JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

OPUS-Mut: Studying the Effect of Protein Mutation through Side-Chain Modeling

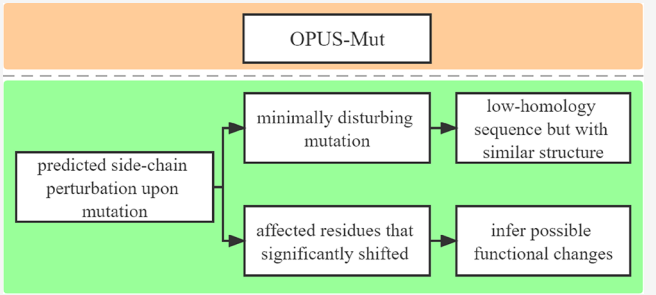
Gang Xu, Qinghua Wang, and Jianpeng Ma*

Cite This: J. Chem. Theory Comput. 2023, 19, 1629−1640

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Predicting the effect of protein mutation is crucial in many applications such as protein design, protein evolution, and genetic disease analysis. Structurally, mutation is basically the replacement of the side chain of a particular residue. Therefore, accurate side-chain modeling is useful in studying the effect of mutation. Here, we propose a computational method, namely, OPUS-Mut, which significantly outperforms other backbone-dependent side-chain modeling methods including our previous method OPUS-Rota4. We evaluate OPUS-Mut by four case studies on Myoglobin, p53, HIV-1 protease, and T4 lysozyme. The results show that the predicted structures of side chains of different mutants are consistent well with their experimentally determined results. In addition, when the residues with significant structural shifts upon the mutation are considered, it is found that the extent of the predicted structural shift of these affected residues can be correlated reasonably well with the functional changes of the mutant measured by experiments. OPUS-Mut can also help one to identify the harmful and benign mutations and thus may guide the construction of a protein with relatively low sequence homology but with a similar structure.



INTRODUCTION

Protein amino-acid mutations play a key role in protein engineering and protein evolution[1] and are also the root causes of many genetic diseases.[2] Therefore, it is crucial to predict the change of mutant properties against their original unmutated counterparts (wild type). In recent years, with the development of deep learning techniques, many successful studies have been performed to predict the change of different properties for mutation, such as thermodynamic stability,[3,4] enzyme activity,[5] and loss or gain of function.[6,7] However, limitations still exist. For example, for protein stability prediction, due to the sequence similarity between proteins used in training and test datasets, the methods tend to overestimate the prediction performance.[8]

From a structural point of view, a mutation is basically the replacement of the side chain of the corresponding residue. In this case, we study the effect of protein mutation through side-chain modeling. By comparing the differences between the predicted unmutated (wild-type) side chains and the predicted mutant side chains, we can infer the extent of structural perturbation and the affected residues whose predicted side chains are significantly shifted due to the mutation. To quantify the extent of structural perturbation, we use the summation of the differences of all predicted side-chain dihedral angles (from χ_1 to χ_4) over all residues between the wild type and mutant. In this paper, for the purpose of clarity, we name this summation of differences over all residues as S_{diff}. It is assumed that larger S_{diff} corresponds to the predicted more severe structural perturbation, which means the mutation is more likely to be harmful. From the affected residues, we can correlate them with the functional changes. In protein engineering, for example, with the possibility of inferring the affected residues for a particular mutation, we may selectively avoid the type of mutations that can potentially impair the functions related to certain residues.

There are some advantages in studying the effect of mutation through a general side-chain modeling method, in oppose to computational methods using the experimentally obtained mutational−functional change data. They are as follows: (1) it has no system dependence and can be generalized to any target without requiring further system-specific data for fine-tuning; (2) it focuses on the side-chain modeling ability and is not limited by the shortage of the experimental data of mutant properties; (3) it is straightforward to interpret since it is based on the predicted structural shift.

In recent years, many protein structure prediction methods have been proposed.[9−12] Among them, AlphaFold2[11] is the

Received: August 18, 2022
Published: February 22, 2023

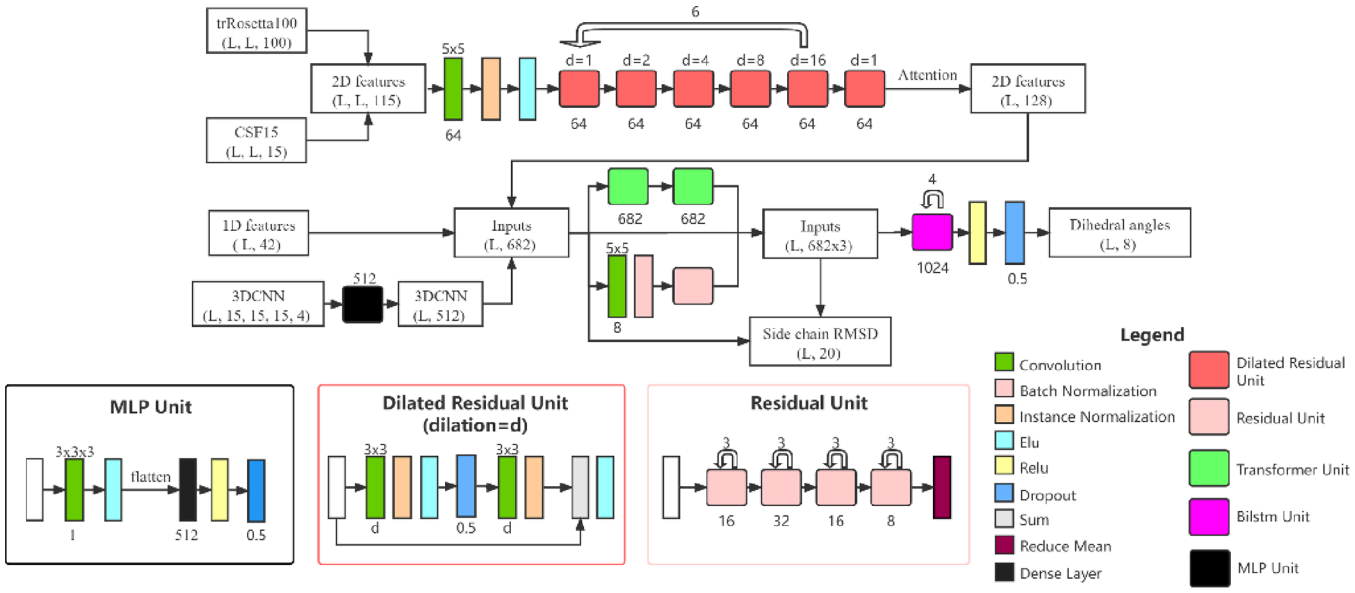


Figure 1. The framework of OPUS-Mut. The predicted four dihedral angles contain eight regression nodes ($\sin(\chi)$ and $\cos(\chi)$ for each of them). Meanwhile, OPUS-Mut also outputs the predicted root mean square deviation (pRMSD) for its predicted side chain on each residue.

best and, in some cases, it delivers the predictions close to the experimental results. Some studies show that AlphaFold2 may not be suitable for the protein mutation task, such as analyzing the stability of the mutant.[13] In our study, we find that the AlphaFold2-predicted backbones for the wild-type sequence and the single-site-mutation sequence are usually identical, maybe because they all depend on the results of multiple-sequence alignment, which are similar to each other when the query sequences are almost the same. Therefore, studying the effect of point mutation through predicted backbone may be not feasible. Since the conformations of side chains are more sensitive to the change of local environment, it is more sensible to study the point mutation from the side chain point of view.

For backbone-dependent side-chain modeling, many successful methods have been developed. For sampling-based methods,[14−17] they sample the rotamers from the rotamer library and use their scoring functions to determine the best rotamer for each residue with the minimal score. However, this kind of methods is limited by the discrete rotamers in the rotamer library and the accuracy of the scoring function. Recently, some deep learning-based methods have been proposed,[18,19] which improve the accuracy of side-chain modeling by a large margin.

In this research, we propose a computational method, named OPUS-Mut, which is mainly based on our previous work OPUS-Rota4,[19] but with some improvements. In OPUS-Mut, we modify our loss function for the predicted dihedral angles according to that in AlphaFold2.[11] To increase the sensitivity of OPUS-Mut toward the point mutation, we add an additional neural network node to predict the root mean square deviation (RMSD) of the predicted side chain against its native counterpart for each residue. Therefore, OPUS-Mut can evaluate the correctness of its prediction at each site. In our study, OPUS-Mut outperforms OPUS-Rota4 and other backbone-dependent side-chain modeling methods on three native backbone datasets. The excellent side-chain modeling ability of OPUS-Mut is the foundation of this research.

For further evaluation, four case studies (Myoglobin, p53, HIV-1 protease, and T4 lysozyme) have been conducted, from either structural or functional perspective. From a structural perspective, we adopt the mutants that have experimentally determined structures and focus on the accuracy of the predicted side chains compared to their experimental counterparts. From a functional perspective, we focus on the extent of structural perturbation and the affected residues whose predicted side chains are significantly shifted due to the mutation, and the correlation between them and the functional changes. The results show that OPUS-Mut can be used in the following cases: (1) it can be used to infer the functional changes of the mutation from a structural perspective; (2) it can be used to infer the affected residues due to the mutation, therefore avoiding the unwanted effect on specific sites in the case of protein engineering; (3) it can help us to identify the harmful and benign mutations. As one of the applications, it may guide us to construct a relatively low-homology mutant sequence but with a similar structure, which thus may be helpful in studying protein engineering and protein evolution.

MATERIALS AND METHODS

Framework of OPUS-Mut. OPUS-Mut is mainly based on our previous work OPUS-Rota4,[19] but with some modifications. In calculation, only the protein residues are taken into consideration; other ligands are omitted. The neural network architecture of OPUS-Mut is shown in Figure 1. trRosetta100 is a 100-d feature, which is used to describe backbone distance (C_β−C_β) and orientation (ω, θ_{ab}, θ_{ba}, φ_{ab}, φ_{ba}) contact information.[10] CSF15 refers to the relative CSF position[20] of the backbone atoms of a specific residue from its contact counterpart.[19] The 2D features go through 32 dilated residual−convolutional blocks and an attention module and output a 128-d vector for each residue. The 1D features are derived from protein backbone conformation. Besides the 41 features used in OPUS-Rota4,[19] an additional feature derived from DLPacker[18] is introduced to denote the side-chain RMSD of prediction from its most similar conformation in the DLPacker's library. 3DCNN is the side-chain density map for each residue, which uses the input from the DLPacker (OPUS) in OPUS-Rota4; it goes through an MPL unit and outputs a

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

512-d vector. The three outputs are combined and go through three modules, which are identical to that in OPUS-TASS2:[9] modified Transformer module,[21] Resnet module,[22] and bidirectional Long-Short-Term-Memory module.[23] All strides in the residual units are set to be one, and the batch size of OPUS-Mut is also set to be one.

In OPUS-Mut, we introduce the auxiliary loss $L_{\text{anglenorm}}$ from AlphaFold2[11] to keep the predicted points close to the unit circle. For a dihedral angle χ, we output two predictions (denoted as o_1 and o_2) for $\sin(\chi)$ and $\cos(\chi)$, respectively. The predicted dihedral angle is recovered from the outputs of the network by $\tan^{-1}(o_1/o_2)$. The loss function is listed as the following:

$$\begin{cases} L = L_{\text{torsion}} + 0.02L_{\text{anglenorm}} \\ L_{\text{torsion}} = (o_1 - \sin(\chi)_{\text{real}})^2 + (o_2 - \cos(\chi)_{\text{real}})^2 \\ L_{\text{anglenorm}} = abs({o_1}^2 + {o_2}^2 - 1) \end{cases}$$

In addition, we add an additional node to predict the RMSD (Å) of the predicted side chain against its native counterpart for each residue. The RMSD is calculated using the side-chain coordinates reconstructed from the predicted dihedral angles. During the training process, we find that the residue-wise side-chain RMSD ranges from 0 to 1 Å for nearly all residues. Therefore, the range of RMSD is set to be 0−1 and then segmented into 20 bins. The cross-entropy loss is used for training.

OPUS-Mut is implemented in TensorFlow v2.4[24] and trained on one NVIDIA Tesla V100. The Glorot uniform initializer and the Adam optimizer[25] are adopted. The initial learning rate is $1e-3$, and it will be reduced by half when the accuracy of the validation set is decreased. The training process will stop after being reduced by four times. We train five models, and the median of their outputs is used to make the final prediction.

Datasets. In OPUS-Mut, the dataset used for training is the same as that in OPUS-Rota4,[19] which contains 10,024 proteins in the training set and 983 proteins in the validation set, culled from the PISCES server on February 2017.[26] Note that none of the original structures and their related mutants mentioned in the four case studies of this paper are present in the dataset that is used to train OPUS-Mut.

For evaluation, we use three native backbone test sets: CAMEO60, collected by OPUS-Rota3,[14] contains 60 hard targets released between January 2020 and July 2020 from the CAMEO website;[27] CASP14, collected by OPUS-X,[9] contains 15 FM targets downloaded from the CASP website (http://predictioncenter.org); and CAMEO65, collected in this study, contains 65 hard targets released between May 2021 and October 2021 from the CAMEO website.

Performance Metrics. To measure the accuracy of side-chain modeling methods, we use the percentage of correct prediction with a tolerance criterion 20° for all side-chain dihedral angles (from χ_1 to χ_4). The residue with more than 20 residues, between which the C_β−C_β distance (C_α for Gly) is within 10 Å, is defined as core residue.

RESULTS

Side-Chain Modeling Performance of OPUS-Mut. For evaluating a side-chain modeling method, we remove all side chains of a target at first then use the side-chain method to repack its side chains. Finally, we measure the differences between the predicted side chains and their experimentally determined counterparts either on all residues or on core residues. In OPUS-Rota4, we use OPUS-Fold2 to optimize the initially predicted dihedral angles from OPUS-RotaNN2 against the predicted side-chain contact map from OPUS-RotaCM to achieve better and final results.[19] In OPUS-Mut, we update OPUS-Rota4 by discarding the time-consuming optimization procedure and modifying the loss function inspired by AlphaFold2.[11] Specifically, we adopt the loss function for angle prediction in AlphaFold2 and introduce an additional node to predict the root mean square deviation (pRMSD) of the predicted side chain against its native counterpart for each residue. The improvement of each modification is shown in the SI Appendix, Figure S1. The effectiveness of using the loss function from AlphaFold2 on angle prediction may bring some insights to the relevant tasks.

We compare OPUS-Mut with two sampling-based methods SCWRL4[16] and OSCAR-star[17] and two deep learning-based methods DLPacker[18] and OPUS-Rota4,[19] on three native backbone test sets CAMEO60, CASP14, and CAMEO65. In terms of the percentage of correct prediction with a tolerance criterion of 20° for all side-chain dihedral angles (from χ_1 to χ_4), OPUS-Mut outperforms other methods either measured by all residues (Figure 2) or measured by core residues only

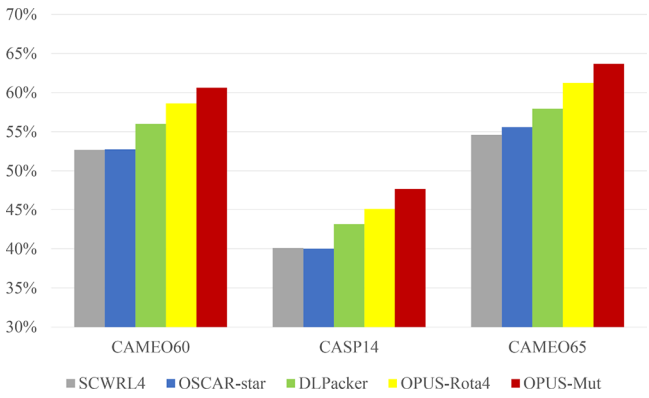


Figure 2. Percentage of correct prediction with a tolerance criterion 20° for all side-chain dihedral angles (from χ_1 to χ_4) of different methods measured by all residues.

(SI Appendix, Figure S2). The percentage of correct prediction for each type of residue on CAMEO65 is also listed in the SI Appendix, Table S1 (all residues), and SI Appendix, Table S2 (core residues).

In addition, the results of the significance value P of the residue-wise percentage of correct prediction with a tolerance criterion 20° of each method compared with OPUS-Mut is listed in the SI Appendix, Table S3. RMSD is calculated residue-wise using the main-chain and side-chain heavy atoms that are involved in side-chain dihedral calculation. The results of RMSD (Å) and significance value P are shown in the SI Appendix, Table S4. Furthermore, the results of the percentage of correct prediction with a tolerance criterion from 5 to 20° for all side-chain dihedral angles (from χ_1 to χ_4) of different methods on CAMEO65 measured by all residues are shown in the SI Appendix, Table S5.

Case Study: Myoglobin. Myoglobin is an essential hemoprotein that regulates the cellular oxygen concentration in striated muscle. It can bind oxygen reversibly by its heme prosthetic group.[28]

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

From a Structural Perspective. A previous study showed that the substitutions (V68A, V68I, V68L, and V68F) at position 68 do not affect the secondary or tertiary structure of the protein significantly and the largest changes caused by these substitutions are the distortions of the interior spaces.[29] Meanwhile, the structures of these mutants (V68A (PDB: 1MLF), V68I (PDB: 1MLM), V68L (PDB: 1MLQ), and V68F (PDB: 1MLJ)) have been determined by Quillin et al.[29] Therefore, examining the side-chain modeling results for these four substitutions is an ideal way to verify the sensitivity of OPUS-Mut toward the single-site mutation.

Following the research from Nienhaus et al.,[30] we download the structure of the wild-type (PDB: 2MGK) sperm whale MbCO.[31] Then, we replace Val at position 68 by Ala, Ile, Leu, and Phe and reconstruct their side chains with OPUS-Mut.

As shown in Figure 3, the side-chain modeling results for each mutant are consistent well with their X-ray crystallographic results.[29] From the side-chain prediction results of mutant V68F (Figure 3f), we can infer that the benzyl side chain that partially fills the Xe4 cavity (region C in Figure 3a) may become a steric barrier to the ligand association. This assumption has been confirmed through experimental approaches.[29,30]

Case Study: p53. The tumor suppressor p53 plays an important role in maintaining the genomic integrity of the cell, and it is mutated in half of all human cancers.[2] In the last few decades, many studies have studied the different properties of its mutations.[32−39]

From a Structural Perspective. Fersht's group constructed *T*-p53C, a superstable quadruple mutant (M133L/V203A/N239Y/N268D) of the human p53 core domain,[38] and determined its structure (PDB: 1UOL).[37] Furthermore, based on *T*-p53C, the experimental structure of a series of mutants were determined,[35,36] including R273C (PDB: 2J20), R273H (PDB: 2BIM), F270L (PDB: 2J1Z), V143A (PDB: 2J1W), Y220C (PDB: 2J1X), R282W (PDB: 2J21), and triple mutant T123A/H168R/R249S (PDB: 2BIQ).

In this study, we download the *T*-p53C (PDB: 1UOL chain A). Then, we substitute the residues according to the corresponding mutations mentioned above and reconstruct their side chains with OPUS-Mut.

As shown in Figure 4 and the SI Appendix, Figure S3, our side-chain modeling results for each mutant are very close to their experimental counterparts, especially for the mutations inside the protein: R273C (Figure 4a), R273H (Figure 4b), F270L (Figure 4c), and V143A (Figure 4d). For the mutation on the loop area Y220C (SI Appendix, Figure S3b), the prediction is also accurate and the biases shown in the figure are mainly caused by the backbone alignment, which means the main chains of the loop area in the original structure (PDB: 1UOL chain A) we used for side-chain modeling are not strictly identical to that in the mutant structure (PDB: 2J1X). For R282W (SI Appendix, Figure S3c), the side-chain predictions for its nearby residues that are close to the deoxyribonucleic acid are less accurate than the others, because OPUS-Mut does not take the effect of deoxyribonucleic acid into account. The results for T123A/H168R/R249S (SI Appendix, Figure S3d) indicate the good performance of OPUS-Mut in modeling the side chains of multiple mutants.

From a Functional Perspective. In this section, we focus on three Zn^{2+} region mutations (R175H, M237I, and C242S) and a DNA region mutation (R282W).[2] We download the p53−DNA complex (PDB: 1TUP chain B)[40] and substitute the residues according to the mutations mentioned above. Then, we reconstruct their side chains with OPUS-Mut. Note that, according to Bullock et al., zinc-binding site consists of four residues: C176, H179, C238, and C242.[2]

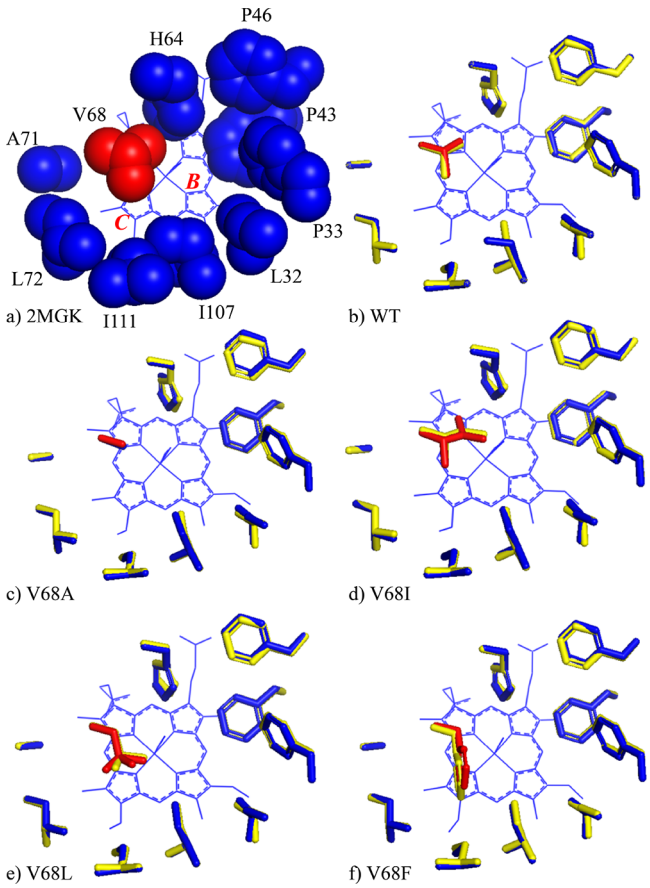


Figure 3. Side-chain modeling results of wild-type MbCO and its mutants. (a) The local environment around V68 in the wild-type MbCO crystal structure (PDB: 2MGK). The Xe4 pocket consists of regions B and C. (b) The side-chain structures in the wild-type structure (PDB: 2MGK) and the side-chains predicted by OPUS-Mut. (c−f) The crystal and predicted side-chain structures for each mutant: (c) the V68A mutant (PDB: 1MLF); (d) the V68I mutant (PDB: 1MLM); (e) the V68L mutant (PDB: 1MLQ); and (f) the V68F mutant (PDB: 1MLJ). The crystal side-chain structure at the mutation site is marked in red, the crystal side-chain structures of the nearby residues are marked in blue, and the predicted side chains are marked in yellow. The results indicate that the predicted mutant side chains (yellow) are consistent with their crystal structures (red and blue). Note that in the side-chain modeling calculation, the heme group is not included. However, for the purpose of illustration, the heme group is shown in each panel.

As shown in Figure 5, all the affected residues are shown and marked in green. Apart from the affected residues, the side chains of other residues in the protein remain relatively unshifted. Figure 5a−c shows that the residues at the zinc-binding site (C238 and C242 in Figure 5a, C242 in Figure 5b, and H179 in Figure 5c) are the affected residues in each mutant, which indicate that the Zn^{2+} region may be modified in these three mutants.

For R282W (Figure 5d), R283 and H115 are its affected residues. According to Rippin et al., R283 belongs to a DNA-binding determinant.[39] We assume that this may be a clue for R282W's binding affinity loss. The predicted side-chain

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

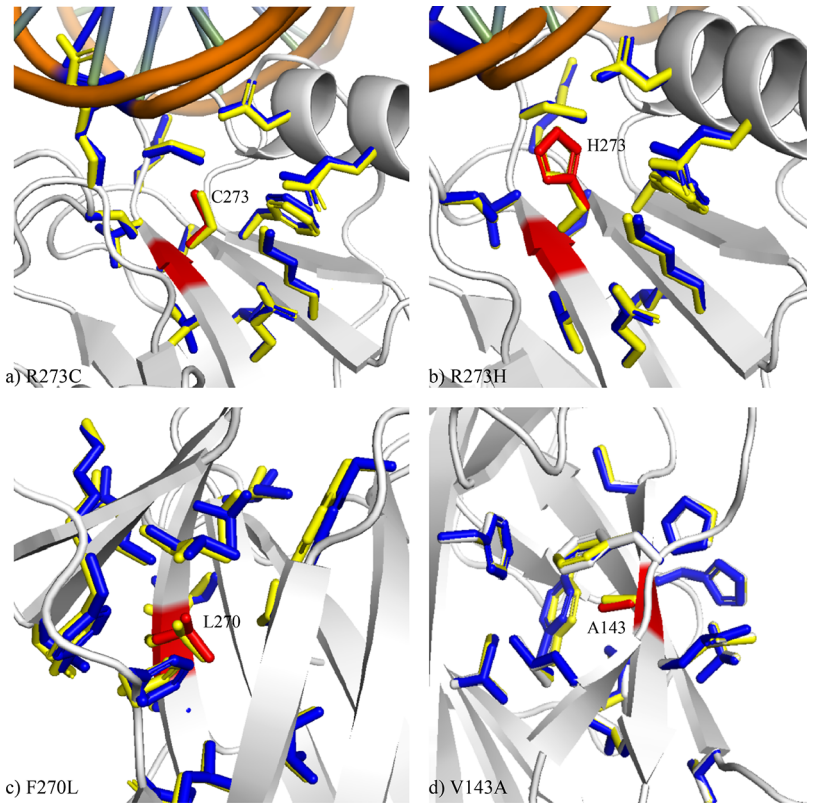


Figure 4. Side-chain modeling results of *T*-p53C mutants. (a) The R273C mutant (PDB: 2J20); (b) the R273H mutant (PDB: 2BIM); (c) the F270L mutant (PDB: 2J1Z); and (d) the V143A mutant (PDB: 2J1W). In all panels, the side chains from the crystal structures are marked in red for the mutation sites, and in blue for neighboring residues within 5 Å from the mutation sites. The side chains predicted by OPUS-Mut are marked in yellow.

conformation changes at H115 may be a prediction bias since it is far away from W282.

Case Study: HIV-1 Protease. HIV-1 protease is an important anti-AIDS drug target,[41] and many studies have been conducted to investigate its different properties.[41−48]

From a Structural Perspective. Ala et al. determined the structure of the mutant HIV-1 proteases (V82F, I84V) complexed with cyclic urea inhibitors.[43] Since the backbone structure of HIV-1 protease may be shifted along with different inhibitors, in this case study, instead of using the wild-type backbone structure and replacing the mutation residues at the corresponding sites, we download the mutant backbone structure (PDB: 1MEU chain A) directly and reconstruct its side chains.

As shown in Figure 6, for mutation residues F82 and V84, the side-chain modeling results are very close to the experimental results. The predictions for their nearby residues (<5 Å) are also accurate, except the predictions for the surface residues R8 and E21.

From a Functional Perspective. Large-scale mutagenesis of HIV-1 protease has been performed by Loeb et al.[48] In these studies, 336 single missense mutations and their corresponding phenotypes were measured. The phenotypes were categorized into three groups based on their ability to process the Pol precursor in *Escherichia coli*: negative (no mature processed products), intermediate (some mature processed products), and positive (processed similarly to the wild-type). In this case study, we collect their results to form an HIV-1 protease mutagenesis dataset that contains 319 results (because some results are hard to distinguish from their paper, especially for uppercase *I* and lowercase *l*). The dataset we collected contains 153 negative phenotype data, 66 intermediate phenotype data, and 100 positive phenotype data (SI Appendix, Table S6).

Following the research from Masso and Vaisman,[5] we download the structure of the wild-type HIV-1 protease with 99 residues in length (PDB: 3PHV).[49] We predict the side chains for all 99 × 19 = 1881 possible single-site mutations. For each mutation, we sum up the differences of all side-chain dihedral angles (from χ_1 to χ_4) between the predicted wild-type side chains and the predicted mutant side chains (here, the difference from the mutated residue itself is excluded) and use it as an indicator for the extent of structural perturbation due to the mutation. Here, besides the summation of differences over all residues S_{diff}, we use $S_{diff\_critical}$ to denote the summation of differences on some critical residues (such as the residues in active sites).

In all 1881 possible single-site mutations, only 319 of them have experimental results (the data in the HIV-1 protease mutagenesis dataset); therefore, we use them to do the statistical analysis. As shown in Figure 7a, the mean of S_{diff} for the mutations in the negative (0), intermediate (1), and positive (2) phenotype groups is 28.3, 23.0, and 21.7, respectively. The median of them is 27.0, 20.5, and 17.0, respectively. Both mean and median of these three groups show that more severely predicted structural perturbation corresponds to a higher probability of loss of function. Among 319 mutations, 7 out of the top 10 mutations with the highest S_{diff} belong to the negative phenotype group, 2 out of the top 10 mutations belong to the intermediate phenotype group, and

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

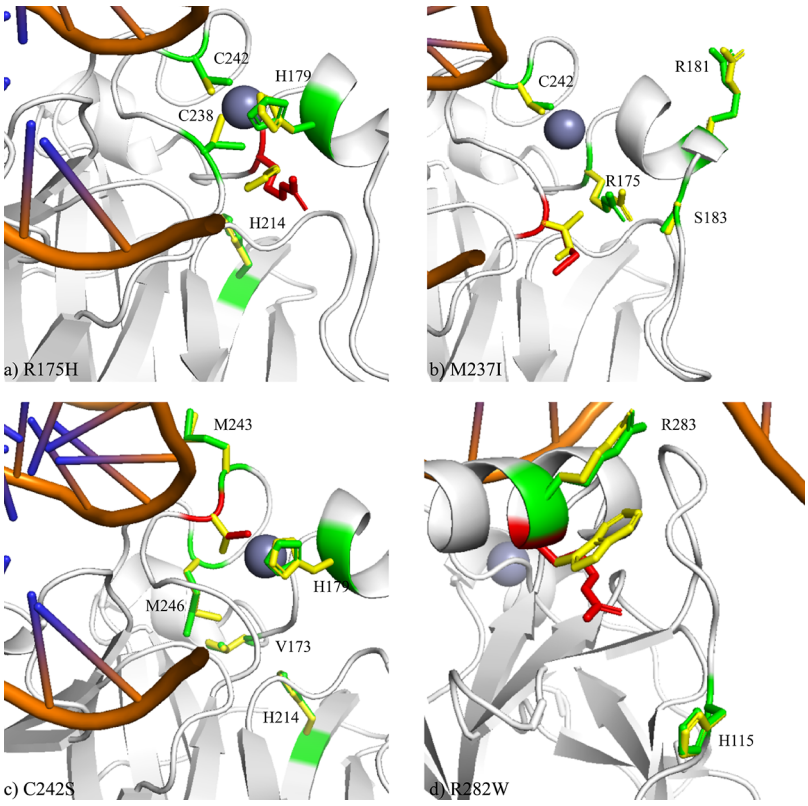


Figure 5. Side-chain modeling results of p53−DNA complex mutations. (a) The R175H mutant; (b) the M237I mutant; (c) the C242S mutant; and (d) the R282W mutant. The mutation site is marked in red, and all the affected residues are shown and marked in green. The affected residue refers to the residue whose mean absolute error of all predicted side-chain dihedral angles between the wild type and mutant is greater than 5°. The predicted side chains of the wild-type (PDB: 1TUP) are marked in green (red at the mutation site), and the predicted side chains of the mutants are marked in yellow.

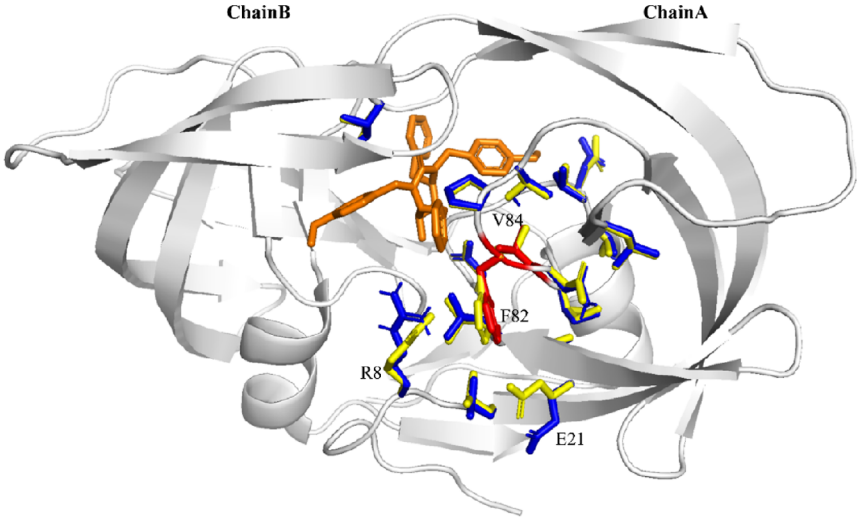


Figure 6. Side-chain modeling results of HIV-1 proteases complexed with cyclic urea inhibitors. The side chains from the crystal structure (PDB: 1MEU) are marked in red for the mutation sites (V82F, I84V) and in blue for neighboring residues within 5 Å from the mutation sites. The cyclic urea inhibitor is shown in brown. The side chains predicted by OPUS-Mut are marked in yellow.

only 1 out of the top 10 mutations belongs to the positive phenotype group.

According to Mager,[42] D25, T26, and G27 are the key residues involved in the active site of HIV-1 protease. To find out the impact of each mutant on the active site, we sum up the differences ($S_{diff\_critical}$) on these three residues exclusively (Figure 7b). Their mean is 1.33, 0.68, and 0.47, respectively. Their median is 0.42, 0.21, and 0.18, respectively. Compared to the three groups obtained by summing up the differences (S_{diff}) over all residues, the groups obtained by summing up the differences ($S_{diff\_critical}$) on three active site residues are more distinguishable. Among 319 mutations, 7 out of the top 10 mutations with the highest $S_{diff\_critical}$ belong to the negative phenotype group, 2 out of the top 10 mutations belong to the

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

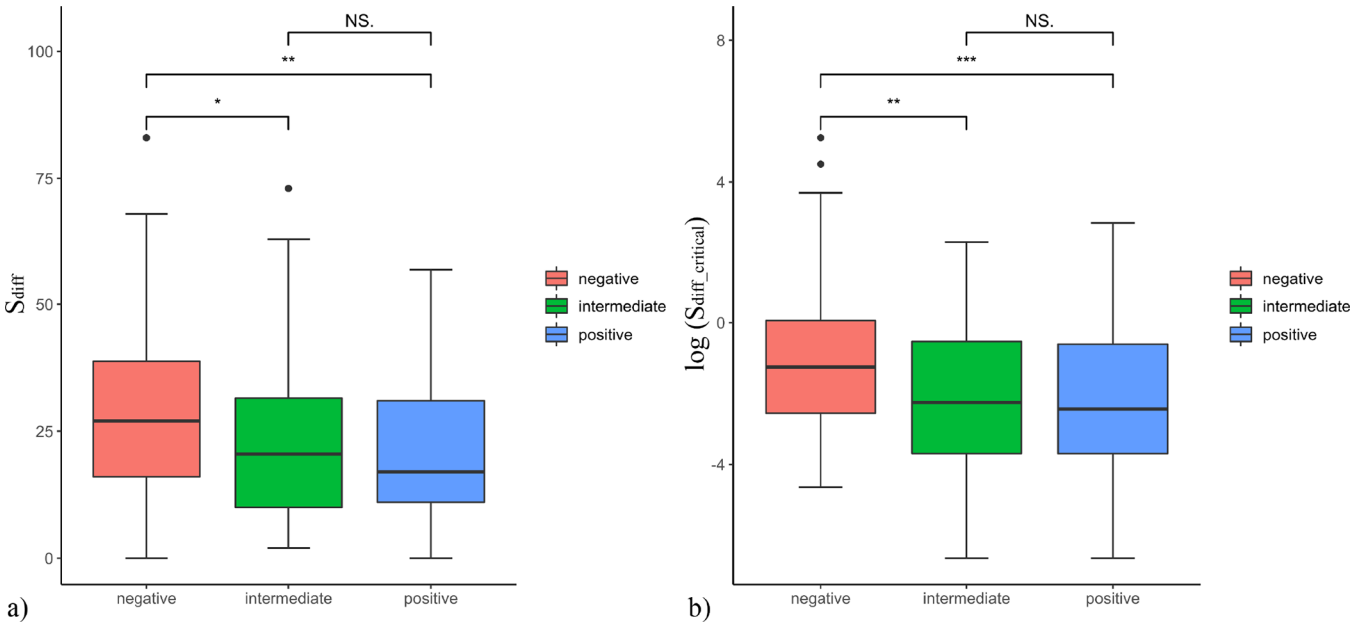


Figure 7. The summation of differences of all side-chain dihedral angles (from χ_1 to χ_4), either summed over all residues (S_{diff}) or summed over three active site residues ($S_{diff\_critical}$), between the predicted wild-type side chains and the predicted mutant side chains on the HIV-1 protease mutagenesis dataset. In the box plot, the line inside the box represents the median of each group. The whiskers extended from the box indicate the variability outside the upper and lower quartiles. Outliers that differ significantly from the rest of the dataset are plotted beyond the whiskers. Mann−Whitney *U*-test is used to analyze the difference between each group: "NS." denotes $P > 0.05$, "*" denotes $P < 0.05$, "**" denotes $P < 0.01$, and "***" denotes $P < 0.001$. The data from three groups (negative (0), intermediate (1), and positive (2) phenotype group) are shown. The mutations on the first and last residues (P1 and F99) are ignored because the side-chain modeling may have a bias on the end of the sequence. (a) shows the differences (S_{diff}) summed over all residues. To avoid the influence of outliers, if the difference between the two residues is smaller than 1°, we set it to be 0; if it is smaller than 5, we set it to be 1; if it is smaller than 10, we set it to be 2; if it is smaller than 20, we set it to be 3; if it is larger than 20, we set it to be 4. (b) shows the differences ($\log(S_{diff\_critical})$) summed over three active site residues (D25, T26, and G27).

intermediate phenotype group, and only 1 out of the top 10 mutations belong to the positive phenotype group.

In Table 1, we list the top 10 mutations with the highest $S_{diff\_critical}$ on three active site residues (D25, T26, and G27) in all 1881 possible single-site mutations. The mutations on residues 24−28 are ignored since they are too close to the active site. Among these predicted top 10 mutations, V82I[47] and V82E[50] have been confirmed to be able to reduce the inhibitor affinity. In addition, V82, I84, and L90 are the hot spots of the drug resistance mutation.[51] Therefore, OPUS-Mut would be a useful tool in finding the possible harmful mutations based on their influences on the protein's critical region.

Table 1. The Top 10 HIV-1 Protease Mutations with the Highest $S_{diff\_critical}$ on Three Active Site Residues (D25, T26, and G27) in all 1881 Possible Single-Site Mutations

mutant	$S_{diff\_critical}$
P9Y	7.32
V82I	7.17
V82G	7.10
I84N	6.76
L90R	6.63
P9F	6.31
I84D	5.69
L90Y	5.68
A22M	5.22
V82E	5.01

For each of 1881 possible single-site mutations, we sum up the differences (S_{diff}) of all side-chain dihedral angles between the predicted wild-type side chains and the predicted mutant side chains over all residues (with the differences on three active site residues multiplied by a factor of 10 in order to emphasize the significance of the active site residues). Then, for each residue site, among its 19 possible mutations, we keep the one with the lowest S_{diff}. Therefore, we get 99 sequences, each with a single-site mutation, that have the smallest structural perturbation. We sort 99 sequences according the value of their S_{diff}. For a desired mutation rate, for example, 10% in type 1, Table 2, we combine the top 10 sequences with the lowest S_{diff} to construct a sequence with 10% mutation rate. For comparison, another type of multiple-mutation sequence is constructed reversely, by combining the sequences with the highest S_{diff} in the procedure mentioned above (type 2, Table 2). We use AlphaFold2[11] without template to predict the

Table 2. The TM-Score of Each HIV-1 Protease (with 99 Residues in Length) Multiple-Mutation Sequence Predicted by AlphaFold2[a]

	wild type	10%	20%	30%	40%	50%
type 1	0.93	0.91	0.90	0.86	0.85	0.85
type 2		0.83	0.82	0.82	0.62	0.50

[a]"Type 1" represents the TM-score of the sequences with 10, 20, 30, 40, and 50% of mutation rate, constructed by preferably using the minimally disturbing mutations with the lowest S_{diff}. "Type 2" represents the TM-score of the multiple-mutation sequences constructed by preferably using the maximally disturbing mutations with the highest S_{diff}.

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

structure for the sequences with 10, 20, 30, 40, and 50% mutation rates. As shown in Table 2, by preferably using the minimally disturbing mutations (with the lowest S_{diff}), the constructed multiple-mutation sequences (type 1, Table 2) tend to cause smaller structural perturbation than those (type 2, Table 2) constructed by preferably using the maximally disturbing mutations (with the highest S_{diff}). For example, the TM-score of the type 1 sequence with 50% mutation rate is 0.85, which is still reasonably high (note that the TM-score of the wild-type structure prediction is 0.93). The details of the structure and sequence of this multiple mutant are shown in the SI Appendix, Figure S4.

Case Study: T4 Lysozyme. T4 lysozyme is a monomeric protein, which contains 164 residues that hydrolyzes peptidoglycan.[52] Many studies have been performed to investigate the different properties of its mutations, especially by Matthews's group.[53−58]

In this case study, we download the structure of the wild-type T4 lysozyme (PDB: 2LZM)[53] and predict the side chains for all 164 × 19 = 3116 possible single-site mutations. According to the PDB file of 2LZM,[53] E11 and D20 are the active site residues, and L32, F104, S117, and N132 are the binding site residues.

From a Functional Perspective (Protein Stability). Masso et al.[59] collected the experimental stability changes of 293 T4 lysozyme mutants at physiological pH from ProTherm.[60] From the figure of the mutational array in their paper,[59] we identify 292 of them, in which 80 of them will increase the stability and 212 of them will decrease the stability (SI Appendix, Table S7).

As shown in SI Appendix, Figure S5, the mean of S_{diff} for the mutations in the decreased stability group and increased stability group is 13.2 and 10.0, respectively. Their median is 10.0 and 7.0, respectively. Among 292 mutations, 9 out of the top 10 with the highest S_{diff} belong to the decreased stability group. The results are consistent with our assumption that more severely predicted structural perturbation corresponds to a higher probability of stability decrease.

Similar to the study on HIV-1 protease, for all 3116 possible single-site mutations, we sum up the differences (S_{diff}) of all side-chain dihedral angles between the predicted wild-type side chains and the predicted mutant side chains on all residues, with the differences on two active site residues and four binding site residues multiplied by a factor of 10. Two types of sequences with 10, 20, 30, 40, and 50% mutation rates are constructed: (1) the first type of multiple-mutation sequences is constructed by preferably using the minimally disturbing mutations (with the lowest S_{diff}); (2) the second type of multiple-mutation sequences is constructed by preferably using the maximally disturbing mutations (with the highest S_{diff}). The results also indicate that the combination of the minimally disturbing mutations may lead to a multiple-mutation sequence with smaller structural perturbation (SI Appendix, Table S8).

From a Functional Perspective (Enzyme Activity). Masso and Vaisman[5] collected the experimental enzyme activities of 2015 T4 lysozyme mutants obtained by Rennell et al.[61] Among them, 1552 mutants are labeled as active, 263 mutants are labeled as inactive.

In all 3116 possible single-site mutations, the top 10 mutations with the highest $S_{diff\_critical}$ on two active site residues (E11 and D20) are the following: two of them have no experimental data, seven of them belong to the inactive group in the experimental enzyme activity data, and only one of them belongs to the active group. For further illustration, in Table 3, we list the top 20 mutations under the following rules: first, we exclude the mutations without experimental data; second, for a residue site with many mutations existing, we show up to three mutations (such as the mutations in Q105 in Table 3). The results show that 15 out of the top 20 mutations belong to the inactive group.

Table 3. The Selected Top 20 T4 Lysozyme Mutations with the Highest $S_{diff\_critical}$ on Two Active Site Residues in all 3116 Possible Single-Site Mutations[a]

rank	mutant	$S_{diff\_critical}$	activity
2	T26Y	15.68	N
3	T142E	15.6	N
4	T26R	15.08	N
5	V149E	14.76	N
7	T26K	14.63	N
8	L7F	14.54	P
9	Q105F	13.99	N
10	Q105Y	13.62	N
12	Q105L	12.31	P
19	L7H	11.63	P
22	L7Y	10.26	P
26	T142Y	9.67	N
27	G30L	9.62	N
30	G30K	8.69	N
32	G30Q	8.17	N
34	L32F	7.51	P
37	T142K	7.38	N
38	F104E	7.35	N
39	F104Q	7.32	N
44	F104S	7.17	N

[a]"Rank" refers to the real rank before selection. "Activity" refers to the experimental enzyme activity, "N" denotes inactive, "P" denotes active.

CONCLUDING DISCUSSION

Modeling protein mutation is crucial for protein design, protein evolution, and genetic disease analysis. The foundation for studying mutation is the ability to accurately model the side-chain conformations. In this research, we study the effect of protein mutation through the extent of predicted structural perturbation upon the mutation, specifically, through the extent of shift of the predicted side chains. To this end, we propose a computational method, namely, OPUS-Mut. As shown in Figure 2, SI Appendix, Figure S1, SI Appendix, Figure S2, SI Appendix, Table S1, SI Appendix, Table S2, SI Appendix, Table S3, SI Appendix, Table S4, and SI Appendix, Table S5, OPUS-Mut significantly outperforms other backbone-dependent side-chain modeling methods including our previous method OPUS-Rota4,[19] either measured by all residues or measured by core residues exclusively. We evaluate the performance of OPUS-Mut on modeling mutation side chains. As shown in Figures 3, 4, and 6 and SI Appendix, Figure S3, by comparing to experimental results, OPUS-Mut is capable of delivering satisfactory results for mutations. In addition, the results also show that, with the help of existing experimentally determined backbone, OPUS-Mut can deliver better side-chain predictions than AlphaFold2, which only uses

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

multiple-sequence alignment as its input (SI Appendix, Table S9).

In addition, the predicted side chains of OPUS-Mut for neighboring residues within 5 Å from the ligand are shown in the SI Appendix, Figure S6. In the SI Appendix, Figure S7, we list the side-chain modeling results of OPUS-Mut on four HIV-1 protease crystal structures with various co-crystallized inhibitors. The results show that for most residues near the ligand, OPUS-Mut is able to deliver the correct predictions. However, there is still room for improvement. For example, the predicted side chains for the residues near the DNA binding site are not accurate enough (SI Appendix, Figure S6b). In the future, we will modify our framework so that the ligand information can be included in our model to further improve the accuracy of such cases.

OPUS-Mut can be used to infer the functional changes of the mutation. As exemplified in Figure 3f, from the side-chain prediction results of myoglobin mutant V68F, we can infer that the benzyl side chain, which partially fills the Xe4 cavity, may become a steric barrier to the ligand association. In addition, as shown in Figure 5, by comparing the differences of all side-chain dihedral angles (from χ_1 to χ_4) between the predicted wild-type side chains and the predicted mutant side chains, we find that the affected residues, whose predicted side chains are significantly shifted due to the mutation, can correspond well with the functional changes of the mutant. For example, the affected residues of the p53 Zn^{2+} region mutations are located at the zinc-binding site (Figure 5a−c), which indicate that the Zn^{2+} region may be affected in these mutants. In this case, OPUS-Mut can also be used to infer the affected residues caused by the mutation, therefore avoiding the unwanted effect on specific sites in the case of protein engineering.

Although some mutations can lead to the perturbation of backbone, in our study, the backbone is set to be fixed as a first-order approximation. Therefore, there can be some bias in our computation. However, we assume that the extent of the predicted side-chain shift can imply the interference of the corresponding mutation on the whole structure. From this point of view, the smaller differences (S_{diff}) between the wild-type side chains and the mutant side chains, the smaller functional perturbation the mutation may cause. Therefore, using the side-chain modeling results with a fixed backbone as a first-order approximation is a rational trade-off for studying protein point mutation.

OPUS-Mut can be used to identify the harmful (maximally disturbing) mutations. For a particular mutation, we sum up the differences (S_{diff}) of all side-chain dihedral angles (from χ_1 to χ_4) between the predicted wild-type side chains and the predicted mutant side chains and use it as an indicator for the extent of structural perturbation due to the mutation. By comparing with the experimentally determined functional change data, the results show that larger structural perturbation corresponds to a higher probability of loss of function (Figure 7) or stability decrease (SI Appendix, Figure S5). Furthermore, as shown in Tables 1 and 3, the harmful mutations are most likely in the top rank of the mutations with the maximal predicted structural perturbation.

OPUS-Mut can be used to identify the benign (minimally disturbing) mutations, which may guide us to construct a relatively low-homology mutant sequence but with a similar structure. We screen all possible single-site mutations for HIV-1 protease and T4 lysozyme. Among them, by preferably using the minimally disturbing mutations (with the lowest S_{diff}), the constructed multiple-mutation sequence tends to cause smaller structural perturbation with respect to the wild-type structure, inferred from the predicted backbone generated by AlphaFold2 (Table 2 and SI Appendix, Table S8). For example, the TM-score of the AlphaFold2 prediction for the HIV-protease mutant (SI Appendix, Figure S4) with a 50% mutation rate is 0.85, which is still reasonably high (the TM-score of the predicted wild-type structure is 0.93).

For HIV-1 protease, the summation of differences (S_{diff}) over all residues between the predicted wild type and experimentally determined wild type is 309. For T4 lysozyme, the summation of differences (S_{diff}) between them is 376. Both of them are larger than the predicted differences of the mutation-induced effect. Although in this paper, we use the predicted perturbation (differences between predicted wild-type and predicted mutant) from OPUS-Mut as the indicator to predict the affected residues and infer the possible functional influence caused by the mutation, this gap between the predicted side chains and the experimental results should also be noted, and the further improvement of the accuracy of side-chain modeling methods is necessary.

The results in this paper show that OPUS-Mut may be a useful tool in studying protein mutation, especially for the target whose critical residues are determined. However, OPUS-Mut has some shortcomings that should be studied in the future: first, it cannot deal with the functional changes caused by the mutated residue itself instead by its influence on other residues upon mutation (e.g., causing a steric barrier to the binding site). Second, since the backbone is set to be fixed as a first-order approximation, there can thus be some bias when the target's backbone is significantly shifted upon mutation. In this case, incorporating the results of OPUS-Mut with other complementary methods may be helpful. Third, there is still a gap between the predicted side chains and the experimental results and the further improvement of side-chain modeling methods is required. In addition, ligand information may be included in the future.

■ ASSOCIATED CONTENT

Data Availability Statement

The code and pre-trained models of OPUS-Mut as well as the datasets used in the study can be downloaded from http://github.com/OPUS-MaLab/opus_mut. They are freely available for academic usage only.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00847.

Figure S1. The percentage of correct prediction with a tolerance criterion 20° for all side-chain dihedral angles (from χ^1 to χ^4) of each modification measured by all residues. Figure S2. The percentage of correct prediction with a tolerance criterion 20° for all side-chain dihedral angles (from χ^1 to χ^4) of different methods measured by core residues. Figure S3. Side-chain modeling results of *T*-p53C and its mutants. Figure S4. The structure and sequence of the HIV-1 protease mutant with 50% mutation rate. Figure S5. The summation of differences of all side-chain dihedral angles (from χ^1 to χ^4) summed over all residues (S_{diff}) between the predicted wild-type side chains and the predicted mutant side chains on the T4 lysozyme stability change dataset. Figure S6. Side-chain modeling results of OPUS-Mut for neighboring

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

residues within 5 Å from the ligand in myoglobin, *T*-p53C, and HIV-1 proteases. Figure S7. Side-chain modeling results of four HIV-1 protease crystal structures with various co-crystallized inhibitors. Table S1. The percentage of correct prediction of different residue types with a tolerance criterion 20° for all side-chain dihedral angles (from χ^1 to χ^4) of different methods on CAMEO65 measured by all residues. Table S2. The percentage of correct prediction of different residue types with a tolerance criterion 20° for all side-chain dihedral angles (from χ^1 to χ^4) of different methods on CAMEO65 measured by core residues. Table S3. The significance value *P* (McNemar test) of the residue-wise percentage of correct prediction with a tolerance criterion 20° of each method comparing with OPUS-Mut. Table S4. The RMSD (Å) results of different side-chain modeling methods and the significance value *P* (paired *t*-test) of each method compared with OPUS-Mut on all residues. Table S5. The percentage of correct prediction with a tolerance criterion from 5 to 20° for all side-chain dihedral angles (from χ^1 to χ^4) of different methods on CAMEO65 measured by all residues. Table S6. The HIV-1 protease mutagenesis dataset (0, 1, and 2 denote negative, intermediate, and positive phenotypes, respectively). Table S7. The T4 lysozyme stability change dataset. 0 and 1 denote the decrease and increase in stability, respectively. Table S8. The TM-score of each T4 lysozyme (with 164 residues in length) multiple-mutation sequence predicted by AlphaFold2. Table S9. The percentage of correct prediction with a tolerance criterion 20° for all side-chain dihedral angles (from χ^1 to χ^4) of different methods measured by all residues. Table S10. The significance value *P* (McNemar test) of the residue-wise percentage of correct prediction with a tolerance criterion 20° of each modification comparing with OPUS-Mut.

AUTHOR INFORMATION

Corresponding Author

Jianpeng Ma − *Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 200433, China; Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 201210, China; Shanghai AI Laboratory, Shanghai 200030, China;* orcid.org/0000-0003-2943-0779; Email: jpma@fudan.edu.cn

Authors

Gang Xu − *Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 200433, China; Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 201210, China; Shanghai AI Laboratory, Shanghai 200030, China*

Qinghua Wang − *Center for Biomolecular Innovation, Harcam Biomedicines, Shanghai 200131, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00847

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work was supported by the Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab. The work was also supported by the National Key Research and Development Program of China (No. 2021YFF1200400).

REFERENCES

(1) Kazlauskas, R. J.; Bornscheuer, U. T. Finding better protein engineering strategies. *Nat. Chem. Biol.* **2009**, *5*, 526−529.

(2) Bullock, A. N.; Henckel, J.; Fersht, A. R. Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene* **2000**, *19*, 1245−1256.

(3) Li, B.; Yang, Y. T.; Capra, J. A.; Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput. Biol.* **2020**, *16*, No. e1008291.

(4) Pires, D. E. V.; Ascher, D. B.; Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014**, *30*, 335−342.

(5) Masso, M.; Vaisman, I. I. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. *Bioinformatics* **2007**, *23*, 3155−3161.

(6) Torng, W.; Altman, R. B. 3D deep convolutional neural networks for amino acid environment similarity analysis. *Bmc Bioinformatics* **2017**, *18*, 302.

(7) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **2020**, *9*, 2927−2935.

(8) Sanavia, T.; Birolo, G.; Montanucci, L.; Turina, P.; Capriotti, E.; Fariselli, P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1968−1979.

(9) Xu, G.; Wang, Q.; Ma, J. OPUS-X: An Open-Source Toolkit for Protein Torsion Angles, Secondary Structure, Solvent Accessibility, Contact Map Predictions, and 3D Folding. *Bioinformatics* **2021**, *38*, 108−114.

(10) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1496−1503.

(11) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583.

(12) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millan, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871.

(13) Buel, G. R.; Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **2022**, *29*, 1−2.

(14) Xu, G.; Wang, Q.; Ma, J. OPUS-Rota3: Improving Protein Side-Chain Modeling by Deep Neural Networks and Ensemble Methods. *J. Chem. Inf. Model.* **2020**, *60*, 6691−6697.

(15) Xu, G.; Ma, T.; Du, J.; Wang, Q.; Ma, J. OPUS-Rota2: An Improved Fast and Accurate Side-Chain Modeling Method. *J. Chem. Theory Comput.* **2019**, *15*, 5154−5160.
(16) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, *77*, 778−795.
(17) Liang, S.; Zheng, D.; Zhang, C.; Standley, D. M. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **2011**, *27*, 2913−2914.
(18) Misiura, M.; Shroff, R.; Thyer, R.; Kolomeisky, A. B. DLPacker: Deep Learning for Prediction of Amino Acid Side Chain Conformations in Proteins. *bioRxiv* **2022**, 1278.
(19) Xu, G.; Wang, Q.; Ma, J. OPUS-Rota4: a gradient-based protein side-chain modeling framework assisted by deep learning-based predictors. *Brief. Bioinform.* **2022**, *23*, bbab529.
(20) Xu, G.; Ma, T.; Zang, T.; Wang, Q.; Ma, J. OPUS-CSF: A C-atom-based scoring function for ranking protein structural models. *Protein Sci.* **2018**, *27*, 286−292.
(21) Xu, G.; Wang, Q.; Ma, J. OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics* **2020**, *36*, 5021−5026.
(22) He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J., Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition; CVPR* 2016, 770−778.
(23) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735−1780.
(24) Abadi, M.; Barham, P.; Chen, J. M.; Chen, Z. F.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X. Q., TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation;* USENIX Association 2016, 265−283.
(25) Kingma, D. P.; Ba, J., Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations;* ICLR 2015.
(26) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics* **2019**, *35*, 2403−2410.
(27) Haas, J.; Barbato, A.; Behringer, D.; Studer, G.; Roth, S.; Bertoni, M.; Mostaguir, K.; Gumienny, R.; Schwede, T. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **2018**, *86*, 387−398.
(28) Ordway, G. A.; Garry, D. J. Myoglobin: an essential hemoprotein in striated muscle. *J. Exp. Biol.* **2004**, *207*, 3441−3446.
(29) Quillin, M. L.; Li, T.; Olson, J. S.; Phillips, G. N., Jr.; Dou, Y.; Ikeda-Saito, M.; Regan, R.; Carlson, M.; Gibson, Q. H.; Li, H.; Elber, R. Structural and functional effects of apolar mutations of the distal valine in myoglobin. *J. Mol. Biol.* **1995**, *245*, 416−436.
(30) Nienhaus, K.; Deng, P.; Olson, J. S.; Warren, J. J.; Nienhaus, G. U. Structural dynamics of myoglobin: ligand migration and binding in valine 68 mutants. *J. Biol. Chem.* **2003**, *278*, 42532−42544.
(31) Quillin, M. L.; Arduini, R. M.; Olson, J. S.; Phillips, G. N., Jr. High-resolution crystal structures of distal histidine mutants of sperm whale myoglobin. *J. Mol. Biol.* **1993**, *234*, 140−155.
(32) Calhoun, S.; Daggett, V. Structural Effects of the L145Q, V157F, and R282W Cancer-Associated Mutations in the p53 DNA-Binding Core Domain. *Biochemistry* **2011**, *50*, 5345−5353.
(33) Garufi, A.; Trisciuoglio, D.; Porru, M.; Leonetti, C.; Stoppacciaro, A.; D'Orazi, V.; Avantaggiati, M.; Crispini, A.; Pucci, D.; D'Orazi, G. A fluorescent curcumin-based Zn(II)-complex reactivates mutant (R175H and R273H) p53 in cancer cells. *J. Exp. Clin. Cancer Res.* **2013**, *32*, 72.
(34) Shah, H. D.; Saranath, D.; Murthy, V. A molecular dynamics and docking study to screen anti-cancer compounds targeting mutated p53. *J. Biomol. Struct. Dyn.* **2022**, *40*, 2407−2416.
(35) Joerger, A. C.; Ang, H. C.; Fersht, A. R. Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15056−15061.
(36) Joerger, A. C.; Ang, H. C.; Veprintsev, D. B.; Blair, C. M.; Fersht, A. R. Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. *J Biol Chem* **2005**, *280*, 16030−16037.
(37) Joerger, A. C.; Allen, M. D.; Fersht, A. R. Crystal structure of a superstable mutant of human p53 core domain - Insights into the mechanism of rescuing oncogenic mutations. *J. Biol. Chem.* **2004**, *279*, 1291−1296.
(38) Nikolova, P. V.; Henckel, J.; Lane, D. P.; Fersht, A. R. Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 14675−14680.
(39) Rippin, T. M.; Freund, S. M. V.; Veprintsev, D. B.; Fersht, A. R. Recognition of DNA by p53 core domain and location of intermolecular contacts of cooperative binding. *J. Mol. Biol.* **2002**, *319*, 351−358.
(40) Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **1994**, *265*, 346−355.
(41) Hornak, V.; Simmerling, C. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discovery Today* **2007**, *12*, 132−138.
(42) Mager, P. P. The active site of HIV-1 protease. *Med. Res. Rev.* **2001**, *21*, 348−353.
(43) Ala, P. J.; Huston, E. E.; Klabe, R. M.; McCabe, D. D.; Duke, J. L.; Rizzo, C. J.; Korant, B. D.; DeLoskey, R. J.; Lam, P. Y.; Hodge, C. N.; Chang, C. H. Molecular basis of HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with cyclic urea inhibitors. *Biochemistry* **1997**, *36*, 1573−1580.
(44) Özen, A.; Lin, K. H.; Kurt Yilmaz, N.; Schiffer, C. A. Structural basis and distal effects of Gag substrate coevolution in drug resistance to HIV-1 protease. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15993−15998.
(45) Masso, M.; Vaisman, I. I. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem. Biophys. Res. Commun.* **2003**, *305*, 322−326.
(46) Liu, Z.; Casey, T. M.; Blackburn, M. E.; Huang, X.; Pham, L.; de Vera, I. M. S.; Carter, J. D.; Kear-Scott, J. L.; Veloro, A. M.; Galiano, L.; Fanucci, G. E. Pulsed EPR characterization of HIV-1 protease conformational sampling and inhibitor-induced population shifts. *Phys. Chem. Chem. Phys.* **2016**, *18*, 5819−5831.
(47) Pawar, S.; Wang, Y. F.; Wong-Sam, A.; Agniswamy, J.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Structural studies of antiviral inhibitor with HIV-1 protease bearing drug resistant substitutions of V32I, I47V and V82I. *Biochem. Biophys. Res. Commun.* **2019**, *514*, 974−978.
(48) Loeb, D. D.; Swanstrom, R.; Everitt, L.; Manchester, M.; Stamper, S. E.; Hutchison, C. A., 3rd Complete mutagenesis of the HIV-1 protease. *Nature* **1989**, *340*, 397−400.
(49) Lapatto, R.; Blundell, T.; Hemmings, A.; Overington, J.; Wilderspin, A.; Wood, S.; Merson, J. R.; Whittle, P. J.; Danley, D. E.; Geoghegan, K. F.; Hawrylik, S. J.; Lee, S. E.; Scheld, K. G.; Hobart, P. M. X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature* **1989**, *342*, 299−302.
(50) Sussman, F.; Villaverde, M. C.; Davis, A. Solvation effects are responsible for the reduced inhibitor affinity of some HIV-1 PR mutants. *Protein Sci.* **1997**, *6*, 1024−1030.
(51) King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Schiffer, C. A. Combating susceptibility to drug resistance: lessons from HIV-1 protease. *Chem. Biol.* **2004**, *11*, 1333−1338.
(52) Poteete, A. R.; Hardy, L. W. Genetic analysis of bacteriophage T4 lysozyme structure and function. *J. Bacteriol.* **1994**, *176*, 6783−6788.
(53) Weaver, L. H.; Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* **1987**, *193*, 189−199.

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640

(54) Bell, J. A.; Wilson, K. P.; Zhang, X. J.; Faber, H. R.; Nicholson, H.; Matthews, B. W. Comparison of the crystal structure of bacteriophage T4 lysozyme at low, medium, and high ionic strengths. *Proteins* **1991**, *10*, 10−21.
(55) Dao-pin, S.; Anderson, D. E.; Baase, W. A.; Dahlquist, F. W.; Matthews, B. W. Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry* **1991**, *30*, 11521−11529.
(56) Heinz, D. W.; Baase, W. A.; Matthews, B. W. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 3751−3755.
(57) Xu, J.; Baase, W. A.; Baldwin, E.; Matthews, B. W. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.* **1998**, *7*, 158−177.
(58) Zhang, X. J.; Matthews, B. W. Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* **1994**, *3*, 1031−1039.
(59) Masso, M.; Alsheddi, T.; Vaisman, I. I., Accurate Prediction of Stability Changes in Bacteriophage T4 Lysozyme Upon Single Amino Acid Replacements. *2009 Ieee International Conference on Bioinformatics and Biomedicine;* IEEE *2009*, 26−30.
(60) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **2004**, *32*, D120−D1121.
(61) Rennell, D.; Bouvier, S. E.; Hardy, L. W.; Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **1991**, *222*, 67−88.

https://doi.org/10.1021/acs.jctc.2c00847
J. Chem. Theory Comput. 2023, 19, 1629−1640