# Response to reviewers for “INTEGRATING GENETIC AND STRUCTURAL FEATURES TO IDENTIFY VARIANTS DISRUPTING PROTEIN-DRUG INTERACTIONS”

# Response Letter

### Overall comment

### We want to thank the reviewers for endorsing our manuscript for publication, recognizing the novelty and importance of our method and study, and offering insightful comments. We have majorly revised the manuscript to address their concerns.

### The specific reviewers’ comments are further addressed below.

### Reviewer #1

### -- Ref1.0 –Positive comments--

|  |  |
| --- | --- |
| Reviewer  Comment | The aim of this paper is important and the topic is currently of great interest. Large scale predictions through data integration toward evaluation of drug-protein interactions - whether toward identification of new targets or as here disrupting protein-drug interactions are increasingly taken up by the community. |
| Author  Response | We thank the reviewer for acknowledging the novelty of our study, and for his/her thorough examination of our manuscript. |

### -- Ref1.1 –Writing quality of manuscript --

|  |  |
| --- | --- |
| Reviewer  Comment | The writing is poor. One example among many "Once personal carried genetic variants are identified"... The manuscript needs serious editing to be considered in a journal like Structure. |
| Author  Response | We thank the reviewer for the comment. We have carefully edited the manuscript thoroughly. Confusing statements and improper usage of terminologies have been modified. |

### -- Ref1.2 –Aim and purpose of paper--

|  |  |
| --- | --- |
| Reviewer  Comment | The aim of the paper is unclear. At the beginning, it is stated "But understanding the biophysical rationales of nsSNVs' implications towards drug efficacy remains difficult". Is the aim prediction or understanding? later on there is no discussion of understanding.  Also, "a well-constructed database that directly links genetic variants to reliable human drug-protein co-crystal structures is in great need". I understood that here the authors present a method, not a database? |
| Author  Response | We thank the reviewer for the comment. We apologize for the ambiguity from our manuscript writing. We have tried to clarify…   1. The main aim of this paper is to describe “hybrid” method in which a physically calculated “pseudo” gold standard set was used to construct our statistical learning model, given that the “real” gold standard test from experimentally measured ligand binding assay results are highly limited. We believe that the GenoDock method is a good addition to the community for rapidly prioritize potential nsSNV candidates that lead to protein-drug biding disruptions. During feature selection and exploration for GenoDock classifier construction, we focused on those physiochemical properties of protein structure and drug ligand molecule such as hydropathy change of amino acid side chain and ligand polar surface area to investigate how an nsSNV affects the binding interaction between ligand and protein residue through altering those properties of the mutant. The Gini distance analysis of feature significance helps reveal some evidence of the relative importance of those biochemical and biophysical properties which may further help us understand the rationales behind drug efficacy change. After all, the main scope of this study is to describe a method, with light touch of physiochemical evidence of how nsSNVs may affect drug efficacy based on our feature engineering and exploration findings. 2. Again, we are sorry for the term usage. We have changed “database” into “dataset” to emphasis that the main scope of the study is to present a classification method. On the other hand, the dataset we construct from scratch is indeed an important part of the study. To our knowledge, the GenoDock dataset is the largest set so far to bridge somatic and germline nsSNVs with high resolution protein-drug co-crystal structures, together with selected features (SNV annotation features and physicochemical features for SNV, protein and ligand molecules) for multi-purpose analysis. One important component of our dataset is the pseudo gold standard set of binding affinity change upon point mutation we calculated for each SNV-Structure-Ligand entry, and this record lays a reliable foundation for constructing our supervised learning model. In addition, the GenoDock dataset itself can independently serve as a valuable data collection set for further research projects involving nsSNV implications on protein-ligand binding activities. From the GenoDock web interface, the dataset is free for the community to download and use. |

### -- Ref1.3 –Reliable validation of GenoDock prediction--

|  |  |
| --- | --- |
| Reviewer  Comment | Prediction, or calculations of affinities is a very difficult problem, with many methods and publications on the subject (The authors need to cite these). Especially, the validation is here unclear. The authors use AutoDock Vina, a state of-the-art method. But they validate their method against a gold-standard also constructed with AutoDock Vina (and some test with Autodock). I understand that experimental data is scant, but here none is given. The results are not compared with other methods either. The absence of reliable validation is the Achilles heel of this paper. It is difficult to know to what extent to trust the predictions of the method. |
| Author  Response | 1. We thank the reviewer for the comments. Evaluation of the performance of a docking method is beyond the scope of this study, and our purpose is to leverage docking calculation results given experimentally measured data is not accessible in order to construct our statistical learning model. We totally agree with the reviewer that the binding affinity change calculations is a hard task, and we have added recent works and reviews on binding affinity calculation in our “Discussion” part of the current version manuscript (Ballester et al., 2014; Smith et al., 2016; Yan et al., 2016). We employ AutoDock Vina to construct our pseudo gold standard since the experimentally measured data is scant. As a well-established program suite, AutoDock Vina is relatively one of the most reliable docking programs available on the market. Here we added citations of relative works which have discussed the difficulties of binding affinity calculation, and they tested the performance of AutoDock calculations with other docking methods (Castro-Alvarez et al., 2017; Wang et al., 2016). 2. We totally agree that solely rely on computational calculations from docking program suites is not enough to evaluate the reliability of GenoDock predictions. We here add an independent test set containing “real gold standard” of binding affinity change for validation purpose. For this experimentally measured binding affinity change dataset, we employ Platinum dataset (Pires et al., 2015). It contains more than 1,000 ligand binding assay results associated with point mutations. We parsed the dataset to make it an independent test set (or, a validation set) for performance benchmark. After removing the entries that are not associated with a human protein, 87 entries left and serve as our independent test set for benchmark purpose. Detailed preparation of this dataset can be seen at “Methods” section. We then applied GenoDock to this parsed Platinum dataset as an independent test set to evaluate the performance of GenoDock. The AUC is 0.62 (Supplementary Figure 12), which is an acceptable result given that the ligand binding measurements recorded in Platinum is highly “heterogeneous” from a variety of labs and experimental setups, and that our gold standard for binding affinity change is based on docking calculations. In addition, we use a cutoff of 0.5 (predicted probability of being a disruptive nsSNV is greater than 0.5, we assign this nsSNV to be “disruptive”; otherwise we assign it as “non-disruptive”) and compare with experiment results from Platinum dataset to calculate the precision of the predictions, which gives a precision of 0.84. This independent test set from Platinum helps to validate the reliability of our method, and we do plan to update our dataset periodically in the future when experimentally measured data is more accessible. |

### Reviewer #2

### -- Ref2.0 – Positive comment --

|  |  |
| --- | --- |
| Reviewer  Comment | The authors developed an algorithm for prediction of the impact of nsSNVs on protein-drug interactions. They used machine learning techniques to tackle a very interesting problem relevant to human genetic variability and precision medicine. I like the overall approach taken, |
| Author  Response | We thank the reviewer for the thorough examination of our manuscript. We have provided additional analyses and updated the website to address the reviewer’s comments. |

### -- Ref2.1 – Training, test and validation set construction --

|  |  |
| --- | --- |
| Reviewer  Comment | The training scheme as shown in Supplementary Figure 8 could potentially overestimate the performance because model selection can be considered as part of hyperparameter optimization, which should not be conducted on the test set. It would be ideal to divide the whole labeled dataset into a training set, a validation set, and a test set, using the training set for parameter optimization, the validation set for model selection, and the test set for evaluation of the performance of the selected classifier. |
| Author  Response | We agree with the review that, ideally, the raw dataset should be divided into three parts: (i) training set, (ii) validation set and (iii) test set. According to the textbook by Hastie et al (Hastie et al., 2016), for a machine learning analysis in a data-rich situation, 50% of the original dataset might be for training, and 25% each for validation and testing.  However, when the original dataset is not large enough, it will be better to split the it into two parts, i.e. training set and test set, followed by “training + cross-validation strategy” model assessment and selection. Practically, many machine learning applications/projects do not contain a “validation set” due to the limited data size of the input dataset.  In our analysis, we have only 735 “disruptive” entries. To prevent bias from unbalanced dataset, we randomly selected equal number of samples from the 9458 “non-disruptive” entries. Thus, the input dataset for the model training is not large. Here, we separated it into training set and test set, and applied cross-validation strategy to select the best model. |

### -- Ref2.2 –Allele frequency cut-off–

|  |  |
| --- | --- |
| Reviewer  Comment | What allele frequency cutoff was used for distinguishing rare and common ExAC variants? How were driver nsSNVs distinguished from passenger nsSNVs in the TCGA data? This information cannot be found in the Methods section as the authors claimed. |
| Author  Response | We thank the reviewer for the comment. The frequency cutoff for distinguishing rare and common ExAC variants was set to be 1. We have modified the Methods section to indicate our cutoff criteria. |

### -- Ref2.3 – Side chain polarity change metric --

|  |  |
| --- | --- |
| Reviewer  Comment | The polarity index of amino acids seems to characterize a mixture of charge and hydrophobicity. Since this feature does not seem to give a strong signal (Fig. 3b), it might be better to use an established amino acid hydrophobicity metric. |
| Author  Response | We appreciate the reviewer’s comments on using an established amino acid hydropathy metric to describing side chain polarity changes. We have assigned each of the amino acid in our database with the hydropathy metric by Kyte & Doolittle (Kyte and Doolittle, 1982). Then we re-processed our analysis. Accordingly, the side chain hydropathy change boxplot in Figure 3b; the ROC curve in Figure 4; case study analysis in Figure 5, and corresponded method section has been modified. |

### -- Ref2.4 – Top disruptive SNV candidates prediction precision --

|  |  |
| --- | --- |
| Reviewer  Comment | In an unbalanced setting, a high auROC does not correlate well with the precision of top X predictions, which is a more useful metric since people are more interested in picking out nsSNVs that do disrupt binding among a large number of variants. What is the precision of say, top 50 or 100, predictions for the classifiers presented? |
| Author  Response | We thank the reviewer for pointing out the issue. We have calculated the precision of the top 100 predictions for the Random Forest classifier. The precision score is 0.92. The calculation process for precision score is stated below:   1. We first apply the trained Random Forest model to predict the binding affinity change (probability based) of the test data (30%), and we got the probability of a positive binding affinity change for each SNV. 2. The cut-off between a “disruptive SNV” and a “non-disruptive SNV” was set to be 0.5. If the predicted probability of positive binding affinity change was equal or greater than 0.5, we assigned this SNV candidate a “disruptive” one (positive class, tagged as “1”); If the predicted probability was smaller than 0.5, we assigned this SNV candidate a “non-disruptive” one (non-positive class, tagged as “0”). 3. We next sorted the probability in descending order and filtered out the top 100 SNV entry with highest predicted probability of a positive binding affinity change. 4. We compared the actual binding affinity change class (disruptive or non-disruptive) with our predict class of those 100 SNV candidates, and we compute the precision score ( precision = TP/(TP+FP) ) and get 0.92. We have added the precision score result in corresponded Results part in our manuscript. |

### -- Ref2.5 – Validation set construction --

|  |  |
| --- | --- |
| Reviewer  Comment | The authors wrote that they prepared a "validation set for specific case studies" (page 20). How was this set constructed? |
| Author  Response | We are sorry for the confusion caused by ambiguous term usage. The “validation set” here does not mean the “validation step” during model selection in machine learning, but an “independent test samples” for evaluation purpose, i.e., the two case studies shown in Figure 5. The drug resistance activity of human EGFR with gefitinib under T790M mutation and of human FPPS with zoledronate under A112H mutation presented in the manuscript are well-studied and validated via ligand binding experiments and clinical trial tests on patients. We applied GenoDock on these two cases aiming to show the reliability of GenoDock predictions.  In the new version of our manuscript, we parsed the Platinum as an independent test set for performance evaluation, and this experimentally measured set serves for the “validation purpose” of our study. |

### -- Ref2.6 – The Platinum dataset --

|  |  |
| --- | --- |
| Reviewer  Comment | The dataset used for this study is based on molecular docking. However, a database of experimentally measured mutation effect on protein-ligand binding, named Platinum, is available (Pires, D.E.V., Blundell, T.L. and Ascher, D.B. Nucleic Acids Research. 2015). How well does GenoDock perform using Platinum as the benchmark set? |
| Author  Response | We thank the reviewer for providing us the information of Platinum dataset. We have parsed the dataset to make it an independent test set for performance benchmark. After removing the entries that are not associated with a human protein, 87 entries left and serve as our independent test set for benchmark purpose. Detailed preparation of this dataset can be seen at “Methods” section. We then applied GenoDock to this parsed Platinum dataset as an independent test set to evaluate the performance of GenoDock. The AUC is 0.62 (Supplementary Figure 12), which is an acceptable result given that the ligand binding measurements recorded in Platinum is highly “heterogeneous” from a variety of labs and experimental setups, and that our gold standard for binding affinity change is based on docking calculations. In addition, we use a cutoff of 0.5 (predicted probability of being a disruptive nsSNV is greater than 0.5, we assign this nsSNV to be “disruptive”; otherwise we assign it as “non-disruptive”) and compare with experiment results from Platinum dataset to calculate the precision of the predictions, which gives 0.84. This independent test set from Platinum helps to validate the reliability of our method, and we do plan to update our dataset periodically in the future when experimentally measured data is more accessible. |

### -- Ref2.7 – Minor comments --

|  |  |
| --- | --- |
| Reviewer  Comment | 1. In page 4, third to last line, "a positive shift in binding affinity" would seem to indicate an increase in binding affinity, which is the opposite of what the authors mean.  2. In page 9, the last paragraph, two-sample Wilcoxon test is not a test of significant difference of the means of two samples. Rather, it is a test the compares the medians without assuming any underlying distribution.  3. Feature names in Fig. 3 and the main text should be kept consistent. Also, the colors in the legend of Fig. 3 are a little bit different from the colors used in the plots.  4. Supplementary figure captions S2, second to last line. Should it be "… and PDID are excluded from our database"? |
| Author  Response |  |

**References:**

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.

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*.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2016). Elements of Statistical Learning: data mining, inference, and prediction, 2nd edition edn (Springer).

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

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*, D387-391.

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.

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.

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.