

# Protein Model Quality Estimation Using Molecular Dynamics Simulation

Jason Kurniawan and Takashi Ishida\*

Cite This: *ACS Omega* 2022, 7, 24274–24281

Read Online

ACCESS |



Metrics &amp; More

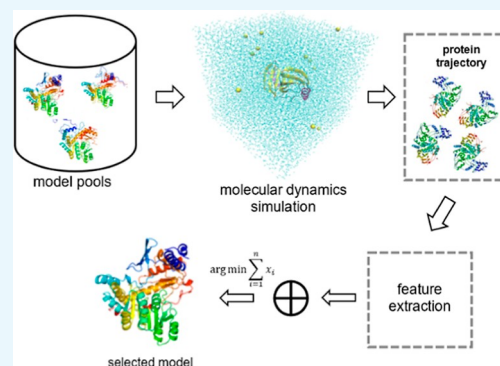


Article Recommendations



Supporting Information

**ABSTRACT:** The estimation of protein model quality remains a challenging task and is important for protein structural model utilization. In the last decade, existing methods that rely on machine learning to deep learning have been developed and shown progressive improvement. Despite utilizing more sophisticated techniques and introducing new features, none of these methods employ explicit protein structure stability information. Hypothetically, protein model quality might be indicated by its structural stability in an in silico system disclosed by the structural difference from its initial structure. One of the possible methods to exploit such information is by implementing molecular dynamics simulations that have shown successful applications in many research fields. We present a novel approach by introducing explicit protein structure stability information using molecular dynamics simulation. Despite using only simple features, small data with no training process required, and a short molecular dynamics simulation time, our method shows comparable performance to the state-of-the-art deep learning-based method.



## INTRODUCTION

The three-dimensional (3D) structure of a protein is the key to understanding its function and has an essential role in drug discovery.<sup>1,2</sup> The structure is typically determined by wet-laboratory experiments, namely, X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy. However, such experimental determination requires a high cost and is time-consuming.<sup>3</sup> To cope with these problems, computational methods have been developed to predict the 3D structure from its primary sequence information. Comparative modeling predicts the 3D structures by identifying one or more known protein structures with a certain degree of similarity (i.e., homologues) with the given query sequence and then maps the residues in the query sequence to residues in the template sequence through alignment. When no homologue sequences are found, de novo modeling predicts the 3D structures by employing the general folding and energetic principles. Nevertheless, the current prediction scheme generates multiple structure models due to multiple template structures found and/or the protein conformational sampling. Thus, it raises the need to select the best model that has the closest conformation to the unknown native structure. This is known as the estimation of (protein) model accuracy or is often referred to as model quality assessment (MQA).

In the past, classical MQA methods made use of scoring functions from statistical potentials such as DFIRE,<sup>4</sup> DOPE,<sup>5</sup> GOAP,<sup>6</sup> and RWplus.<sup>7</sup> However, the considerable success in various research fields has shifted the trend for MQA method development to the utilization of machine learning techniques. This can be seen as in the last decade, MQA methods majorly

employed various machine learning to deep learning techniques in their pipeline due to the increasing number of known 3D structures and available standard data sets. The methods typically extract features from the protein structure and/or sequence information and then use a supervised learning technique to predict the model quality accuracy using specific evaluation metrics. For example, ProQ<sup>8</sup> uses a combination of structural and evolutionary information as features to train a support vector machine, ProQ3<sup>9</sup> uses similar techniques with the addition of Rosetta energy terms,<sup>10</sup> and RFMQA<sup>11</sup> trains random forest using statistical potential and energetic property information. Recently, the utilization of deep learning techniques has achieved top performance and has become the basis of state-of-the-art MQA methods.<sup>12,13</sup> Deep learning can take advantages of both high-level and particularly low-level features using specific network architectures. For instance, ProQ3D<sup>14</sup> and DeepQA<sup>15</sup> extract the high-level features obtained from the output of other methods and feed them to a multilayer perceptron. Other methods exploit convolutional neural network (CNN) variations, including 3DCNN<sup>16–19</sup> and graph CNN,<sup>20</sup> that learn low-level structural information. DeepAccNet<sup>21</sup> estimates per-residue accuracy using a deep

Received: March 11, 2022

Accepted: June 21, 2022

Published: July 5, 2022



residual network consisting of 3D followed by 2D convolutions to evaluate local environments and global context. QDeep<sup>22</sup> introduces inter-residue distance combined with multiple sequence alignment information and then trains the extension of CNNs using deep residual network architecture.

Despite the application of more sophisticated algorithms, the principally existing methods extract geometric, sequence, and energetic features calculated from static structure information. Protein is not static but flexible and dynamic.<sup>23</sup> This is shown by the change in the shape of the protein, known as conformational change. The conformational change is often induced by environmental factors like temperature and denaturants. A single-protein structure might have multiple conformational shapes before going back to its equilibrium state, occurring on different length and time scales. High-quality structures or accurate models tend to maintain their stability by keeping the folding shape in the equilibrium state. In contrast, low-quality structures have significant deviation and drastic changes from their initial structure caused by instability. Model inaccuracies significantly impact protein stability due to its structural changes over time in an *in silico* system. This is indicated by the deviation from the original structure, which is directly correlated with the loss of quality of the model.<sup>24</sup> Thus, protein structural stability information might be useful for the MQA task and has not been employed by any existing methods. To derive such information, one of the possible methods involves applying a molecular dynamics (MD) simulation technique.

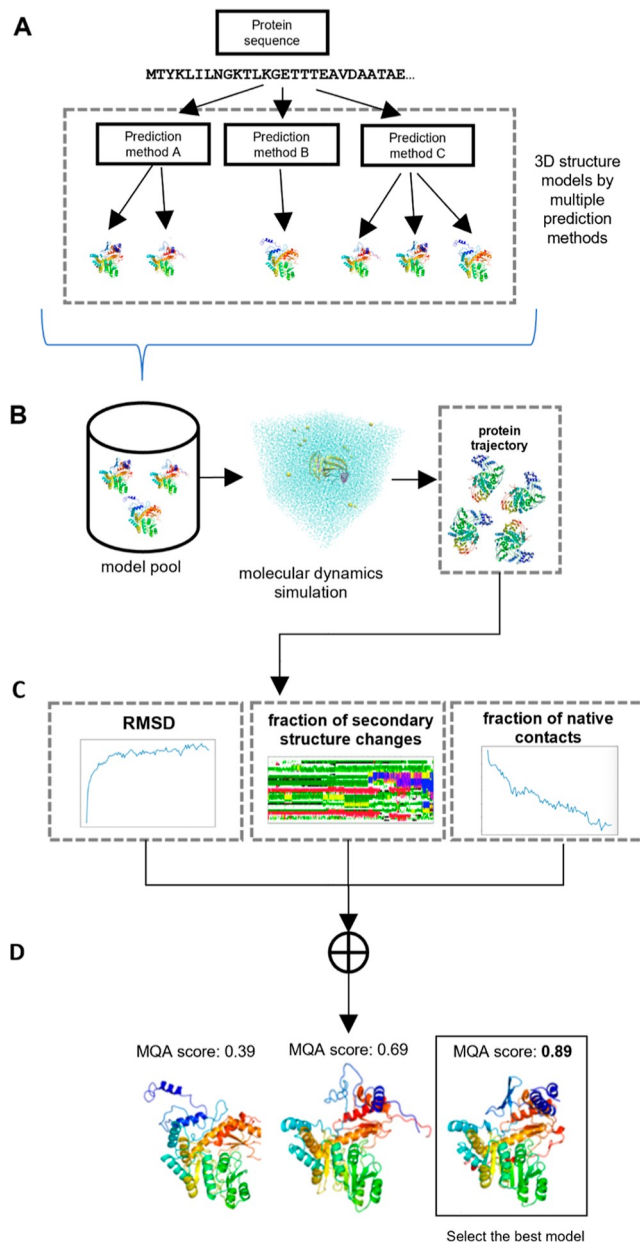
MD simulation (of protein) is the process of computing forces iteratively for a fixed time to solve the Newton equation of motion using a known protein structure (e.g., predicted structure) and the selected force field. It has shown many successful applications in vast research fields, such as understanding allosteric regulation, docking strategies for drug design, and protein structure refinement.<sup>24</sup> Other studies show the feasibility of MD simulation to determine protein structure stability,<sup>25</sup> which reveals consistent results with the stability determined experimentally.<sup>26</sup> The output of MD simulation is known as a “trajectory”, which contains the atom position, velocity, and energy information over time. This information is useful for analyzing the structural changes over time in an *in silico* system. Motivated by the success of MD simulation applications, we propose a novel MQA method that discloses protein structural stability information by incorporating MD simulation.

In this work, we propose a novel approach for protein model quality estimation using protein structural stability information obtained from MD simulations. We introduce three features: root-mean-square deviation (rmsd), the fraction secondary structure, and the fraction of native contacts to the initial structure. The main hypothesis of this work is that the predicted structure model quality affects the structure's stability in an *in silico* system disclosed by the structural difference from its initial structure. We believe that structural stability information might be useful for MQA. Despite using only simple feature combinations obtained from a short and uniform MD simulation setup, our proposed method shows comparable performance with state-of-the-art methods that are usually trained on multiple large critical assessment of structure prediction (CASP) data sets and complex deep learning models. This can be advantageous in cases where the training data sets or sequence homology information is not available. Moreover, this method can be easily implemented even for people with no prior expertise in MD simulation. The main contribution of our

proposed method is that this work is the first to utilize protein structure stability information for the MQA task.

## MATERIALS AND METHODS

The proposed method consists of three steps: protein trajectory generation through MD simulation, feature extraction, and best model selection (Figure 1).



**Figure 1.** Schematic diagram of the proposed method: (A) model pool generation by structure prediction methods, (B) protein trajectory generation through MD simulation, (C) feature extraction, and (D) best model selection with GDT\_TS values as an example.

**Data Sets.** CASP data sets are one of the standard benchmark data sets broadly used to evaluate MQA method performance.<sup>27</sup> The data sets are updated every 2 years and can be accessed through the CASP website ([https://predictioncenter.org/download\\_area/](https://predictioncenter.org/download_area/)). A single CASP data set contains numerous protein pools. Each pool consists of multiple predicted structure models from different structure

prediction methods (Figure 1A). MQA methods generally train and evaluate using multiple CASP data sets, including ten to a hundred thousand predicted structures. However, due to the different approaches of our method and the computational cost limitation of MD simulation, we only chose sample pools from a single CASP. In this work, we selected the test data from QDeep, which consisted of 20 protein pools with various protein sizes from CASP13 stage 2. Each pool here contains 150 predicted models, except for T0951, with 149 models.

**MD Simulation.** MD simulation requires protein structure information as the input. This information is provided by the predicted structures in the CASP data set. On the other hand, knowledge about the protein environment is rarely available.<sup>17</sup> To compensate for such limitations, we make the simulation parameters and environment uniform by following the relatively simple setup described by Lemkul, J., which solvates the proteins in a cubic box filled with water molecules and added ions.<sup>28</sup> Protein folding simulations typically require long simulation times that range from tens to hundreds of nanoseconds (ns) for the conformational space searches.<sup>29</sup> Here, we perform a relatively short 1 ns simulation. MD simulation is commonly performed at room temperature of 300 K. A previous study shows that helical proteins in explicit water tend to destabilize faster within the same timeframe while being simulated at higher temperatures.<sup>30</sup> Thus, we use a higher temperature of 500 K to speed up the stabilization effect, allowing us to get such information in short simulations (Table 1). We then perform

**Table 1. Chosen Parameters for the MD Simulation<sup>a</sup>**

parameter	value
force field	OPLS/LA <sup>34</sup>
water model	SPC/E <sup>35</sup>
ions	Na (+), CL (-)
temperature	500 K
energy minimization	steepest descent

<sup>a</sup>The command lines to execute the simulation and GROMACS MDs parameters (.mdp files) are provided in the Supporting Information file.

post-processing in the final step by removing the periodic boundary condition after centering the protein molecule to avoid boundary effect problems. The MD simulation is implemented using GROMACS version 2019.4.<sup>31</sup> The final output of the simulation is protein trajectories containing raw information, including structural and positional changes of the protein over time. To validate the simulation results, we calculate the rmsd value of the final to the initial structure with the threshold of 2.5 Å.<sup>32</sup> A simulation is marked as invalid if the rmsd value of the final trajectory to the initial structure is larger than 2.5 Å. This is to ensure that there are no unrealistic movements during the simulation as a result of drastic structure changes or clashes. The final output of the MD simulation is the protein trajectory required for the feature extraction step (Figure 1B).

**Feature Extraction.** Our method suggests that protein structure stability in an in silico system might indicate structural quality. To incorporate such structure stability information, we define features extracted from the protein trajectories. The results of previous work<sup>25,33</sup> reveal that the rmsd and fraction of native contacts to the initial structure are useful to monitor structure stability. We also define new stability information: the fraction of the secondary structure to the initial structure. These

three types of structural change information are extracted from the MD trajectories and defined as features (Figure 1C). The feature extraction step is implemented using the MDTraj library.<sup>36</sup> These features are later used as the input for the best model selection method. In the pilot phase of this work, we also calculated other potential features from the MD trajectories, such as the radius of gyration and solvent accessibility surface area. However, these features did not significantly correlate to the structure quality.

**Root-Mean-Square Deviation.** rmsd is an evaluation metric to measure structural similarity by calculating the average distance of the selected atoms between a trajectory of structures to one reference state. The reference is often defined as the initial structure of the trajectory. This can provide insight into the overall structure movement from the initial structure. Stable rmsd values can measure structural stability and conformational convergence. Low-quality models hypothetically have unstable structures with more significant atom position deviations than the initial structure. In contrast, high-quality predicted models ideally have lower rmsd values. For the featurization, we calculate the rmsd value of the last trajectory to the initial structure and then normalize it by using 1 Å cutoff. As the final step, we invert the rmsd values as  $1 - \text{rmsd}$ . Thus, from this featurization, the best model is selected based on the highest feature value.

**Fraction of Secondary Structure Changes.** The protein stability can be examined through the secondary structure-type changes among conformational-state transitions. For example, the unstable structures might have numerous secondary-type changes between a trajectory state and the initial structure. Conversely, stable structures tend to maintain their secondary structure type. To represent this information as a feature, we define the fraction of secondary structures as follows

$$SS(X) = \frac{\sum c^X - c^0}{n} \quad (1)$$

where  $X$  is a conformation state,  $n$  is the total number of residues, and  $c$  is a  $1 \times n$  binary matrix, where the value of 1 represents the secondary structure type that is not changed compared to the initial structure. Here, we define eight secondary structure types as determined using the DSSP program.<sup>37</sup> High-quality models hypothetically have stable structures with a higher fraction value. Thus, the best model is selected based on the highest value. Like rmsd, we calculate the value change between the last trajectory and the initial structure.

**Fraction of Native Contacts.** The results of a previous work revealed that the fraction of native contacts is useful to monitor structure stability alongside rmsd.<sup>25</sup> Hence, we apply this information as an additional feature. Native contacts are formed during the transition between two conformational states. The stability of the proteins is reflected by the fraction of native contacts in the presence of denaturants. This feature is computed according to the definition:<sup>38</sup>

$$N(X) = \frac{1}{|S|} \sum_{(i,j) \in S} \frac{1}{1 + \exp[\beta(r_{ij}(X) - r_{ij}^0)]} \quad (2)$$

where  $X$  is a conformation,  $r_{ij}(X)$  is the distance between atoms  $i$  and  $j$  in conformation  $X$ ,  $r_{ij}^0$  is the distance from heavy atom  $i$  to  $j$  in the native-state conformation,  $S$  is the distance from  $e$  the set of all pairs of heavy atoms  $(i, j)$  belonging to residues  $\theta_i$  and  $\theta_j$ , such that  $|\theta_i - \theta_j| > 3$  and  $r_i^0 < 4.5$  Å,  $\beta = 5$  Å<sup>-1</sup>,  $\lambda = 1.8$ , and  $X$  is the structural conformation of the last trajectory. Thus, the

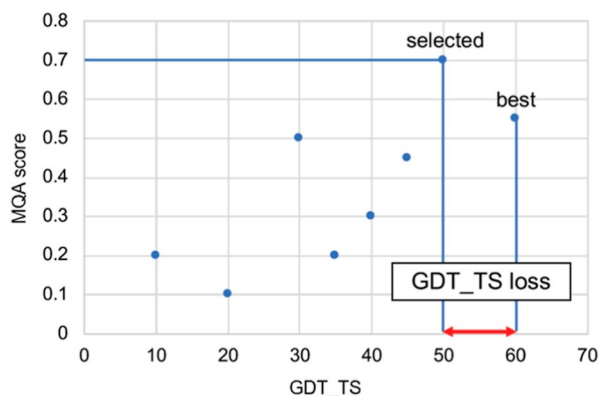
higher its value, the more stable the structure. Similar to previous features, the best model is the model with the highest feature value.

**Best Model Selection.** As mentioned in each feature definition, high-quality models hypothetically have a stable structure. Stable structures ideally have low atom position deviation, less secondary structure-type changes, and high native heavy atom contacts to their initial structure. These are represented by inverted rmsd, a fraction of secondary structure changes, and a fraction of native contacts, respectively. Thus, the best model in a prediction pool is the model with the highest feature value, defined as follows:

$$\text{Best model}(T) = \operatorname{argmax} \sum_{i=1}^n x_i \quad (3)$$

where  $T$  is the selected pool,  $n$  is the total number of models in each pool, and  $x_i$  is the selected feature. We also investigate the model selection results from the combination of the features. The features are combined by simple addition, and the best model is selected based on the highest value (Figure 1D).

**Evaluation Method.** The quality of the predicted structure model is quantified using global distance test total scores (GDT\_TS). GDT\_TS is an accuracy-like score that indicates the structure similarity between the predicted models and the native structure.<sup>39</sup> We evaluate the performance of MQA methods by calculating the GDT\_TS loss, that is, the difference between true GDT\_TS of the model selected by the MQA methods and GDT\_TS of the best/most accurate model in a protein pool (Figure 2). A lower loss indicates better performance. This evaluation method measures the ability of the MQA methods to find the best model in protein pools.



**Figure 2.** Illustration for GDT\_TS loss. MQA methods generally select the best model using certain scoring methods.

## RESULTS

**Simulation Results.** During the energy minimization step, not all prediction models from the HMSCraper-refiner group could be simulated due to the incompleteness/missing atoms. Thus, we excluded models from this group in all pools except for T1016 since it does not contain prediction models from the group. Additionally, a few different prediction models from other groups could not be simulated due to the occurrence of overlapping atoms. Only a small percentage of unsuccessful simulations in each pool (excluding prediction from the HMSCraper-refiner group) were found. Theoretically, a larger protein size tends to have higher difficulty in the simulation. This

is found in the increasing percentage of unsuccessful simulations on larger proteins, particularly when the number of residues is > 300 (Table S1). The results also show that our proposed MD simulation setup is effective for all pools with an average simulation success rate larger than 90%. Since the ratio of unsuccessful simulations is less than 10%, we omit these data from the simulation results. In addition, the simulation validation results show that 7 of 20 pools contain invalid simulations (rmsd > 2.5 Å) with the ratio relative to the number of successful simulations that is smaller than 10% (Table S2). Since the ratio of invalid simulations is less than 10%, we also exclude them from the simulation results.

**Feature Combination.** The best model in each pool is selected based on the highest feature values. In the experiment, we evaluate the performance of each singular feature and all possible feature combinations. The results show that combining all three proposed features led to the best performance according to the average top 1 GDT\_TS loss and the number of actual best models in each pool (Table S3). Even though the combination between rmsd and native contacts features results in a slightly lower average GDT\_TS loss, the difference is only 0.004, and it only successfully selected the actual best model in one pool, while the combination of all the three features selected two actual best models in two pools. Thus, the combination of all of the three features is selected as the main proposed method.

**Performance at Different Simulation times and Temperatures.** Performing simulation at unusually high temperatures like 500 K might be harmful to the protein structure stability even for high-quality models. However, this also might accelerate the destabilization effect in shorter simulation length and could reduce the computational cost. To confirm this effect, we perform additional MD simulations at lower temperatures of 300K and 400 K. We then compare the performance results at different temperatures and simulation lengths. The results show that simulation at unusually high temperature and shorter length fastened the destabilization effect as the 500 K and 0.5 ns simulation achieved the best performance with the lowest average top 1 GDT\_TS loss and the highest number of the actual selected best model (Table S4). However, both results of 400 K simulations show worse performance than 300K simulation within the same simulation length. This might be because the temperature difference is not sufficiently high enough and thus the performance did not significantly change. The best performance results from 500 K and 0.5 ns simulation then are taken as the main results for the proposed method.

**Performance Evaluation.** For this scenario, we compare the main results of the proposed method with the results from QDeep. The proposed method shows comparable performance to QDeep, where it achieved a lower average top 1 GDT\_TS loss with 0.008 difference (Table 2). The individual pool results also show comparable performance, where the method achieves eight wins, four draws, and eight losses. In several pools, our method significantly outperforms QDeep. For instance, in T1008, our method attains a GDT\_TS loss of 0.110, while QDeep achieves a poor performance of 0.455. This comes from the disadvantage of using the MSA feature-based method, including QDeep, where the alignment depth of the MSA for T1008 is zero with no identifiable homologous sequences.<sup>22</sup> Our method does not rely on such information since the features are acquired solely from the protein structural information. The results also show that our proposed method successfully selected the actual best model with zero GDT\_TS loss in three pools

**Table 2. Top 1 Model GDT\_TS Loss Comparison Between Our Proposed Method and QDeep on the CASP13 Stage 2 Data set<sup>a</sup>**

pool name	baseline	proposed	QDeep	GDT_TS of actual best model
T0950	<u>0.215</u>	0.246	0.030	0.385
T0951	0.167	<b>0.008</b>	0.057	0.943
T0953s1	0.171	0.067	0.041	0.489
T0953s2	<u>0.267</u>	0.358	0.028	0.631
T0954	0.239	<b>0</b>	0	0.699
T0955	0.308	<b>0.043</b>	0.171	0.951
T0957s1	0.259	<b>0</b>	0.151	0.544
T0957s2	0.267	<b>0.019</b>	0.261	0.610
T0958	0.253	<b>0.127</b>	0.133	0.740
T0960	<u>0.101</u>	0.102	0.078	0.484
T0963	0.124	<b>0.121</b>	0.121	0.516
T0966	0.171	0.098	0.006	0.611
T0968s1	0.323	<b>0.057</b>	0.057	0.667
T0968s2	0.387	<b>0</b>	0.130	0.713
T1003	0.110	<b>0.047</b>	0.047	0.895
T1005	0.154	0.063	0	0.558
T1008	0.449	<b>0.179</b>	0.455	0.870
T1009	0.140	0.016	0.003	0.673
T1011	0.171	0.105	0.043	0.686
T1016	0.055	<b>0.005</b>	0.014	0.816
Average	0.217	<b>0.083</b>	0.091	0.674

<sup>a</sup>The underlined marks indicate that the baseline method performs better than the proposed method. The bold marks indicate that our proposed method achieves better/draw performance than QDeep.

(T0954, T0957s1, and T0968s2) while QDeep was successfully selected in two pools (T0954 and T1005). To compare the performance of our method to random selection, we add the GDT\_TS loss of the baseline method, which is the average GDT\_TS of each pool. It is shown that our method achieves significantly superior overall performance. However, in T0950, T0953s1, and T0960, our method shows worse performance than the baseline method. In addition, we computed the Wilcoxon signed-rank test with  $\alpha = 0.05$  between the proposed versus baseline method and the proposed method versus QDeep results. The statistical test results between the proposed versus baseline method reject the null hypothesis with  $p$ -value = 0.0001, which means that the proposed method is significantly different from random selection. On the other hand, the test results of the proposed method versus QDeep fail to reject the null hypothesis with  $p$ -value = 0.87. This means that the proposed method achieves comparable performance with QDeep.

## DISCUSSION

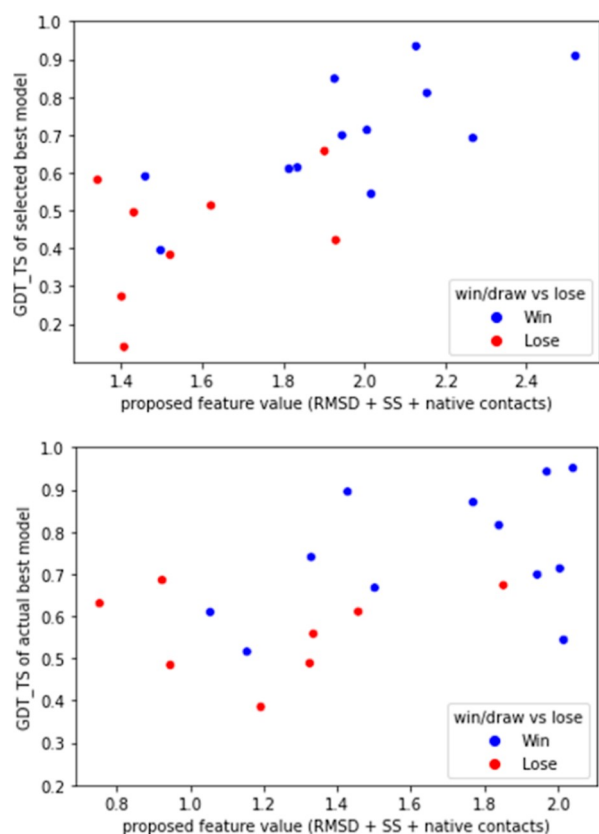
This work shows the possibility and potential application of using protein stability information to estimate the quality of a protein model. This information is derived from the structural change information over time obtained through MD simulation. We propose three features representing the protein stability information: rmsd, a fraction of the secondary structure, and a fraction of native contact information to the initial structure. Thus far, no previous MQA method has utilized the stability information explicitly. Our approach does not use any additional predictive features or evolutionary information, such as the predicted secondary structure or sequence profiles from multiple sequence alignment homologues. Furthermore, our method does not rely heavily on machine learning methods that

require training on tens to hundreds of thousands of models, that is, training on multiple CASP data sets.

**Quality of Unsuccessful and Invalid Simulated Structures.** Our method requires protein trajectory information in the first step by conducting MD simulation. A small percentage of the models could not be simulated successfully and was omitted from the simulation results. However, the omission might discard the top models in each pool related to the selection of the best model in each pool. Hypothetically, poor-quality models and/or models with structure incompleteness are the main causes of unsuccessful simulations; thus, the omission will not discard good-quality models. To prove this hypothesis, we compare the model quality distribution between successful and unsuccessful data (Figure S1). It is shown that the omission of unsuccessful simulation data “filters” the low-quality models from each pool, especially for low-size proteins. This is plausible because low-quality models typically have structural problems as mentioned above and are found in the omitted models.

Running MD simulation under extreme physical conditions like high temperature might cause drastic structural deviation even in a short simulation. We further investigate whether invalid simulations are coming from low-quality models in the pools. The results show that all these invalid simulations are found in pools with no high-quality models (GDT\_TS > 80) and models with lower quality relative to other models in the same pool, except for T0960 and T0963 (Figure S2). They also have a larger percentage of invalid simulations compared to the other five pools. This is because these two models did not contain any outstanding prediction models, as the GDT\_TS score ranged between 30 and 50 s, unlike other pools with larger ranges. Nevertheless, none of the actual highest quality models in these two pools were marked as invalid simulations.

**Case Study.** The poor performance in the T1008 pool induced the huge disadvantage of QDeep when there were no identifiable homologous sequences. Correspondingly, our method achieves significantly poor performance compared to QDeep with the difference of GDT\_TS loss larger than 0.1 in T0950 and T0953s2. Our method suggests that the structural stability might indicate the quality of the model, which is represented by the feature values. When there are no high-quality models in a pool, it is more difficult for our method to select and distinguish between bad- and good-quality models since there is no significant feature value difference between them. This is found in two pools where the proposed method shows significantly worse performance than QDeep, and none of the pools has high-quality models (Figure S3). On further inspection, we find consistent results with our hypothesis, where the proposed feature values can select the best model in the winning cases if high-quality models (GDT\_TS > 70) are available in pools (Figure 3). This is also found in the comparison between the proposed feature value and the GDT\_TS of the actual best model. Thus, our method has a major advantage when there are high-quality models in the prediction pools as found in T0951, T0955, T1003, T1008, and T1016. This is more useful and applicable for real applications, where selecting the best, high-quality models is the primary goal. Interestingly, the winning cases are also found in the pools where high-quality models are not available such as in T0957s1 and T0957s2. We then investigate further by comparing the best model between the two winning cases with the lose cases whose GDT\_TS is larger than 50. In protein MQA, the GDT\_TS value of larger than 50 often indicates that the majority of secondary structure



**Figure 3.** Top: proposed feature value versus GDT\_TS of the best model selected by the proposed method. Bottom: proposed feature value versus GDT\_TS of the actual best model.

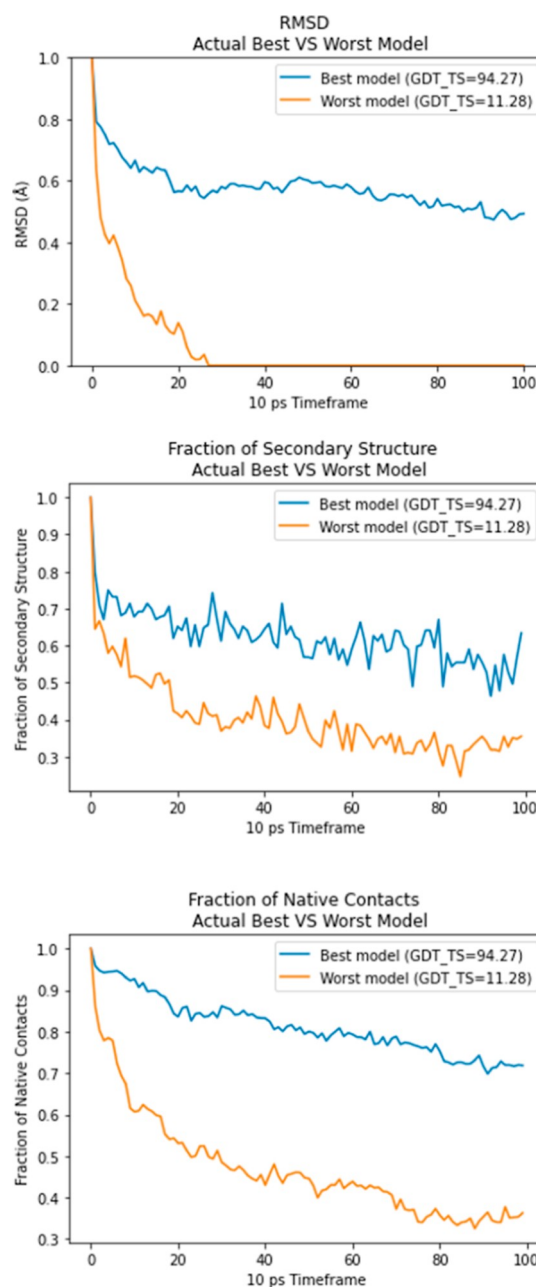
composition is correctly predicted. We found that the structure of the actual best model in the two winning pools has fewer random and long terminal coils, unlike those from the five lose cases (Figure S4). This might highly affect the method's performance in the lose cases since the proposed features, rmsd and the fraction secondary structure, are sensitive to the fluctuation bias that comes from these coil regions.

Additionally, we also investigate each feature value between low- and high-quality models. A high-quality model should have lower and stable fluctuations over time. We took the T0951 pool as an example since this pool contains a large number of high-quality models and also low-quality models. Each singular feature value between the actual best and worst model in the T0951 pool shows a significant feature value change over time (Figure 4).

**Feature Weight Ratio.** Each feature might have more contributions than the others. To examine this, we define the weights for the features as follows

$$\text{Best model } (T) = \operatorname{argmax} \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i} \quad (4)$$

where  $T$  is the selected target,  $n$  is the total number of models in each target,  $x_i$  is the selected feature, and  $w_i$  is the weight for each feature. Using various weight ratio combinations, the currently proposed method with balanced weight ratios achieves the best performance with the lowest average top 1 GDT\_TS loss (Table S5). Our proposed method feature combination is already appropriate for the best model selection. The weight optimization itself might be contrary to the advantages of the



**Figure 4.** Each singular feature value between the actual best versus the worst model in the T0951 pool.

proposed method that does not require any training or optimization steps, making the method data set-dependent.

**Computing Time for Model Quality Assessment.** The proposed method uses an MD simulation, and thus, it needs more computing resources than machine learning-based MQA methods, whose running time is generally less than a minute. In this research, we used 1 NVIDIA Tesla V100 SMX2 GPU for the MD simulation, and the running times of the proposed method generally ranged between 1000 to 3000 s (Figure S5). The running time slightly depends on the protein size, but each pool has a different running time although the proposed method employs uniform MD simulation parameters for all models. Especially, there were some outliers represented by extremely long running time. We found that extremely long running times were mainly caused by the unusual structures. For instance, the prediction model MUFold\_server\_TS4 in T1003, whose

running time was the longest, had long coil terminals (Figure S6). The long coil terminals caused larger energy minimization, and the simulation system became much larger than the others. Thus, to reduce the computing cost, we may need to remove such regions before applying the proposed method.

Furthermore, the previous work shows that a systematic MD simulation study of temperature dependency requires numerous temperature parameters that run on ten to hundreds of nanosecond simulations.<sup>40</sup> This becomes the limitation of our proposed method since such experiments demand huge computational resources and time. Despite the fact that the current temperature and simulation length parameters have shown promising results, further systematic studies using different simulation conditions might be necessary to re-evaluate and improve the performance of the methodology.

## CONCLUSIONS

We propose a novel approach for model quality estimation by introducing explicit protein structure stability information derived from MD simulation. In this work, we use relatively simple, uniform parameters and a short MD simulation time to extract the stability information as features. A combination of the features is useful for selecting the best prediction model. Despite using only simple feature combinations and short MD simulation time, our proposed method shows comparable performance with existing state-of-the-art deep learning-based methods typically trained on large, multiple-CASP data sets. Thus, the introduction of explicit protein stability information might be a valuable addition to the existing MQA methods.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.2c01475>.

Percentage of successful simulation, percentage of invalid simulation, performance of each singular and feature combination, performance at different temperatures and simulation lengths, average top 1 GDT\_TS loss using different ratios, GDT\_TS distribution between successful and unsuccessful simulation, 3D structure and GDT\_TS of win and lose cases, MD simulation running time, and 3D structure of prediction models with the longest running time (PDF)

GROMACS configuration (.mdp) files and parameters to run the standard MD simulation (ZIP)

## AUTHOR INFORMATION

### Corresponding Author

Takashi Ishida – Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo 152-8550, Japan; [orcid.org/0000-0002-9478-3223](https://orcid.org/0000-0002-9478-3223); Email: [ishida@c.titech.ac.jp](mailto:ishida@c.titech.ac.jp)

### Author

Jason Kurniawan – Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo 152-8550, Japan; [orcid.org/0000-0002-6471-8332](https://orcid.org/0000-0002-6471-8332)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acsomega.2c01475>

## Author Contributions

J.K provided the methodology, experiments, implementation, and wrote the original draft. T.I supervised the conduct of this study. All authors reviewed the manuscript draft and revised it critically on intellectual content.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This study was carried out using the TSUBAME3.0 supercomputer at the Tokyo Institute of Technology.

## REFERENCES

- (1) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93–96.
- (2) Jacobson, M.; Sali, A. Comparative protein structure modeling and its applications to drug discovery. *Annu. Rep. Med. Chem.* **2004**, *39*, 259–276.
- (3) Wallner, B.; Elofsson, A. Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* **2005**, *21*, 4248–4254.
- (4) Zhou, H.; Zhou, Y. Distance scaled, finite ideal gas reference state improves structure derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2002**, *11*, 2714–2726.
- (5) Shen, M.-y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507–2524.
- (6) Zhou, H.; Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **2011**, *101*, 2043–2052.
- (7) Zhang, J.; Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **2010**, *5*, No. e15386.
- (8) Ray, A.; Lindahl, E.; Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinf.* **2012**, *13*, 224.
- (9) Uziela, K.; Shu, N.; Wallner, B.; Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. *Sci. Rep.* **2016**, *6*, 33509.
- (10) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **2011**, *487*, 545–574.
- (11) Manavalan, B.; Lee, J.; Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS One* **2014**, *9*, No. e106542.
- (12) Won, J.; Baek, M.; Monastyrskyy, B.; Kryshtafovych, A.; Seok, C. Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning. *Proteins: Struct., Funct., Bioinf.* **2019**, *87*, 1351–1360.
- (13) Kwon, S.; Won, J.; Kryshtafovych, A.; Seok, C. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1940–1948.
- (14) Uziela, K.; Menéndez Hurtado, D.; Shu, N.; Wallner, B.; Elofsson, A. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **2017**, *33*, 1578–1580.
- (15) Cao, R.; Bhattacharya, D.; Hou, J.; Cheng, J. DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinf.* **2016**, *17*, 495.
- (16) Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* **2018**, *34*, 4046–4053.
- (17) Pagès, G.; Charmettant, B.; Grudinin, S. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics* **2019**, *35*, 3313–3319.

- (18) Sato, R.; Ishida, T. Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network. *PLoS One* **2019**, *14*, No. e0221347.
- (19) Takei, Y.; Ishida, T. P3CMQA: single-model quality assessment using 3DCNN with profile-based features. *Bioengineering* **2021**, *8*, 40.
- (20) Baldassarre, F.; Menéndez Hurtado, D.; Elofsson, A.; Azizpour, H. GraphQA: protein model quality assessment using graph convolutional networks. *Bioinformatics* **2021**, *37*, 360–366.
- (21) Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **2021**, *12*, 1340.
- (22) Shuvo, M. H.; Bhattacharya, S.; Bhattacharya, D. QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. *Bioinformatics* **2020**, *36*, i285–i291.
- (23) Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **2014**, *43*, 5051–5066.
- (24) Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpi, J. L. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinf. Chem.* **2015**, *8*, 37–47.
- (25) Zhang, D.; Lazim, R. Application of conventional molecular dynamics simulation in evaluating the stability of apomyoglobin in urea solution. *Sci. Rep.* **2017**, *7*, 44651.
- (26) Luo, Y.; Kay, M. S.; Baldwin, R. L. Cooperativity of folding of the apomyoglobin pH 4 intermediate studied by glycine and proline mutations. *Nat. Struct. Biol.* **1997**, *4*, 925–930.
- (27) Moulton, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, ii–iv.
- (28) Lemkul, J. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1*, 5068.
- (29) Bowman, G. R. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *J. Comput. Chem.* **2016**, *37*, 558–566.
- (30) Duan, L.; Guo, X.; Cong, Y.; Feng, G.; Li, Y.; Zhang, J. Z. H. Accelerated molecular dynamics simulation for helical proteins folding in explicit water. *Front. Chem.* **2019**, *7*, 540.
- (31) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (32) Tsai, H. H.; Tsai, C. J.; Ma, B.; Nussinov, R. In silico protein design by combinatorial assembly of protein building blocks. *Protein Sci.* **2004**, *13*, 2753–2765.
- (33) Aier, I.; Kumar Varadwaj, P.; Raj, U. Structural insights into conformational stability of both wild-type and mutant EZH2 receptor. *Sci. Rep.* **2016**, *6*, 34984.
- (34) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (35) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, A. J. *Intermolecular forces*; D. Reidel Publishing Company, 1981.
- (36) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.
- (37) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (38) Best, R. B.; Hummer, G.; Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17874–17879.
- (39) Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **2003**, *31*, 3370–3374.
- (40) Zhuang, X.; Makover, J. R.; Im, W.; Klauda, J. B. A systematic molecular dynamics simulation study of temperature dependent bilayer structural properties. *Biochim Biophys Acta* **2014**, *1838*, 2520–2529.