., Yatabe, Y., Endoh, H., Yoshida, K., Hida, T., Tsuboi, M., Tada, H., Kuwano, H., and Mitsudomi, T. (2006). Analysis of epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer and acquired resistance to gefitinib. Clin. Cancer Res. *12*, 5764–5769.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Kumar, A., Rajendran, V., Sethumadhavan, R., and Purohit, R. (2013). Molecular dynamic simulation reveals damaging impact of RAC1 F28L mutation in the switch I region. PLoS One *8*, e77453.

Kumar, S., Clarke, D., and Gerstein, M. (2016). Localized structural frustration for evaluating the impact of sequence variants. Nucleic Acids Res. *44*, 10062–10073.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. *161*, 269–288.

Kunzmann, V., Bauer, E., and Wilhelm, M. (1999). Gamma/delta T-cell stimulation by pamidronate. N. Engl. J. Med. *340*, 737–738.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Laing, R.E., Hess, P., Shen, Y., Wang, J., and Hu, S.X. (2011). The role and impact of SNPs in pharmacogenomics and personalized medicine. Curr. Drug Metab. *12*, 460–486.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Levitt, M., and Warshel, A. (1975). Computer simulation of protein folding. Nature *253*, 694–698.

Liu, Y.L., Lindert, S., Zhu, W., Wang, K., McCammon, J.A., and Oldfield, E. (2014). Taxodione and arenarone inhibit farnesyl diphosphate synthase by binding to the isopentenyl diphosphate site. Proc. Natl. Acad. Sci. U S A *111*, E2530–E2539.

Madian, A.G., Wheeler, H.E., Jones, R.B., and Dolan, M.E. (2012). Relating human genetic variation to variation in drug responses. Trends Genet. *28*, 487–495.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Martin, M.B., Grimley, J.S., Lewis, J.C., Heath, H.T., 3rd, Bailey, B.N., Kendrick, H., Yardley, V., Caldera, A., Lira, R., Urbina, J.A., et al. (2001). Bisphosphonates inhibit the growth of *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania donovani*, *Toxoplasma gondii*, and *Plasmodium falciparum*: a potential route to chemotherapy. J. Med. Chem. *44*, 909–916.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature *267*, 585–590.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics *10*, 213.

Meyer, U.A., Zanger, U.M., and Schwab, M. (2013). Omics and drug response. Annu. Rev. Pharmacol. Toxicol. *53*, 475–502.

Meyer, M.J., Beltran, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nat. Methods *15*, 107–114.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comput. Chem. *30*, 2785–2791.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: progress and new directions in round XI. Proteins *84* (*Suppl 1*), 4–14.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. *17*, 53.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Pao, W., Miller, V.A., Politi, K.A., Riely, G.J., Somwar, R., Zakowski, M.F., Kris, M.G., and Varmus, H. (2005). Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med. *2*, e73.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Peterson, T.A., Doughty, E., and Kann, M.G. (2013). Towards precision medicine: advances in computational approaches for the analysis of human variants. J. Mol. Biol. *425*, 4047–4063.

Pires, D.E., Blundell, T.L., and Ascher, D.B. (2015). Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res. *43*, D387–D391.

Raphael, B.J., Dobson, J.R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. *6*, 5.

Reichold, M., Zdebik, A.A., Lieberer, E., Rapedius, M., Schmidt, K., Bandulik, S., Sterner, C., Tegtmeier, I., Penton, D., Baukrowitz, T., et al. (2010). KCNJ10 gene mutations causing EAST syndrome (epilepsy, ataxia, sensorineural deafness, and tubulopathy) disrupt channel function. Proc. Natl. Acad. Sci. U S A *107*, 14490–14495.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Rose, P.W., Prlic, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., et al. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. *43*, D345–D356.

Russell, R.G. (2011). Bisphosphonates: the first 40 years. Bone *49*, 2–19.

Schärfe, C.P.I., Tremmel, R., Schwab, M., Kohlbacher, O., and Marks, D.S. (2017). Genetic variation in human drug-related genes. Genome Med. *9*, 117.

Sethi, A., Clarke, D., Chen, J., Kumar, S., Galeev, T.R., Regan, L., and Gerstein, M. (2015). Reads meet rotamers: structural biology in the age of deep sequencing. Curr. Opin. Struct. Biol. *35*, 125–134.

Shipman, C.M., Croucher, P.I., Russell, R.G., Helfrich, M.H., and Rogers, M.J. (1998). The bisphosphonate incadronate (YM175) causes apoptosis of human myeloma cells in vitro by inhibiting the mevalonate pathway. Cancer Res. *58*, 5294–5297.

Smith, R.D., Damm-Ganamet, K.L., Dunbar, J.B., Jr., Ahmed, A., Chinnaswamy, K., Delproposto, J.E., Kubish, G.M., Tinberg, C.E., Khare, S.D., Dou, J., et al. (2016). CSAR benchmark exercise 2013: evaluation of results from a combined computational protein design, docking, and scoring/ranking challenge. J. Chem. Inf. Model. *56*, 1022–1031.

Spear, B.B., Heath-Chiozzi, M., and Huff, J. (2001). Clinical application of pharmacogenetics. Trends Mol. Med. *7*, 201–204.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. J. Mol. Biol. *425*, 3919–3936.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.

Szpiech, Z.A., Strauli, N.B., White, K.A., Ruiz, D.G., Jacobson, M.P., Barber, D.L., and Hernandez, R.D. (2017). Prominent features of the amino acid mutation landscape in cancer. PLoS One *12*, e0183273.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461.

De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. J. Med. Chem. *59*, 4035–4061.

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., and Hou, T. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. Phys. Chem. Chem. Phys. *18*, 12964–12975.

Ward, A.B., Sali, A., and Wilson, I.A. (2013). Biochemistry. Integrative structural biology. Science *339*, 913–915.

Webb, B., and Sali, A. (2016). Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci. *86*, 291–2937.

Wilkinson, G.R. (2005). Drug metabolism and variability among patients in drug response. N. Engl. J. Med. *352*, 2211–2221.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

Wood, J., Bonjean, K., Ruetz, S., Bellahcene, A., Devy, L., Foidart, J.M., Castronovo, V., and Green, J.R. (2002). Novel antiangiogenic effects of the bisphosphonate compound zoledronic acid. J. Pharmacol. Exp. Ther. *302*, 1055–1061.

Yan, C., Grinter, S.Z., Merideth, B.R., Ma, Z., and Zou, X. (2016). Iterative knowledge-based scoring functions derived from rigid and flexible decoy structures: evaluation with the 2013 and 2014 CSAR benchmarks. J. Chem. Inf. Model. *56*, 1013–1021.

Yun, C.H., Boggon, T.J., Li, Y., Woo, M.S., Greulich, H., Meyerson, M., and Eck, M.J. (2007). Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell *11*, 217–227.

Zhan, Y., and Guo, S. (2015). Three-dimensional (3D) structure prediction and function analysis of the chitin-binding domain 3 protein HD73_3189 from *Bacillus thuringiensis* HD73. Biomed. Mater. Eng. *26* (*Suppl 1*), S2019–S2024.

Zhang, Z., Miteva, M.A., Wang, L., and Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. Comput. Math. Methods Med. *2012*, 805827.

Zhang, Z., Norris, J., Kalscheuer, V., Wood, T., Wang, L., Schwartz, C., Alexov, E., and Van Esch, H. (2013). A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. Hum. Mol. Genet. *22*, 3789–3797.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U S A *111*, E455–E464.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
GenoDock Dataset	This paper	http://genodock.molmovdb.org/download
Structure of proteins in GenoDock dataset	RCSB PDB (Berman et al., 2000)	https://www.rcsb.org/
Structure of Human Epidermal Growth Factor Receptor	RCSB PDB (Yun et al., 2007)	PDB: 2ITY
Structure of Farnesyl Diphosphate Synthase	RCSB PDB (Liu et al., 2014)	PDB: 4P0W
Germline exonic variants	ExAC release (Lek et al., 2016)	ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/
Somatic exomic variants	TCGA network (http://cancergenome.nih.gov)	http://portal/gdc.cancer.gov/repository
Genomic Evolutionary Rate Profiling	Davydov et al., 2010	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html
FDA-approved drug ligands	DrugBank (Wishart et al., 2018)	https://www.drugbank.ca/
Drug ligand physiochemical features	PubChem (Kim et al., 2016)	https://pubchem.ncbi.nlm.nih.gov/
Platinum Dataset	Pires et al., 2015	http://biosig.unimelb.edu.au/platinum/
Software and Algorithms		
GenoDock method	This paper	http://genodock.molmovdb.org/download
SNV-PDB mapping algorithm	Kumar et al., 2016	N/A
AutoDock Vina	Trott and Olson, 2010	http://vina.scripps.edu/
AutoDock 4	Morris et al., 2009	http://autodock.scripps.edu/
Mdock	Yan et al., 2016	http://zoulab.dalton.missouri.edu/mdock.htm
Modeller	Webb and Sali, 2016	https://salilab.org/modeller/
Scikit-learn package	Pedregosa et al., 2011	https://scikit-learn.org/
PyMOL molecular graphics system	ver. 2.0, Schrodinger, LLC	https://pymol.org/2/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Mark Gerstein (mark@gersteinlab.org).

METHODS DETAILS

GenoDock Database Preparation

Germline exonic variants were collected from the ExAC release 1 (Lek et al., 2016) (download source: ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/). Somatic exonic variants came from TCGA network (http://cancergenome.nih.gov; download source: http://portal/gdc.cancer.gov/repository). "Simple Nucleotide Variation", "Masked Somatic Mutation", and "MuTect2 Variant Aggregation and Masking" served as filters for "Data Category", "Data Type", and "Workflow Type", respectively. The list of FDA-approved drug ligands was directly obtained from DrugBank (Wishart et al., 2018). Human protein 3D structures with a resolution better than 3.0 Å were downloaded from the PDB (https://www.rcsb.org/) (Berman et al., 2000). A careful curation to filter structures in the PDB that contain FDA-approved drug molecules was conducted. The mapping of the variants from both the ExAC and TCGA datasets to the curated co-crystal PDB structures was done using a modified version of a previously published method (Kumar et al., 2016). For tagging common and rare variants from the ExAC dataset, a variant occurrence cutoff of 1 was used to differentiate rare SNVs from common ones: if a variant occurred only once in the ExAC dataset, we tagged this SNV as rare; if a variant occurrence count is greater than 1, we tagged it as a common SNV. For tagging driver and passenger SNVs from TCGA dataset, SNVs were tagged as enriched-in-driver variants if they were variants in cancer driver genes listed in the Catalogue of Somatic Mutations in

Cancer (COSMIC), version 83. If a variant was not in a cancer driver gene, we tagged it as a passenger variant. Not all SNVs in driver genes are driver variants, but they are more likely to be driver variants, which is sufficient for the purpose of this study.

Mutant Structure and Binding Affinity Change Calculation

For each entry recorded in our database, we generated a mutant structure associated with that SNV through homology modeling using Modeller (ver. 9.18) (Webb and Sali, 2016), using the corresponded native co-crystal structure as a template. During the modeling process, adjustments were made to the target residue under stereo-chemical and homology-derived constraints, followed by a minimization step of the constraints to deliver the final mutant structure. In this project, 10,283 mutant PDB structures were generated in total.

For each native-mutated protein structure pair, we used Vina (Trott and Olson, 2010) to evaluate the change in drug binding affinity to setup the pseudo gold-standard set: $\Delta BA = \Delta G(MUT) - \Delta G(WT)$, in kcal/mol, where $\Delta G(MUT)$ and $\Delta G(WT)$ are binding affinities of the drug with the mutated and wild-type protein target, respectively, evaluated using Vina. During the calculation, we fixed the protein structure to avoid concerns from protein flexibility. ‘‘Local optimization’’ was applied for the ligand-binding model, and the ‘‘Vina score’’ was set as the scoring function. Due to the lack of experimentally measured LBA data for every entry in GenoDock dataset, we validated the calculations of Vina by applying the same procedure with AD4 (ver. 6.2.6) (Morris et al., 2009) and MDock (Yan et al., 2016) to check the consistency of the results. The three programs evaluate the binding affinity scores via different score functions. Vina employs a score function derived from both the conformational preferences of protein-ligand complexes and the experimental binding affinities, which thus combines the advantages of empirical scoring functions and knowledge-based potentials. AD4 employs a purely empirical scoring function with the parameters derived only by fitting the experimental affinity measurements. MDock employs statistical potentials derived from the protein-ligand complex structures. If for a given structure pair, ΔBA values calculated by two out of the three scoring methods were of the same sign (both positive, indicating both tools assigned a drug binding disruptive role to the SNV; or both non-positive, indicating both tools assigned a drug binding non-disruptive role to the SNV), then we regard the result as consistent between these two methods. Vina and AD4 achieved a consistency of 84%; Vina and MDock achieved a consistency of 93%. The two sets of results from Vina and AD4 reached a Pearson product-moment correlation of 0.89, and that of 0.94 between Vina and MDock (Figure S5B), indicating strong consistency.

To further access the rigorousness of our ligand-binding model on binding affinity change calculations to evaluate the pseudo gold-standard used for our core statistical model, we conducted docking calculations using available Platinum experimental records. Modeller and Vina were used to generate the mutant structure and to score the binding affinity change. We followed the same pipeline to get binding affinity change as we did for the GenoDock pseudo gold-standard set. For the ~120 entries in this test, the docking calculation and the experimentally measured set reached a consistency of 71.3% with a Matthews Correlation Coefficient of 0.272, indicating a reasonable scientific rigorousness of our docking calculations to construct the pseudo gold-standard set.

Features Extraction and Construction for the Machine-Learning Method

SNV Features

SIFT and PolyPhen-2: SIFT score and Polyphen-2 score for somatic and germline exonic SNVs in our study were directly extracted from the ‘‘INFO’’ column of VCF files from ExAC consortium and TCGA project.

GERP: GERP scores were retrieved directly from the Sidow lab (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html) (Davydov et al., 2010).

Ligand Features

Ligand features including molecular weight, H-bond donor and acceptor count, rotatable bond count, and polar surface area for each drug molecule in our database were extracted from PubChem database (Kim et al., 2016).

Structure Features

Amino acid side chain volume change index: defined as $\Delta V_{index} = log_2\left(\frac{V_{MUT}}{V_{WT}}\right)$, where V_{MUT} and V_{WT} stand for van der Waals volume (Darby and Creighton, 1993) of the mutant and wild-type protein residue, respectively.

Amino acid side-chain hydropathy change: for each amino acid, we employed the amino acid hydropathy scale by Kyte and Doolittle (Kyte and Doolittle, 1982) as the hydropathy metric. Amino acid side chain hydropathy change index is defined as $\Delta \text{hydropathy} = \text{hydropathy}(\text{mutant}) - \text{hydropathy}(\text{WT})$.

Distance between mutation and drug ligand: The distance between a protein residue to a ligand was defined as the shortest distance of a heavy atom of that residue to a heavy atom of the associated ligand.

Binding site (‘‘on’’/‘‘off’’): This is a binary feature describing whether the mutation is inside or outside the binding site. If a residue had a distance less than 8Å from the target ligand in the co-crystal structure, we considered that this residue is in the binding pocket. Though this feature is grouped into the structure feature set, it could still be used when only one out of the drug ligand structure and protein structure is available. We constructed the ‘‘SNV + Ligand’’ model and ‘‘SNV + Structure’’ model under two scenarios: ‘‘Binding Site’’ is known and ‘‘Binding Site’’ is unknown. The former model was trained with the ‘‘Binding Site’’ feature included, and users need to tell GenoDock whether the SNV of interest is associated with residues on or off the binding pocket. The latter model was trained with the ‘‘Binding Site’’ feature excluded. In practice, with the ‘‘Binding Site’’ being ‘‘on’’, we are able to predict the maximal probability of the target SNV being ligand-binding disruptive. Users are also free to set the ‘‘Binding Site’’ to ‘‘off’’ if they want the prediction for protein residues of associated variants that are not in binding sites. When the user does not care about binding status when applying the "SNV + Ligand" or "SNV + Structure" models, they can use remove this feature and set the "Binding Site" to unknown. We engineered GenoDock source scripts for both application scenarios.

During feature engineering and exploration, p-values between the median values of disruptive and non-disruptive SNV groups for each feature were calculated via a two-sample Wilcoxon test.

Protein-Ligand Complex Visualization

All figures regarding protein-ligand complex were generated by the PyMOL molecular graphics system, Version 2.0 Schrödinger, LLC.

QUANTIFICATION AND STATISTICAL ANALYSIS

Training, Testing, and Evaluating the Performance of a Machine-Learning Method

GenoDock dataset was separated into a training set (70%) and test set (30%) in a random manner. To avoid potential bias raised from imbalanced composition of the two classes of samples in our dataset (735 entries for disruptive SNVs; 9,458 entries for non-disruptive SNVs), we counted the number of disruptive SNV samples ($\Delta BA > 0$) and randomly selected an equal number of non-disruptive SNV samples from ($\Delta BA \leq 0$) to make up the balanced training set. The Scikit-learn package (Pedregosa et al., 2011) was used for learning model development. We tested classification methods including Lasso Regression, Support Vector Machine, Random Forest, and Gradient Boosting Decision Tree. We trained each learning model through a tenfold grid-search cross-validation process. For each training set, the additional 30% data was tested for performance evaluation. Based on the AUC values, Random Forest had the highest AUC among all methods (Figure S4B). Feature selection was performed by evaluating the AUC for each feature. If the selection power of a feature was near or worse than random selection, we removed it from our feature pool (e.g., allele frequency). With the same procedure, we trained and optimized a random forest model for each of the four feature combinations (SNV only; SNV + Structure; SNV + Ligand; SNV + Structure + Ligand) for GenoDock.

Curation of the Independent Test Set Based on Experimental Measurements

We also prepared an independent benchmark dataset comprising experimentally measured binding affinity change upon mutation from the Platinum database (Pires et al., 2015). Briefly, the full Platinum database content was downloaded as a flat comma-separated file from http://biosig.unimelb.edu.au/platinum/. Amino acid mutations other than single-point mutations and those found in species other than human were excluded. In addition, mutations that do not result from SNV were also removed because GenoDock uses the GERP score as one of the predictive features, and the GERP score is position specific. Further, mutations that could not be mapped onto their associated UniProtKB canonical amino acid sequences were discarded. In the end, 86 unique data points were obtained (two data points with the same mutation but different ligands were considered to be different) and used as the independent test set. Each data point in this set was labeled as "disruptive" if its associated fold change in binding affinity upon mutation, $\left(\frac{BA_{WT}-BA_{MUT}}{BA_{WT}}\right)$, was negative or "non-disruptive" otherwise. Note that the curation of this test set was conducted in a manner that was completely blinded from the training of GenoDock. Ligand features and structure features were assigned for each mutation entry for the database to run GenoDock. We then applied "SNV + Structure + Ligand" model (auROC = 0.97) on the dataset to evaluate the reliability of GenoDock predictions.

DATA AND SOFTWARE AVAILABILITY

To make our method easily accessible, we provide a web interface called the GenoDock web server (http://genodock.molmovdb.org/). We tailored GenoDock to four separate models (with different levels of prediction accuracy) based on the accessibility of input features to broaden the applicability of our tool. Users can import their sample data using the GenoDock graphic user interface with different feature-set combinations: SNV feature only; SNA feature and structure feature; SNV feature and ligand feature, and all three groups of features. The predicted result will be returned in the form of an HTML webpage. The calculation page can be accessed at http://genodock.molmovdb.org/calculation/0. Users can also download our open-source Python code to run large-scale inputs on local computers or on high-performance computing clusters.

Structure, Volume *27*

Supplemental Information

Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions

Bo Wang, Chengfei Yan, Shaoke Lou, Prashant Emani, Bian Li, Min Xu, Xiangmeng Kong, William Meyerson, Yucheng T. Yang, Donghoon Lee, and Mark Gerstein

Supplementary Figures and Tables

S1. Detailed flowchart of SNV mapping and ligand-binding modeling (related to Figure 1)

The construction of the GenoDock dataset mainly consists of three steps. First, we mapped the SNVs onto PDB structures, together with their annotation scores including polyphen-2, GERP and SIFT. Second, for each associated amino acid residue on a PDB structure, we model the point mutant structure and conduct physical calculations for binding affinity change. Third, we construct other physiochemical features on the structure side including polarity change, side chain volume change, etc. to finish the GenoDock data source.

S2. Comparison of protein-drug co-crystal structure covered by GenoDock database and those by three alternative databases including PDBBind_DB, PDID, and DrugBank (related to Figure 1)

In order to compare our GenoDock database with existing protein-drug interaction databases, we match PDB entries of PDB Bind DB, Drug Bank and Protein-Drug Interaction DB (PDID) with the FDA drug list to filter out protein-drug crystal-structures that meets our criteria. The Venn Diagram depicts the overlap between the 4 datasets. The GenoDock database is filtered from the Whole RCSB PDB library, it covers the most protein-drug structures (795) and drug molecules (264) compared with the other three. Part of structures from PDBBind_DB and PDID are excluded from our database, and this is due to the resolution of PDBs did not meet our 3.0Å baseline.

S3. GenoDock dataset characterization (related to Figure 1, Figure2, and Figure3)

S3 (a) Heat map for amino acid mutation landscape in GenoDock database (related to Figure 1)

X-axis and y-axis refer to types of mutated amino acids and wild type amino acids, respectively. Different counts for each mutation pair are colored from white to cyan. The percentage distribution in wildtype and mutated amino acid pools are shown outside of the heat map in green and purple, respectively. In the heat map, the two most abundant mutation pairs are arginine to cysteine and arginine to histidine, which are referred as "mutation signatures" in previous literatures. Analyzing the mutational landscape of our database is very useful for the ensuing study of how a point mutation affects drug efficacy, which is further tailored to be sensitive to how side-chains interact with ligands differently before and after mutational replacement. Within the GenoDock database, we found that the two most abundant mutations are arginine to cysteine and arginine to histidine. This observation is consistent with previous expectations: first, arginine is the most frequently occurring native amino acid among the somatic mutations and germline variants that can be mapped onto a PDB structure in our protein pool (14% in wild-type distribution,; second, the arginine-to-cysteine mutation is also the most common mutation identified that causes human disease in disease-associated variant datasets such as the Human Gene Mutation Database (Stenson et al., 2014), the Online Database of Mendelian Inheritance in Man (Hamosh et al., 2005), and ClinVar (Hamosh et al., 2005; Landrum et al., 2014; Peterson et al., 2013; Stenson et al., 2014); third, many cancer mutation signatures are enriched in the arginine-to-histidine mutation (Peterson et al., 2013). Previous literature shows that mutation from arginine to histidine can confer protein pH sensitivity to the mutant and thus alter protein function leading to disease (Reichold et al., 2010; Szpiech et al., 2017; Zhang et al., 2012).

S3 (b) Percentage distribution of polarity from amino acid point mutation by SNV in TCGA and ExAC dataset (related to Figure 1)

We group amino acid residues into three groups according to their side chain polarity, and the polarity index assigned to each group of residues are listed below:

C (charged residue; polarity index = 1 or -1) including GLU, ASP, ARG, LYS;

P (polar residue; polarity index = 0.5) including GLN, ASN, HIS, SER, THR, TYR;

H (hydrophobic residue; polarity index = 0) including ALA, ILE, LEU, MET, PHE, VAL, PRO, GLY.

In the figure, we visualize the distribution of mutation pairs in terms of polarity change between the native and mutated residue. Overall, we observed that around one third of somatic SNVs lead to point mutations from a charged amino acid residue to a polar one; whereas among the germline variants, the most frequently occurred mutations are between two hydrophobic amino acids.

S3 (c) Percentage distribution of positive and non-positive samples in GenoDock data source from ExAC and TCGA database (related to Figure 2)

We can see that in the ExAC and TCGA datasets, the frequency of disruptive SNVs (positive binding affinity change) is 6.0% and 8.9%, respectively.

S3 (d) Comparison of top common and rare SNV pairs with most positive binding affinity changes in ExAC dataset of GenoDock data set (related to Figure2)

We compare the top range of common and rare nsSNVs from EXAC to see the difference of two sets of germline variants in terms of protein-drug binding destabilization. We select the same number of nsSNVs from the common set and the rare set each time and compare the difference. We observe that rare variants have a greater extent of positive binding affinity shift than common ones, and the difference is significant from the two-body Wilcoxon test. Also, the p-value is decreasing as we compare a larger number of nsSNVs from the two sets.

S3 (e) Boxplot distribution between disruptive SNV (binding affinity change >0) and non-disruptive SNV (binding affinity change <0 or =0): other 4 features in model (related to Figure 3)

Among various features we select for final feed for our classifier, some of the features are not strong enough in terms of sensitivity of prioritizing disruptive nsSNV candidates to be kept in our final model. We here list the distributions between positive class and non-positive class of nsSNVs in our database in term of H-bond acceptor/donor count and rotatable count of drug ligands associated with each nsSNVs, as well as the allele frequency of nsSNV.

S4. Statistical learning model construction, performance benchmark and feature significance analysis (related to Figure 4)

S4 (a) Flowchart of machine learning method for GenoDock

We start our supervised learning model construction by dividing the GenoDock data source into 70% and 30% for training and testing purpose, respectively. We tune the sample nsSNV count of "Class 1" ($\Delta BA >$

1) and “Class 2” ($\Delta BA \leq 0$) to be 1:1 to avoid potential bias. During the training process, we employ a 10-fold cross-validation training and testing process on different learning algorithms (Random Forest, Lasso Regression, Support Vector Machine, and Gradient Boosting Decision Tree) to pick up the one with best performance. To make our features more efficient, we conduct a feature selection process by evaluating AUC of each feature independently to make sure our features are sensitive enough.

S4 (b) Model Selection among RF, LR, SVM and GBDT

We follow the same pipeline to evaluate the sensitivity of each model to pick up the best one in terms of prioritizing disruptive nsSNV candidates. Random Forest classifier stands out.

The Scikit-learn package is used for learning model development:

random forest: RandomForestClassifier ;
lasso regression: linear_model.Lasso;
support vector machine: SVR;
gradient boost decision tree: GradientBoostingClassifier

S4 (c) Performance comparison of GenoDock with other SNV annotation tools

S4 (d) Gini distance for relative feature significance in four models

We employ Gini distance as a measurement for feature importance in 4 models of GenoDock. We find GERP score, amino acid side chain volume change, polar surface area of drug ligand, distance between mutated amino acid residue and drug ligand are the most important features in SNV annotation features, PDB features, ligand features, and structure features, respectively. With more features feeding into our classifier, the significance of each feature is stable across different models. Particularly, binding site is an important feature if there is at least one structural component (protein PDB, drug ligand or co-crystal structure) present during the classification process of GenoDock. If the protein residue associated with the nsSNV of interest is not on binding pocket, the probability of this nsSNV to disrupt the drug-protein binding is much smaller than those nsSNVs that are associated with binding pocket residues.

S5. Validation of statistical learning model on experimental dataset and rigorousness assessment of physical calculations on binding affinity change (related to Figure 1)

S5 (a) Performance evaluation of GenoDock on Platinum Experimental Dataset

We parse Platinum dataset which contains experimental assay data on binding affinity change upon point mutations, and only keep proteins from homo sapiens. For the 86 data entries parsed from Platinum, we apply GenoDock (“SNV annotations + Ligand + Structure” model) to process the dataset and plot the ROC. The AUC reaches 0.62, which is a reasonable result that helps validate the reliability of GenoDock predictions, given that the data source collected in Platinum is highly heterogeneous from across different labs, publication years, assay methods, and experiment pipelines. We also evaluate reliability of GenoDock by calculate precision of predictions based on our parsed Platinum dataset. We observe that by setting the cutoff to be 0.5 (when predicted probability of $\Delta BA > 0$ is greater than 0.5, we assign this data entry as “disruptive”; otherwise we assign the data entry as “non-disruptive”), we can reach a precision of 1.84 and a recall of 0.59 after comparing this predicted result with experimental binding affinity change recorded in Platinum. Among the 86 samples, we have True Positive count = 7, True Negative count = 13, False Positive count = 7, False Negative count = 27.

S5 (b) Binding affinity calculation validation by AutoDock Tools (ADT) and MDock

In order to check the reliability of the binding affinity change calculations between native and mutated amino acid residues by Vina, we evaluate the consistency between results given by Vina with those from ADT and MDock. Vina, ADT and MDock have different scoring functions. The Pearson product-moment correlation (PPMC) between two sets of results by Vina and ADT for each nsSNV entry in our GenoDock database is 0.89, and that between Vina and MDock reaches 0.94. It indicates a strong consistency of the results given by Vina. Due to the lack of experimental data from ligand binding assays for SNV-Ligand-PDB entries in our database at the current stage, checking the ΔBA prediction reliability via calculation consistency with another package is a practical way. When more experimental data are available, we will update the experimental binding affinity change data into our pipeline for better classification results.

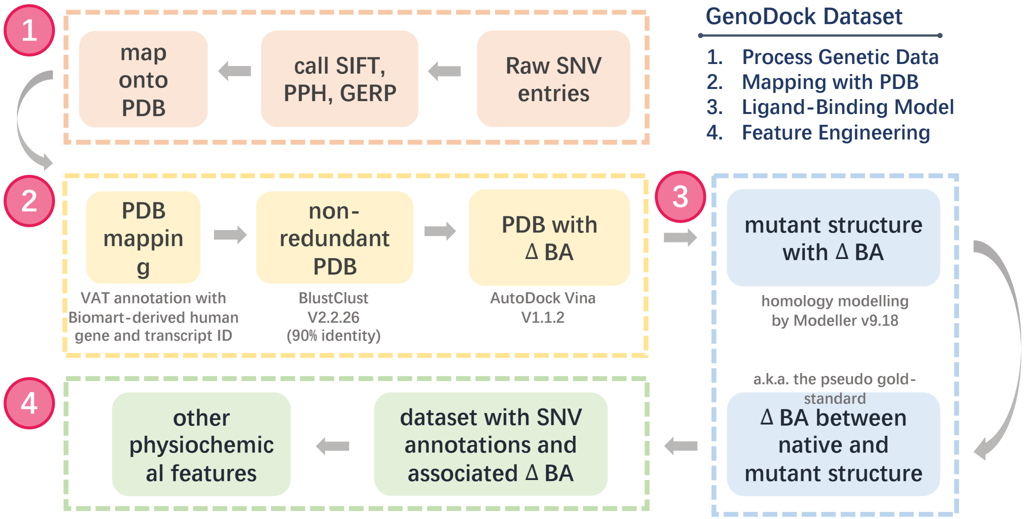
Supplementary Table 1. AUROC list of different GenoDock models and of other benchmark methods (related to Figure 4)

Supplementary Table 2. List of models and datasets in the study (related to Figure 1)

We list the statistical model (Model #1) and the ligand binding model (Model #2) with instructions of their role, validation method, and how we parameterize them. Model #1 is the core model and the main method we aim to deliver in this study. We also list the two datasets used for parameterization and for validation purpose for training and testing the core model. Note that we though did not directly validate Model #2 containing the pseudo gold standard due to unavailable experimental measurements, we did access the rigorousness of Model #2 by testing the accuracy of docking calculations on an independent experiment dataset (see Methods for details of evaluating Model #2), and the reasonable rigorousness of the model indicates a good reliability of our pseudo gold standard set.

S1

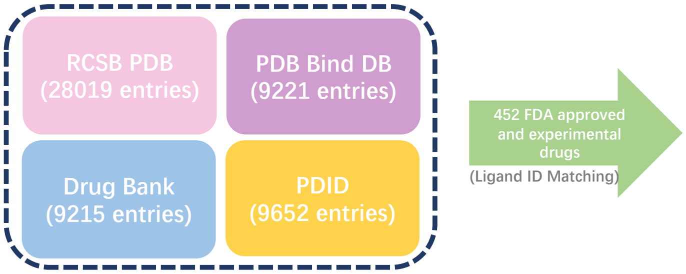
Detailed Flowchart of SNV Mapping and Ligand-binding Modeling


1
map onto PDB
call SIFT, PPH, GERP
Raw SNV entries
GenoDock Dataset
1. Process Genetic Data
2. Mapping with PDB
3. Ligand-Binding Model
4. Feature Engineering
2
PDB mapping
VAT annotation with Biomart-derived human gene and transcript ID
non-redundant PDB
BlustClust V2.2.26 (90% identity)
PDB with Δ BA
AutoDock Vina V1.1.2
3
mutant structure with Δ BA
homology modelling by Modeller v9.18
a.k.a. the pseudo gold-standard
Δ BA between native and mutant structure
4
other physiochemical features
dataset with SNV annotations and associated Δ BA

*Genomic coordinates of variants were standardized to the human reference assembly GRCh37

S2

Comparison of Protein-Drug Co-Crystal Structure Covered by RCSB PDB and these by Three Other Databases



RCSB PDB (28019 entries)

PDB Bind DB (9221 entries)

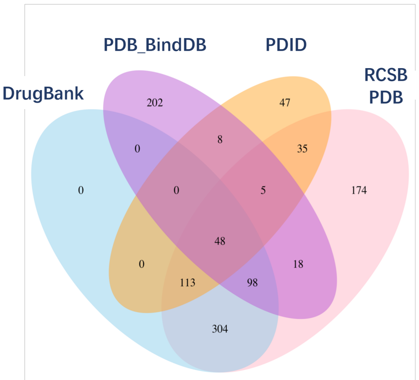
Drug Bank (9215 entries)

PDID (9652 entries)

452 FDA approved and experimental drugs

(Ligand ID Matching)

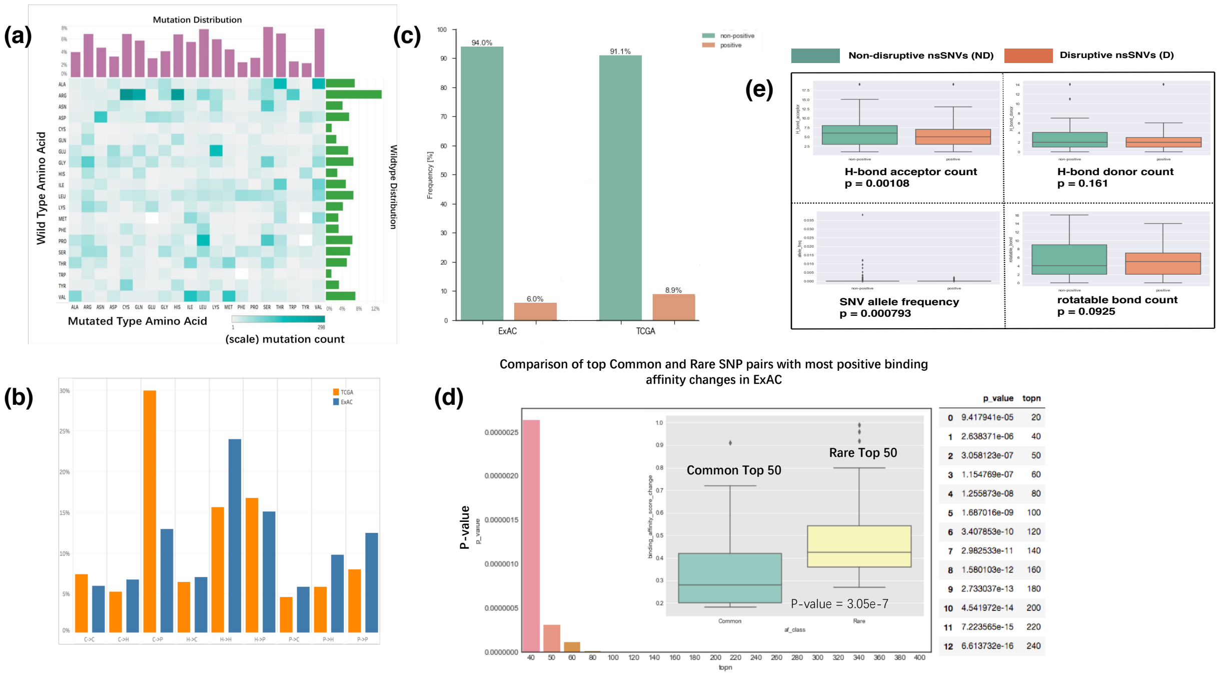
Whole RCSB PDB pre-filter: Resolution better than 3.0Å; Human Protein only; No DNA/RNA/hybrid chain



Data Set	PDB covered	Drug covered
RCSB PDB	**795**	**264**
PDB Bind DB	379	177
Drug Bank	563	209
PDID	256	112

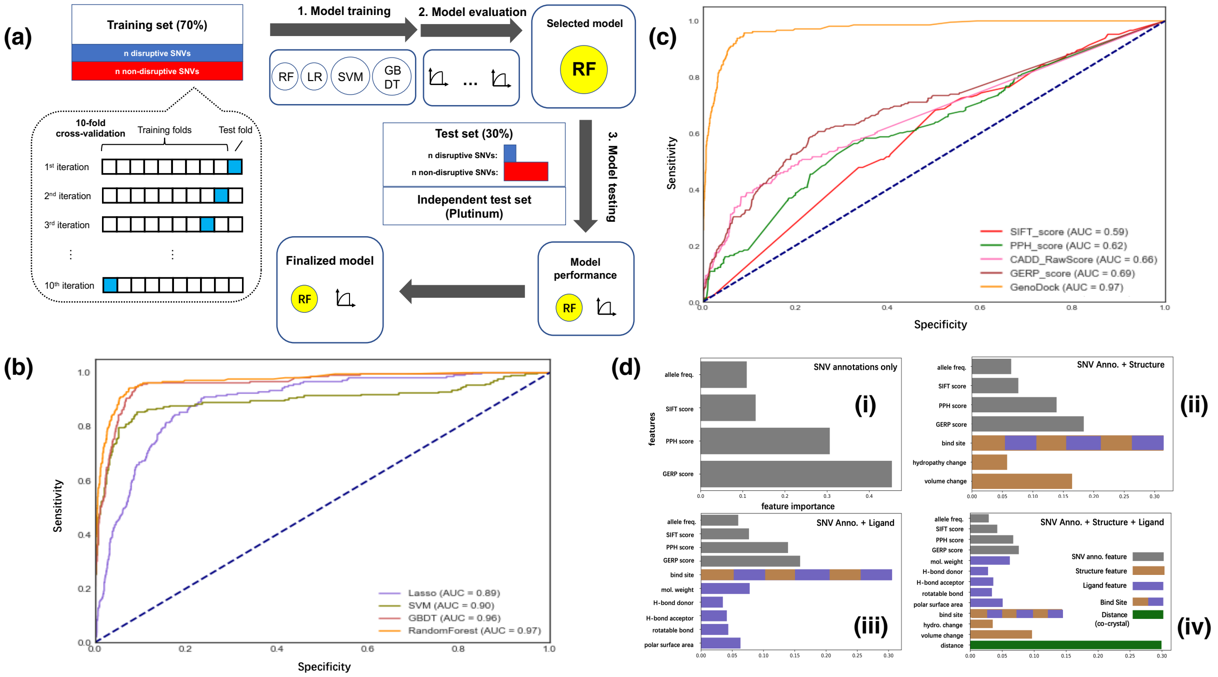
- RCSB PDB covers the most co-crystal structures with FDA approved drugs
- Further mapping ExAC and TCGA SNVs on to the curated RCSB PDB dataset to get GenoDock dataset

S3

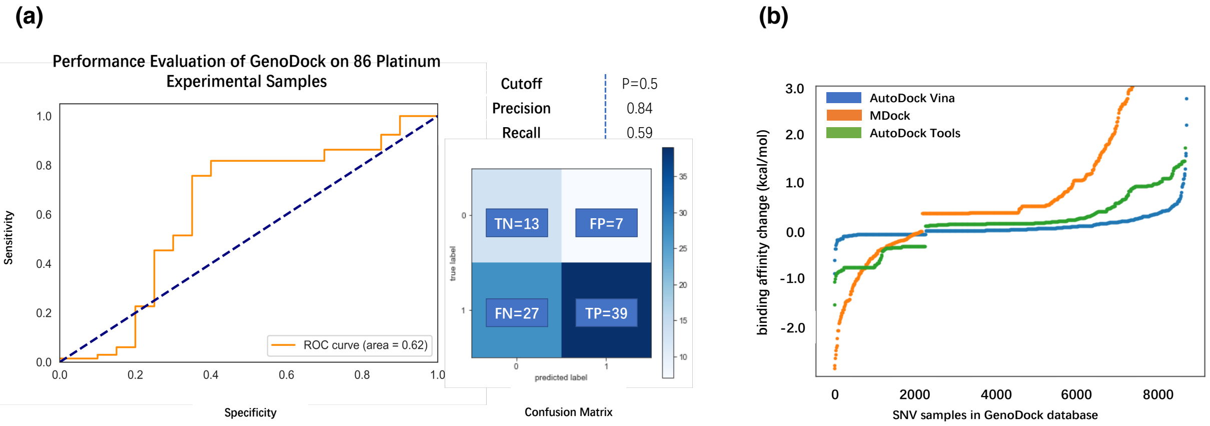


	p_value	topn
0	9.417941e-05	20
1	2.638371e-06	40
2	3.058123e-07	50
3	1.154769e-07	60
4	1.255873e-08	80
5	1.687016e-09	100
6	3.407853e-10	120
7	2.982533e-11	140
8	1.580103e-12	160
9	2.733037e-13	180
10	4.541972e-14	200
11	7.223565e-15	220
12	6.613732e-16	240

S4


(a)
Training set (70%)
n disruptive SNVs
n non-disruptive SNVs
10-fold cross-validation
Training folds
Test fold
1st iteration
2nd iteration
3rd iteration
10th iteration
1. Model training
RF
LR
SVM
GB DT
2. Model evaluation
...
Selected model
RF
3. Model testing
Test set (30%)
n disruptive SNVs:
n non-disruptive SNVs:
Independent test set (Plutinum)
Model performance
RF
Finalized model
RF
(b)
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Lasso (AUC = 0.89)
SVM (AUC = 0.90)
GBDT (AUC = 0.96)
RandomForest (AUC = 0.97)
(c)
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
SIFT_score (AUC = 0.59)
PPH_score (AUC = 0.62)
CADD_RawScore (AUC = 0.66)
GERP_score (AUC = 0.69)
GenoDock (AUC = 0.97)
(d)
SNV annotations only
(i)
features
allele freq.
SIFT score
PPH score
GERP score
0.0
0.1
0.2
0.3
0.4
feature importance
SNV Anno. + Structure
(ii)
allele freq.
SIFT score
PPH score
GERP score
bind site
hydropathy change
volume change
0.00
0.05
0.10
0.15
0.20
0.25
0.30
SNV Anno. + Ligand
(iii)
allele freq.
SIFT score
PPH score
GERP score
bind site
mol. weight
H-bond donor
H-bond acceptor
rotatable bond
polar surface area
0.00
0.05
0.10
0.15
0.20
0.25
0.30
SNV Anno. + Structure + Ligand
(iv)
allele freq.
SIFT score
PPH score
GERP score
mol. weight
H-bond donor
H-bond acceptor
rotatable bond
polar surface area
bind site
hydro. change
volume change
distance
SNV anno. feature
Structure feature
Ligand feature
Bind Site
Distance (co-crystal)
0.00
0.05
0.10
0.15
0.20
0.25
0.30

S5


(a)
Performance Evaluation of GenoDock on 86 Platinum Experimental Samples
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
ROC curve (area = 0.62)
Cutoff
P=0.5
Precision
0.84
Recall
0.59
TN=13
FP=7
FN=27
TP=39
0
1
true label
predicted label
35
30
25
20
15
10
Confusion Matrix
(b)
AutoDock Vina
MDock
AutoDock Tools
3.0
2.0
1.0
0.0
-1.0
-2.0
binding affinity change (kcal/mol)
0
2000
4000
6000
8000
SNV samples in GenoDock database

Supplementary Table S1: AUROC list of different GenoDock models and of other benchmark methods (related to Figure 4)

	Prediction Model	AUROC
benchmark methods	CADD	0.66
	SIFT	0.59
	Polyphen-2	0.62
	GERP	0.69
GenoDock models	“SNV only”	0.73
	“SNV + Structure” (Binding Site known)	0.92
	“SNV + Structure” (Binding Site unknown)	0.76
	“SNV + Ligand” (Binding Site known)	0.96
	“SNV + Ligand” (Binding Site unknown)	0.79
	“SNV + Structure + Ligand”	0.97

Supplementary Table S2: List of models and datasets in the study (related to Figure 1)

Model	Role	Parameterization	Validation	Description
1	Core Model	Statistical model from Ψ_{GS}	Platinum	Supervised learning model using the pseudo gold-standard set as target feature. The direct validation of this model is to apply the model to an independent, experiment-based validation dataset.
2	Auxillery Model	Physically based	-	A physical-based, previously published computational ligand-docking model to calculate binding affinity change for the pseudo gold standard set.

Dataset	Role	Size	Source	Description
Ψ_{GS}	Trains 1	~10k	Built from 2	Core dataset constructed for training the statistical model. Contains pseudo gold standard set as the target feature.
Platinum	Validates 1	86	Experiment	The human protein subset from Platinum. used as direct validation dataset of our statistical method.