nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23692-x OPEN

Assessment of protein–protein interfaces in cryo-EM derived assemblies

Sony Malhotra [1,2]✉, Agnel Praveen Joseph [2], Jeyan Thiyagalingam [2] & Maya Topf [1,3]✉

Structures of macromolecular assemblies derived from cryo-EM maps often contain errors that become more abundant with decreasing resolution. Despite efforts in the cryo-EM community to develop metrics for map and atomistic model validation, thus far, no specific scoring metrics have been applied systematically to assess the interface between the assembly subunits. Here, we comprehensively assessed protein–protein interfaces in macromolecular assemblies derived by cryo-EM. To this end, we developed Protein Interface-score (PI-score), a density-independent machine learning-based metric, trained using the features of protein–protein interfaces in crystal structures. We evaluated 5873 interfaces in 1053 PDB-deposited cryo-EM models (including SARS-CoV-2 complexes), as well as the models submitted to CASP13 cryo-EM targets and the EM model challenge. We further inspected the interfaces associated with low-scores and found that some of those, especially in intermediate-to-low resolution (worse than 4 Å) structures, were not captured by density-based assessment scores. A combined score incorporating PI-score and fit-to-density score showed discriminatory power, allowing our method to provide a powerful complementary assessment tool for the ever-increasing number of complexes solved by cryo-EM.

[1] Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, London, UK. [2] Scientific Computing Department, Science and Technology Facilities Council, Didcot, UK. [3] Centre for Structural Systems Biology, Leibniz-Institut für Experimentelle Virologie and Universitätsklinikum Hamburg-Eppendorf (UKE), Hamburg, Germany. ✉email: sony.malhotra@stfc.ac.uk; m.topf@cryst.bbk.ac.uk

A majority of proteins are known to interact in order, to perform their functions, and sustain the activities of cells. Unveiling the molecular details underlying these functions provides crucial structural, and functional insights. In recent years, cryo-EM has become a prominent technique for solving the structures of complex biological systems, such as polymerases, transmembrane receptors, viral assemblies and ribosomes, by overcoming some of the limitations of X-ray crystallography and NMR spectroscopy[1]. Cryo-EM techniques usually require a small amount of sample, are more forbearing on sample purity, and the rapid freezing of the sample maintains its closeness to native state. Due to these strengths, cryo-EM provides an alternative to X-ray crystallography for large complexes. Recent advances in instrumentation and image processing methods in structure determination using cryo-EM and tomography of sub-cellular structures have pushed the resolution limit of structures towards the near-atomic range. However, the average resolution of structures solved using single-particle cryo-EM per year (since 2002) is worse than 5 Å (for example, 5.6 Å for 2019 and 6.3 Å for 2020), and determining structures at near-atomic resolution is still a challenge[2,3]. Additionally, many of the cryo-EM maps associated with a near-atomic global resolution have regions at intermediate (~4–8 Å) resolutions (or even lower), owing to the variability in local resolution.

The resolution of the cryo-EM map dictates the approach to be adopted for model building, fitting, refinement and validation to a great extent[4]. Regardless of the resolution of the map, upon model building and/or fitting, assessment of the atomistic model is crucial to ensure its overall reliability, and thus should be independent of the score(s) optimised during the fitting stage.

The most commonly used global score to optimise the fitting is the cross-correlation coefficient (CCC) between the cryo-EM map and the simulated density of the fitted model. Apart from some variations of the CCC with masks and filters, there are other global scores, such as the mutual information[5]. Local scores are very useful in identifying the regions of poor fit in the models, which can be further refined to obtain a better fit. Local mutual information, TEMPy local scores-SMOC[6] (Segment-based Manders' Overlap Coefficient) and SCCC[7] (Segment-based Cross Correlation), Q-scores[8], and EMRinger[9] can guide the fitting at different structure levels, such as residues, domains, secondary structure elements, and loop regions. Additionally, there are other metrics that assess the geometry of models, such as MolProbity[10] and CaBLAM[11]. These metrics, however, do not include the assessment of the quaternary structure in terms of quality of the interface between subunits.

Some of the common scenarios that may result in sub-optimal protein–protein interfaces in cryo-EM derived models are as follows:

- fitted models are usually built sequentially, *i.e.* one chain is fitted into the map at a time, independent of the others;
- map segmentations are an integral part of model building, but segmentation techniques are not accurate enough to identify boundaries between the subunits;
- building the model of only one protomer and applying symmetry operations; and
- integrating models of subunits built in maps reconstructed by refinement focused on certain segment(s) of the macromolecule.

The features that characterise the interfaces can be used to build a model quality assessment metric. While being density-independent, this metric may provide a complementary quality measure of modelled assemblies in cryo-EM maps, especially for cases such as those listed above. Some of the features that have been shown to be discriminatory in identifying biological (*'native-like'*) interfaces are, conservation of residues at interface[12–16], shape[17,18] and electrostatic complementarity[19,20], residue contact pairs[21,22], types of interactions[23–26], and interface size and area[25,27–29].

Although the interface features listed above are useful for identifying the quality of interfaces, these derivations rely on the dataset of protein complexes used in the study and ignore one major aspect, that is, the extraction and reuse of the knowledge from different datasets. Machine learning (ML)-based approaches, on the other hand, are inherently data-centric, and can accumulate knowledge from various datasets. ML is a class of algorithms that learns from the data and are *trained* on several datasets prior to using them on real datasets (inferencing).

A number of approaches have utilised ML methods for predicting protein–protein interactions, which can vary in terms of exact algorithms used, datasets (*i.e.*, protein–protein complexes), and more importantly, on the set of features used for training. The most commonly used features for predicting protein–protein interactions using ML-based methods include physicochemical properties, evolutionary features, secondary structures, solvent-accessible area and binding energies among others. The choice of algorithms for training ML models include support vector machines (SVM), random forest (RF), neural networks (NN)[30]. Combination of different features and ML algorithms lead to a very rich set of methods that one can rely on. Recent reviews[30,31] provide an elaborate comparison of structure- and sequence-based existing methods, detailing the performance and availability of these techniques.

In this article, we present a systematic assessment of protein–protein interfaces in cryo-EM derived assemblies using a new metric called Protein Interface-score (PI-score). The score was developed based on various features describing protein–protein interfaces in high-resolution crystal structures from the Protein Data Bank (PDB). These derived features were further used to train a ML-based classifier in order to distinguish 'good' (native/native-like) and 'bad' interfaces. To assess the applicability and performance of the trained model to cryo-EM derived assemblies, we used PI-score to assess the quality of interfaces in CASP-13 cryo-EM targets, EM model challenge targets (2016 and 2019), and PDB entries associated with Electron Microscopy Data Bank (EMDB) (4913, as of Aug 2020). A combined score incorporating PI-score and fit-to-density score showed discriminatory power, especially in the cases where there is a disagreement between these two scores.

Results

In this section, we discuss the workflow of building a training and testing dataset for a machine learning (ML)-based model to assess protein–protein interfaces. The derived score (PI-score) is then applied to assess the quality of interfaces in models submitted for CASP13 targets (https://predictioncenter.org/casp13/), EM model challenge targets (https://challenges.emdataresource.org/), and PDB entries (https://www.rcsb.org/) associated with EMDB (https://www.emdataresource.org/). We discuss the examples from each of these datasets. We also compared the performance of PI-score with density based-scores and statistical interface potentials.

Building the dataset. A total of 3926 high-resolution complexes obtained from PDB[32] were subjected to an *in-house* python script to assign the interfaces using a distance-based threshold ('Methods'). To avoid the over-representation of similar interfaces in the dataset, structurally similar interfaces within a

quaternary structure were filtered out using interface similarity score calculated with *iAlign*[33], resulting in 2858 interfaces from 2314 complexes. Various interface features, namely: number of interface residues, contact pairs, surface area, shape complementarity, number of hydrophobic, charged, polar and conserved residues at the interface and other interface properties evaluated by PISA (protein interfaces, surfaces and assemblies), were computed for the dataset. These features were successfully calculated for 2406 interfaces, which form 'positive dataset 1' (PD1, see 'Methods').

To train the ML classifiers of our choice on data closer in quality to models fitted on cryo-EM maps (especially at intermediate-to-low resolutions), noise was added to PD1 by slightly perturbing the relative positions and orientations of the interacting subunits. This was performed using a protein–protein docking method[34] and then selecting the poses with high fraction of aligned interface residues/interface residues in native complex (f_{Nal}), and low interface RMSD (iRMSD) (see 'Methods' for cut-offs). This set, which contains 3743 interfaces, is referred to as 'positive dataset 2' (PD2).

A 'negative dataset' (ND), containing 3578 interfaces, was also derived using docking and includes complexes in which the interfaces are structurally different, i.e., 'far' from native interfaces (low f_{Nal} and high iRMSD, see 'Methods' for cut-offs). The schematic of the procedure to collate the datasets and workflow is summarised in Fig. 1.

Ranking of interface features. Various interface features (listed in 'Methods') were computed for the above-described datasets. As the number of derived features (12) was manageable and computationally not very expensive, we used all the features to train our classifiers. To identify the top-ranking (or most influential) interface feature(s), we ranked them using different methods namely, Ridge, Random Forest, Recursive Feature Elimination, Linear Regression and Lasso. Our exploration showed that the top-ranking features were shape complementarity, number of polar interface residues, number of charged interface residues, and interface solvation energy (Fig. 2a).

Training the classifier and cross-validation. To develop a better understanding of the performance, we evaluated ML classifiers, namely, support vector machine (SVM), random forest (RF), plain vanilla neural network (NN), or simply, multi-layer perceptron (MLP) and gradient boost (GB) using the Scikit-learn Python package[35] (*scikit*).

We used the following combination of datasets described above to train two high-level models, namely, Model A and Model B (referred to as models henceforth), using the interface features ('Methods') (Fig. 1). Each of these models relies on different classifiers, described above (SVM, RF, NN and GB).

Model A: Positive and negative datasets derived using docking (PD2 and ND, respectively). Model B: Positive dataset constitutes high-resolution complexes and computationally derived docked complexes (PD1 + PD2) and ND as negative dataset.

While training and testing both models (Model A and Model B), to minimise the bias of the classifiers, which can easily become an issue with unbalanced datasets, the dataset was

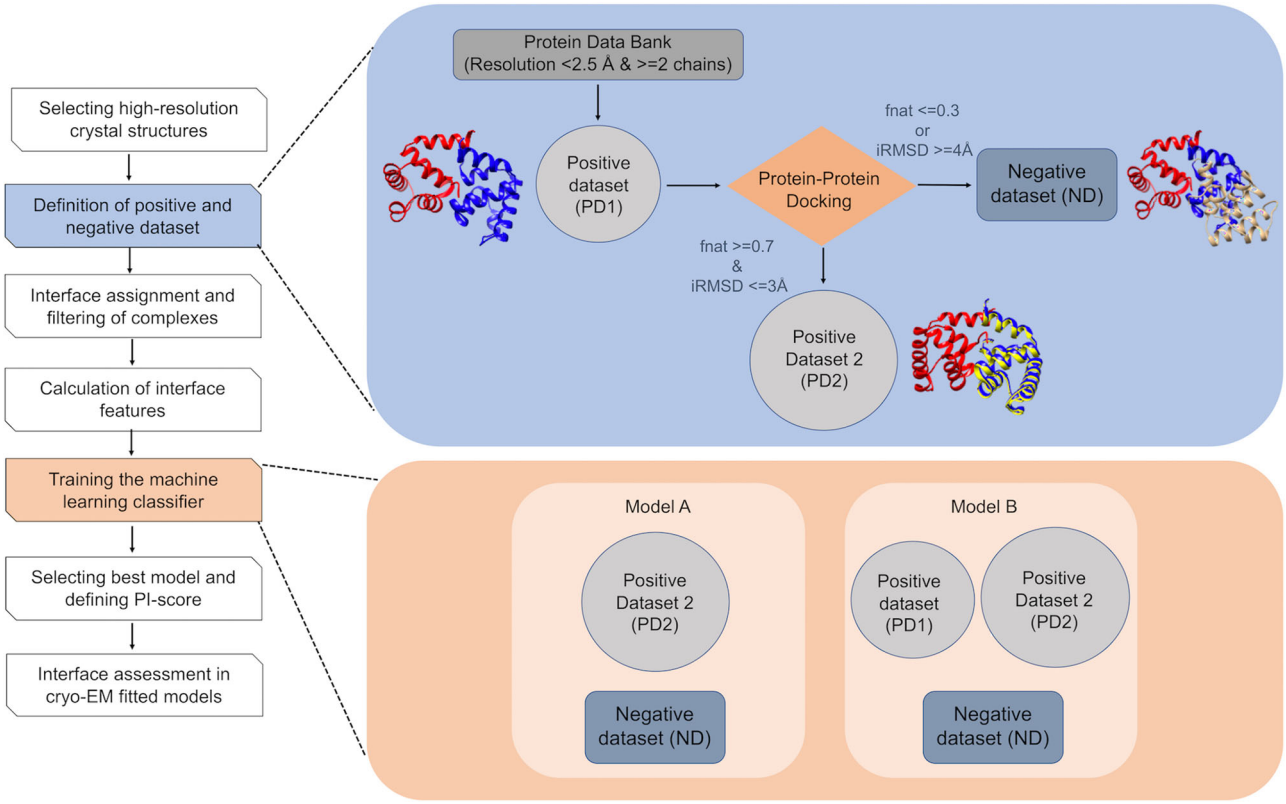

Fig. 1 Workflow for developing a protein–protein interface-based score (PI-score) to assess macromolecular assemblies derived using cryo-EM. High resolution complexes (with > = two chains) were obtained from the PDB and are referred to as the 'positive dataset 1' (PD1). Protein–protein docking was used to derive structurally close (to PD1) complexes that form the 'positive dataset 2' (PD2). The complexes obtained upon docking that have a higher interface RMSD (iRMSD) and lower fraction of aligned native residues (f_{Nal}) at the interface form the 'negative dataset' (ND). Interface features are calculated on all the complexes and are used as an input to train a supervised machine-learning classifier, which is further used to predict the class labels of the benchmark dataset.

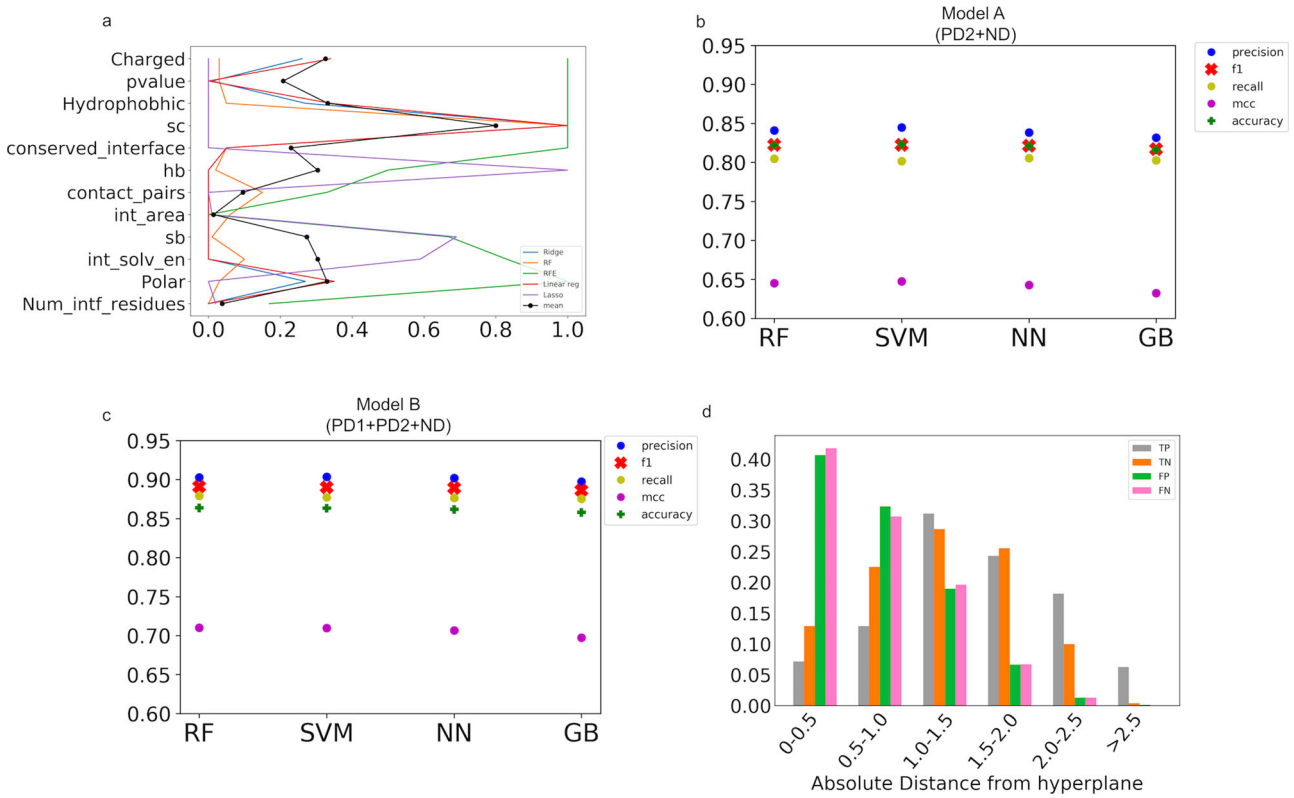

Fig. 2 Machine learning-based classifier to assess the quality of protein–protein interfaces. a Importance of interface features in distinguishing the 'native-like' interface. The ranks calculated using different methods (Ridge, Random Forest (RF), Recursive feature elimination (RFE), Linear regression (Linear reg) and Lasso) were normalised between 0 and 1 and the mean feature rank is plotted in black. **b**, **c** Performance of different classifiers on the training dataset: RF (random forest), SVM (support vector machine), NN (neural networks), and GB (gradient boost) are used to perform supervised learning on the training dataset using *stratified shuffle split* as a means of cross-validation with ten splits. The performance is evaluated using accuracy, precision, F1, recall scores and Matthews correlation coefficient. Performance measures of Model A (**b**): trained on docking-derived positive dataset (PD2) and negative dataset (ND). Performance measures of Model B (**c**): trained using both high-resolution and docking-derived positive datasets (PD1 + PD2) and negative dataset (ND). **d** Fraction of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in different PI-score thresholds. The fractions (*Y*-axis) are averaged over the ten splits (*stratified shuffle split*) of the data. The different PI-score thresholds (*X*-axis) are indicated in absolute values.

split into training set (70%) and test set (30%) using *stratified shuffle split (scikit learn)* and a ten-fold cross-validation was performed. In both scenarios, the performance, which was measured based on ML- and classifier-specific metrics, namely, accuracy, precision, recall, F1 and Matthews correlation coefficient (MCC), was comparable between the three methods (Fig. 2b and c).

The different classifiers gave a comparable performance and SVM-trained model, with a validation accuracy of 86% (Model B), was selected to assess the quality of protein–protein interfaces modelled in cryo-EM maps. The SVM classifier finds a hyperplane that maximises the inter-class variance, and enables the use of the distance of a given data-point from this hyperplane as the machine learning-based score (PI-score) for a given prediction (interface). The farther a point (interface) is located from the hyperplane (more negative or positive), the more confident is the prediction using the SVM model[36]. We assessed the performance of the PI-score at different thresholds by analysing the number of false positives (FP) and false negatives (FN). For the ten test sets (30%) obtained using the *stratified shuffle split* (for cross validation purpose), the fractions of FP (41%) and FN (42%) were observed to be highest in the PI-score ranges of (0 to 0.5] and (−0.5 to 0], respectively (Fig. 2d). We also estimated the measures of performance in different PI-score thresholds and observed that the PI-scores >1 and < −1 (for the positive and negative class label, respectively), were more reliable, based on the low false positive rate (FPR) and high true positive rate (TPR) in the respective bins (Table 1).

Application to CASP targets (high resolution targets). We applied the above-trained models to make predictions on the quality of protein–protein interfaces in cryo-EM targets from the CASP13 competition[37]. Three of the targets (T1020o, T0995o, and T0984o) were classified as 'easy targets', with many high-accuracy models deposited by the participating groups and were also evaluated for the goodness-of-fit to the experimental cryo-EM maps[38]. For each of these targets, a submitted pool of models, an experimentally solved structure (target) and a density-based score for assessing the goodness-of-fit are available from the CASP13 website. Therefore, these targets form an ideal dataset for assessing the performance of PI-score.

We used the CASP multimeric scores (https://predictioncenter.org/casp13/multimer_results.cgi), namely, F1, Jaccard index, lDDT(oligo) and GDT(o) (see 'Methods': comparison with CASP13 oligomeric scores) to define true positives (TP), true negatives (TN), FP and FN for CASP targets[39]. If any of the four CASP13 multimeric scores was equal or greater than (>=) 0.5, and the model was scored positive by our classifier, it was treated as TP. TN were defined as model structures that did not have any

Table 1 Performance in different bins of the scores using the SVM machine learning-based classifier.

Scores' bins	TPR (True Positive Rate)	FPR (False positive Rate)	Precision	Specificity
(−/+)[0.0 to 0.5]	0.15	0.76	0.15	0.24
(−/+) [0.5 to 1.0]	0.29	0.58	0.29	0.41
(−/+) [1.0 to 1.5]	0.61	0.39	0.63	0.61
(−/+) [1.5 to 2.0]	0.78	0.21	0.78	0.78
(−/+) [2.0 to 2.5]	0.93	0.11	0.93	0.99
>=2.5 and < = −2.5	1.0	0.18	0.97	0.81

The following bins according to the listed thresholds and the measure TPR, FPR, precision and specificity are averaged values over the test datasets obtained from ten-fold cross-validation:
0.0–0.5: True positives present in the score range of 0.0 to 0.5 and true negative in score range of −0.5 to 0.0. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score assigned between 0.0 and 0.5 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score between −0.5 and 0.0.
0.5–1.0: True positives present in the score range of 0.5 to 1.0 and true negative in score range of −0.5 to −1.0. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score assigned between 0.5 and 1 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score between −0.5 and −1.0.
1.0–1.5: True positives present in the score range of 1.0 to 1.5 and true negative in score range of −1.0 to −1.5. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score assigned between 1.0 and 1.5 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score between −1.0 and −1.5.
1.5–2.0: True positives present in the score range of 1.5 to 2.0 and true negative in score range of −1.5 to −2.0. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score assigned between 1.5 and 2.0 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score between −1.5 and −2.0.
2.0–2.5: True positives present in the score range of 2.0 to 2.5 and true negative in score range of −2.0 to −2.5. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score assigned between 2.0 and 2.5 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score between −2.0 and −2.5.
>=2.5 and < =−2.5: True positives with a score > = 2.5 and true negative with score < = −2.5. False positives are the complexes from negative dataset (ND) that are predicted positive with a PI-score >2.5 and false negatives are positive complexes from either positive dataset 1 or 2 (PD1 or PD2) that are predicted as negative with the PI-score < −2.5.

of the CASP13 scores > = 0.5 and were scored negative by the classifier. The models which were scored > = 0.5 by any of the four CASP13 scores and negative using our classifier score were defined as FP and models which were scored negative by the classifier but had at least one of the CASP13 score > = 0.5 were FN.

For the target T1020o, 3.3 Å resolution homo-trimer structure of an S-type anion channel from *Brachypodium distachyon*, nine of the assessed 111 submitted models (with 329 interfaces) were predicted to have at least one 'negative' interface (negative PI-score) in the complex. These nine models were also scored low on the CASP multimeric assessment scores[39] (Supplementary Table 1). With a more systematic comparison of PI-scores against the oligomeric assembly assessment scores from CASP13[39], we achieve 82% accuracy for this target (see 'Methods': comparison with CASP13 oligomeric scores).

All interfaces in the target structure (Fig. 3a) and in the top-ranked model based on the cross-correlation of the model with the cryo-EM density (CCC) (TS004_2o, Fig. 3b) have positive PI-score. Out of the nine negatively-scoring models, TS008_4o and TS135_3o have negative PI-score for all the three interfaces (Fig. 3c and d, respectively). When these models are compared to the target structure, all three interfaces have high iRMSD and low fraction of aligned native residues (average iRMSD of 2.93 Å and 3.33 Å for TS008_4o and TS135_3o, respectively, Table 2). For model TS208_1o, two of the interfaces (formed by chains, AC and BC) have negative PI-score (Fig. 3e). PI-score was not calculated for the third interface, as only 9 residues in chain A and 8 residues in chain B are forming the interface in this case, which is less than our cut-off for defining an interface (see 'Methods').

For the model TS208_1o (CCC = 0.34, target structure CCC = 0.77), we generated density maps at resolutions lower than the target map: 5, 8, 10 and 12 Å using the *low pass filter* utility in CCP-EM suite (https://www.ccpem.ac.uk/). Since the CCC does not have a defined absolute cut off value to differentiate between good and bad fits at any given resolution, it is difficult to identify 'target-like' models (Supplementary Table 2). On the other hand, PI-score, which is a density-independent metric, can be very useful to distinguish 'target-like' interfaces in the modelled complex(es).

The PI-score for the target structure, T0995o, a 3.15 Å resolution homo-octamer (A8) of cyanide dehydratase, was positive for the dimer interface (Supplementary Fig. 1a). We calculated the PI-score for 657 interfaces in the 118 CASP13 models for this target and assessed the quality of the dimer interface between all subunits. 123 interfaces in 37 models were observed to have negative PI-score. The top-ranked model (after target) in terms of CCC was TS008_2o (Supplementary Fig. 1b), which is calculated to have a positive PI-score for the equivalent dimer interface (iRMSD = 1.55 Å). Examples for the models with a negative PI-score are TS117_1o (iRMSD = 4 Å, Supplementary Fig. 1c) and TS008_5o (iRMSD = 2.76 Å, Supplementary Fig. 1d). The models with interfaces having negative PI-score using our classifier were also scored low for the CASP13 multimeric scores (Supplementary Table 1).

By comparing it with the multimeric scores in CASP13, we achieve an accuracy of 67% for this target. This target has higher stoichiometry and more interfaces than T1020o, and therefore it is expected to achieve a lower accuracy against the CASP13 assembly scores, which are calculated per complex (while our classifier is per interface and hence this may not be a direct comparison).

For the target T0984o- 3.4 Å dimer of a calcium channel, 145 models were assessed, and all were observed to have a positive PI-score for the interface (Supplementary Data 1).

PI-scores for the assessed interfaces in models for CASP13 cryo-EM targets are provided in Supplementary Data 1. Given the nature of CASP experiments where the participating groups model the complexes without the knowledge of cryo-EM map, protein–protein interface assessments such as PI-scorecan provide additional insights into model quality.

Application to EM model challenge. Next, we calculated PI-scores for the models submitted for the targets from two EM validation challenges (https://challenges.emdataresource.org/), namely, 2016 EM model challenge and 2019 model metrics challenge (Supplementary Data 2).

Target T0002 (from model challenge 2016) is a 3.3 Å resolution cryo-EM map of the 20S proteasome (EMD-5623). We assessed the ten submitted models (with 175 interfaces) based on the interfaces in the target structure (PDB ID: 3J9I). In three of the models- EM164_1, EM189_1 and EM189_2, there was at least one interface that obtained a negative PI-score.

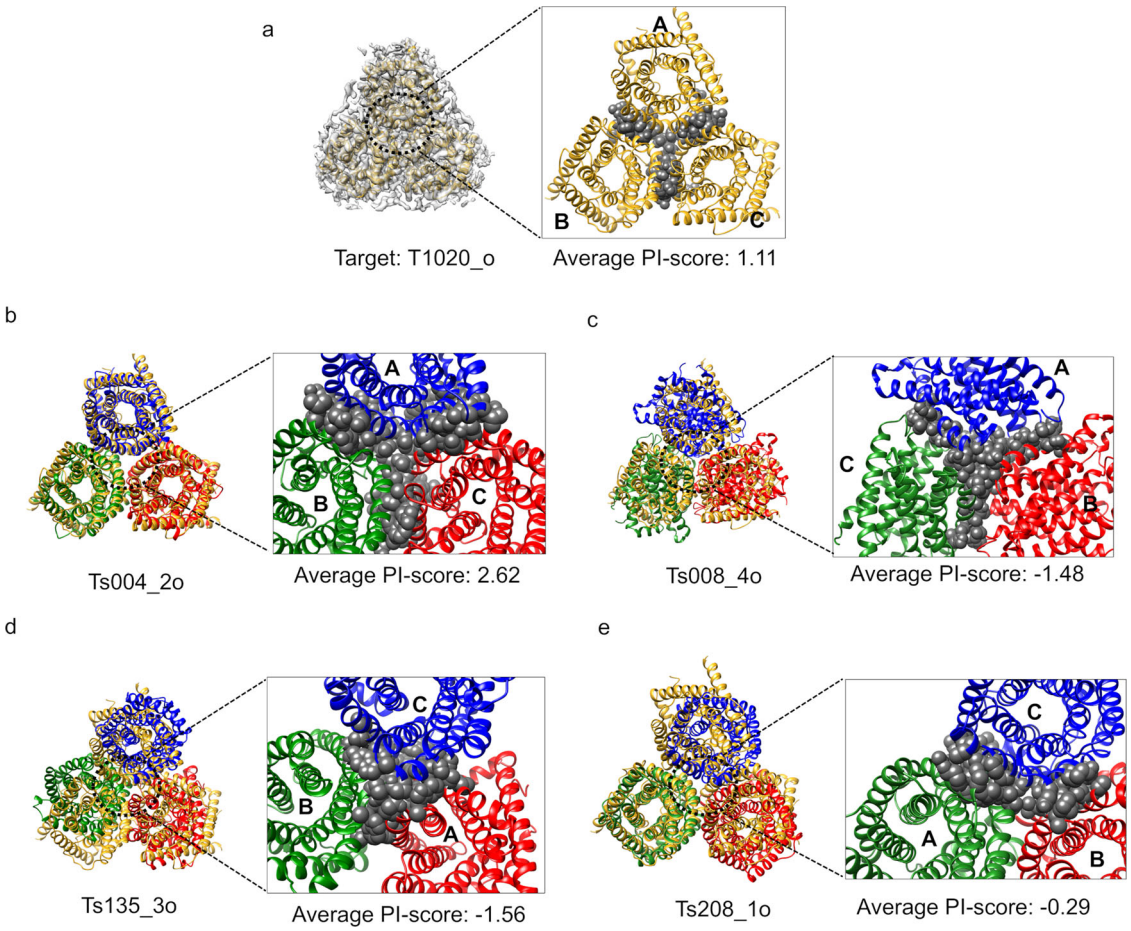

Fig. 3 Scoring the interfaces in the oligomeric CASP13 target T1020o. The target structure is shown in gold in all the panels and the model structures being assessed are shown in red, green and blue. The chains are labelled accordingly. **a** Target structure within the cryo-EM map. The interface residues from the three chains are shown as grey spheres. **b** Model TS004_2o, with a positive PI-score for all the three interfaces in the trimeric assembly. **c**–**e** Models TS008_4o, TS135_3o and TS208_1o, respectively, for which interfaces are scored negatively with PI-score.

Table 2 Assessment of interfaces in the models of CASP13 cryo-EM target T1020o.

Model ID	Model interface	Target interface	iRMSD (Å), f_{nal}	Predicted class	Score
TS004_2o	AB	AB	2.2, 0.81	Positive	2.6
	BC	BC	2.5, 0.75	Positive	2.6
	AC	AC	2.1, 0.82	Positive	2.7
TS008_4o	AB	AC	3.16, 0.42	Negative	−1.5
	BC	BC	2.82, 0.48	Negative	−1.5
	AC	AB	2.81, 0.48	Negative	−1.5
TS135_3o	AB	BC	3.08, 0.56	Negative	−1.6
	BC	AB	3.4, 0.52	Negative	−1.6
	AC	AC	3.51, 0.6	Negative	−1.6
TS208_1o	AB	BC	2.6, 0.63	Not ranked (Interface residues from model 9 and 8)	NA
	BC	AC	2.5, 0.54	Negative	−0.2
	AC	AB	2.6, 0.52	Negative	−0.39

The model and equivalent target chains forming the interface are listed along with the interface RMSD (iRMSD), fraction of aligned native interface residues (f_{Nal}) and predicted class using our model.

As an example, we chose model EM164_1, for which most the interfaces in the alpha and beta subunits were scored negative (Supplementary Data 2). In the alpha ring, the two subunits in the model (chains F and C, shown in red and green Fig. 4a) were scored negative by our classifier (PI-score: −2.27). The interface conformation is slightly different as compared to the target structure (iRMSD = 0.86, f_{Nal} = 0.54). This interface is loosely packed (Fig. 4) and smaller than the equivalent interface in the target structure (23 interface residues in model and 37 in the target structure). Due to its small size the iRMSD is low, and therefore is not a good indicator of the quality of the modelled interface in this case. Due to the offset in the modelled interface, the shape complementarity at the interface drops significantly to 0.32, as opposed to 0.73 for the interface in the target structure. We also checked the multimeric scores from CASP assessment and EM164_1 is scored high for QS-global (0.88) and lDDT

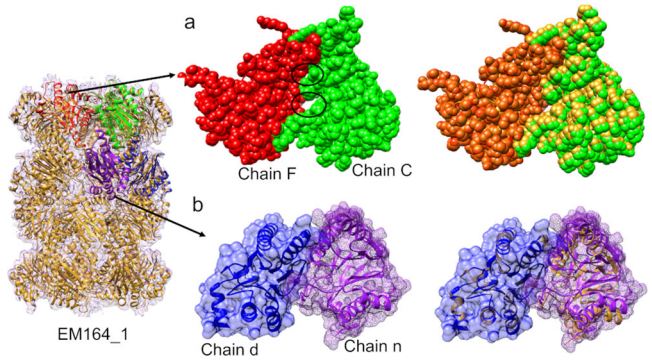

Fig. 4 Scoring the interfaces in the target T0002 from 2016 EM model challenge. a Scoring the interfaces between the alpha subunits' ring in the model structure EM164_1, the chains (F and C), which form a negatively-scoring interface is shown in red and green and the target structure is shown in golden. The surface for the interface forming chains (F and C) are shown as spheres and the loose packing at the interface is marked with black ovals. **b** Scoring the interfaces in the beta subunits' ring of the 20S proteasome, one of the targets in EM model challenge (T0002). The target structure is shown in golden and the chains forming the interface (n and d) in the model structure (EM164_1) being assessed are shown in blue and purple. The surface (mesh) of the chains forming interface are shown to highlight the clashes at the interface formed by chains n and d in the model.

(0.98)) scores but these scores reflect the quality of a multimeric structure as a whole rather than per interface. Other interface assessment scores (from CASP13) such as F1 and Jaccard index, which are calculated per interface, are not reported in the EM Model challenge 2016 website for this model.

Structurally equivalent subunits in the target structure (chains P and Q) have a CCC of 0.85 and the model subunits (chains F and C) have a CCC of 0.73 (Supplementary Table 3) and local score (SMOC) averaged over interface residues are 0.73 and 0.23 for target and EM164_1, respectively. Our model has rightly predicted this interface as 'negative' as reflected by the loose packing at the interface and lower local density-based score for the modelled interface.

Further, we calculated the density-based scores (global and SMOC) at different resolutions (map simulated using *low pass filter utility* in CCP-EM, Supplementary Table 3). The scores assessing the fit of the model (with interface offset) are comparable at resolution worse than 5 Å. Therefore, especially at intermediate-low resolution, our proposed density-independent PI-score can be a crucial model validation tool.

The interfaces between the beta-subunits were also scored negative (TS164_1) by our classifier. This is reflected by the presence of steric clashes at the interface (blue and purple in Fig. 4). The clashes present at the interface resulted in lower shape complementarity score for the interface in model (0.28) as opposed to a higher score (0.62) for the equivalent interface in the target structure. The subunits (chains X and Y) have a CCC of 0.85 whereas the subunits (chains n and d) of model have a CCC of 0.65. This model interface also has a much lower SMOC score than the equivalent interface in the target at all resolutions (Supplementary Table 3).

Recently, model metrics challenge (2019) was open, and we applied our score for assessing the only multimeric target -T0104 (Horse liver alcohol dehydrogenase, 2.9 Å, dimer). We assessed the reference structure (PDB ID: 6NBB) and 17 submitted models using PI-score. Two models (T0104EM060_1; PI-score −0.31 and T0104EM060_2; PI-score 0.13), were scored low (Supplementary Data 2), which is in agreement with the CASP multimeric scores (QS and lDDT scores, https://challenges.emdataresource.org/?q=model-metrics-challenge-2019).

Application to fitted entries in EMDB. We divided this dataset into three sets: high resolution (better than 4 Å, high resolution), 4–8 Å (intermediate resolution) and 8–12 Å (low resolution). As we have described above the performance of PI-score using high-resolution complexes from CASP and the EM model challenge targets, in this section we will focus more on selected examples from intermediate and low resolution cryo-EM maps. The fitted models were also compared with the interfaces in the corresponding crystal structures. For completeness, we also provide the PI-scores of our SVM model for the interfaces fitted at high resolution (better than 4 Å) in Supplementary Data 3.

Chikungunya virus: the available cryo-EM map with an associated fitted model is resolved at 5 Å (EMD-5577, PDB: 3J2W, shown in green and red in Fig. 5a, with interface residues shown as grey circles). The envelope1–envelope2 (E1–E2) heterodimer was observed to have a negative PI-score (−1.67). The available crystal structure (PDB: 3N44, 2.35 Å) for the E1–E2 subcomplex (chains B and F, coloured in gold and interface residues in grey spheres, Fig. 5a) is scored positive (PI-score: 1.67). The interface between E1 and E2 is slightly shifted as compared to the crystal structure (Supplementary Table 4).

We also calculated the density-based scores (global and local) for the E1–E2 subcomplex to assess the fitted model and crystal structure. The E1–E2 subcomplexes from both the fitted model and crystal structure have a CCC of 0.64 and average SMOC over interface residues of 0.72, and hence are indistinguishable with these scores. The plot for the SMOC score (per residue) is shown in Fig. 5a, for both chains and the average SMOC score per chain is shown with a blue dashed line. Interestingly, the interface residues (grey circles) are observed to score higher than the per-chain average, especially for chain B. Therefore, at intermediate resolution interface-based scores such as PI-score can prove useful to distinguish the offsets in the modelled protein–protein interfaces that are indistinguishable with the density-based scores.

PI-scores for the interfaces in the fitted models at the intermediate resolution range derived from EMDB are available as Supplementary Data 4.

TFIID complex: the available cryo-EM map with an associated fitted model is available at 9.8 Å resolution (EMD-9302, PDB: 6MZD, shown in green and red in Fig. 5b, with interface residues shown as grey circles). The interface between subunits 9 and 5 in the fitted model (LF) was scored negative (PI-score: −1.99). This interface is shifted when compared to the corresponding crystal structure (Supplementary Table 4) at 2.5 Å (PDB: 6F3T, chains F and A, shown in golden with interface residues marked as grey circles, Fig. 5b).

Next, we calculated the density-based scores (global and local) to assess the fitted model and crystal structure. The CCC of the fitted model is 0.54 and CCC of the crystal structure is 0.63 upon local optimization of the fit in the map, whereas the average SMOC score over interface residues is 0.84 and 0.87 for the fitted model and crystal structure, respectively (Fig. 5b). In this example, we see again (but this time with low resolution maps) that PI-score can provide additional complementary assessment when density-based scores alone are not sufficient to identify the offsets in the modelled interfaces.

PI-scores for the interfaces in the fitted models at the low-resolution range derived from EMDB are available as Supplementary Data 5.

Application to SARS-CoV-2 cryo-EM derived complexes. We also assessed the fitted models in the SARS-CoV-2 cryo-EM maps

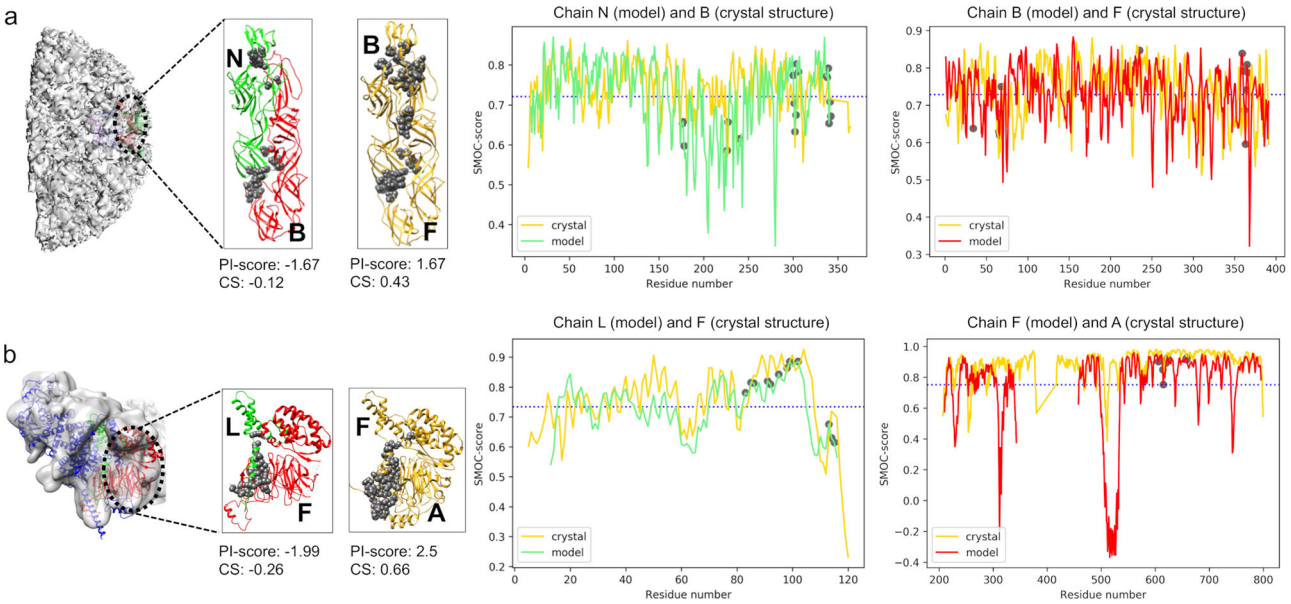

Fig. 5 Application to the fitted models in EMDB at intermediate-low resolution. The chains from the crystal structure are in gold and the chains from the modelled structure are in red and green. The interface residues are shown as grey spheres. The plot of the local density-based score (SMOC) is shown for the chains forming an interface in the model and the equivalent chain in the crystal structure. The X-axis is numbered as per residue numbers in the crystal structure. The average SMOC over the model chain is shown as a blue dashed line. **a** 5 Å resolution structure of Chikungunya virus and the subcomplex envelope1-envelope2 heterodimer (E1–E2) (EMD-5577; fitted PDB: 3J2W). The corresponding 2.35 Å resolution crystal structure is PDB: 3N44 (gold). **b** 9.8 Å resolution structure of the TFIID subunit 5 and 9 sub-complex (EMD-9302, fitted PDB: 6MZD, cyan and green). The 2.5 Å corresponding crystal structure for subunit5-subunit 9 is PDB: 6F3T (gold). PI-score and CS (the weighted combined score, see 'Methods') are listed.

using PI-score. 108 fitted models were downloaded from EMDB: (https://www.ebi.ac.uk/pdbe/emdb/searchResults.html/?EMDB Search&q=text:(ncov%20OR%20SARS-CoV-2). Out of the 108 models, we were able to successfully compute interface features and PI-scores for 55 complexes (149 interfaces). Of these 149 interfaces, 12 were observed to have a negative PI-scores (Supplementary Data 6), with 11 of these being antibody-antibody or protein–antibody interfaces. As our machine learning classifier is not trained on such interfaces (which are reported to have different shape complementarity from other protein–protein interfaces[17]), we decided not to further investigate these cases.

However, the interface between small subunits (S28-S5) of a human 40S ribosome bound to SARS-Cov2 *nsp1* (blue spheres, Supplementary Fig. 2a) protein (EMD-11301, PDB ID: 6ZMT) obtained a negative PI-score of −0.04 (Supplementary Fig. 2b). We next inspected this complex using the validation suite in CCP-EM. The sub-complex S28-S5 was found to have a clashscore of 7.20 with severe clashes reported at the interface. We used *real space refine zone* and *auto fit rotamer*, with backrub rotamers switched on to fix the steric clashes at the interface using Coot[40]. Upon re-refinement in Coot, the clashscore dropped to 6.20 and PI-score improved to 0.25 (Supplementary Fig. 2c). The improvement in PI-scores is most likely due to resolving the clashes between the interface residue pairs R63-A138, V55-34S and L59-R122 (Supplementary Fig. 2b) from chain d and K, respectively.

Comparison with protein–protein interface statistical potentials. Next, we compared PI-score to the existing protein–protein interface-based statistical potentials (PIE[41] and PISA[42]) commonly used for protein–protein docking. PIE and PISA scores provide residue and atomic potentials, respectively, and we also used a combination (0.1*PISA + (−0.8)*PIE + PISA*PIE) of these, which is reported to perform better in identifying 'native-like' complexes[42]. We used the 30% randomly selected test dataset from the entire set (PD1 + PD2 + ND) to calculate the statistical potentials (combined PIE-PISA score) for the interfaces. Different weights for SVM-based score and statistical potentials were tried ranging from 0 to 1, with an increment of 0.1. For this dataset, PI-score separates the ND and PD2 (both derived using docking) better than the combined statistical potential score (Supplementary Fig. 3).

Application of a combined score. As PI-score is a density-independent interface quality measure, providing an additional validation metric for a fitted cryo-EM assembly, we further developed a weighted combined score, which consists of a PI-score term and a quality of fit-in-map term (iCCC) (see 'Methods'). The performance of the combined score was assessed on the pool of models associated with two targets from the EM Model Challenge 2016 (T0002: archaeal 20S proteasome, 3.3 Å and T0003: GroEL, 4.1 Å). The predictions based on the combined score were assessed using iRMSD (Supplementary Fig. 4a and d). We have highlighted two cases in each of these examples where the combined score was proven helpful.

For the target T0002, the majority of the modelled interfaces (83%) with iRMSD <1 Å have a combined score of at least 0.3. For the interface between chains N and M in one of the models for the target T0002 (EM133_1), it is difficult to interpret the quality of the modelled interface, due to the disagreement between the two scores (PI-score is positive while iCCC is negative: 2.1 and −0.04, respectively). Clearly, the interface between chains N and M is modelled reasonably well, with iRMSD = 0.81 (Supplementary Fig. 4b). However, it is fitted poorly within the density (Supplementary Fig. 4c). In this scenario, the combined score of 0.39, which is lower than other good models (<1 Å iRMSD), shown with black outline in Supplementary Fig. 4a), helps to interpret the quality implying that the interface is reasonable despite the poor fit.

For the target T0003, all the modelled interfaces with <1 Å iRMSD were observed to have a combined score of at least 0.5. For the interface between chain E and F in one of T0003 associated models (EM164_1), PI-score indicates that it is modelled with an offset (PI-score = −0.78, iRMSD = 1.73, Supplementary Fig. 4e) but iCCC score is 0.44 (Supplementary Fig. 4f). Modelled interface (EF, EM164_1 for T0003) gets a low combined score of 0.26, and hence provides a better metric combining both the information on quality of protein interface and fit-in-map.

We further applied the combined score for the fitted model and the crystal structure for the complexes described in Fig. 5. For both the examples namely E1–E2 complex of Chikungunya virus (Fig. 5a) and the complex between the subunits 5 and 9 of the TF2D (Fig. 5b), the weighted combined score was correctly able to reflect the fit-in-map and interface quality for the fitted model. For the E1–E2 interface, both the fitted model and crystal structure are scored 0.27 for iCCC whereas the combined score was correctly able to distinguish between the two (−0.12 for the model vs 0.43 for the crystal structure). For the sub-complex between the subunit 5 and 9, iCCC for the fitted model was 0.16 whereas for the crystal structure it was 0.33. The combined score is correctly able to distinguish between the fitted model (−0.27) and the crystal structure (0.66).

Discussion

Density independent PI-score to assess modelled assemblies in cryo-EM maps. Machine learning-based methods trained using interface features have proven to be discriminatory in identifying the 'native-like' complexes, and are routinely used for protein interface sites and hotspots prediction using sequence and structure-based features[30]. So far, such methods have not been applied to the models derived from cryo-EM data, where errors at the interface are likely. Here, we have developed a density independent metric to assess the quality of protein–protein interfaces in cryoEM derived models (PI-score), using a machine learning-based method trained on interface features. We carefully collated high-resolution crystal structures of the protein–protein complexes and annotated them with interface features, which were further used to train a machine learning-based classifier.

In total, 12 features were calculated for 9727 interfaces in our dataset. A 9727*12 vector was used as an input to train a classifier using random forest, support vector machines and neural networks. Shape complementarity at the interface, which is a well-known feature to discriminate 'native-like' complexes[17], was observed to be the most discriminatory feature (see the section 'Training the classifier'). We first tried the combination PD1 as positive and ND as negative set, and we achieved a training accuracy of 96% as it was too simple a problem for a classifier (no noise), therefore did not proceed with this. Using both PD1 and PD2 as positive labels and ND as negative class labels, we were able to achieve a validation accuracy of 86% using a ten-fold cross-validation.

We show that PI-score can help in identifying native-like fits from a pool of candidate models (see sections 'Application to CASP targets' and 'Application to EM model challenge targets').

Importance of interface validation in cryo-EM maps. Most of the structures (~95%) derived using cryo-EM have at least two protein chains. Hence, it becomes crucial to model protein–protein interfaces in such structures.

With the recent advances in technology, single-particle reconstructions are getting to near-atomic resolution, where the modelling of protein–protein interface is becoming more accurate. However, several complexes in the EMDB are in the intermediate-to-low resolution range. The average resolution achieved in 2019 is still less than 5 Å, where the models are likely to be less reliable, especially at regions with less-resolved density. Additionally, there are plenty of maps where the nominal reported resolution is high, but the local resolution varies significantly. In EMDB, we have identified 107 interfaces in 54 complexes (at resolution better than 4 Å), 508 interfaces in 171 complexes (at 4–8 Å), and 51 interfaces in 23 complexes (less than 8 Å), with a negative PI-score, implying potential modelling issues at the protein–protein interface. Investigating these cases revealed that the errors at the interface could be of different types, including steric clashes, loose interface packing, smaller interface size and lower shape complementarity (see section on 'Application to fitted models from EMDB').

Comparison with other scores. As we have demonstrated, PI-score is density-independent and is especially useful to distinguish the native-like interfaces at low-to-intermediate resolution, where density-based scores alone become less informative (see section on 'Application to fitted models from EMDB'). Most studies calculate the global CCC, which will not reflect minor changes at local regions of the structure (such as interface regions). Local scores can be more informative in this respect; however, they require a well-resolved density around the interface. PI-score captures different types of information, specifically assessing interface quality, and therefore brings an added value.

To explore the intention that a combination of both approaches could be extremely beneficial in guiding model fitting and validation (see section 'Application to fitted entries in EMDB') we examined the performance of weighted combined score on some of the targets in the benchmark. We find that our weighted combined score captures the information from PI-score and iCCC and is able to provide a metric for distinguishing 'target-like' models. For example, the combined score specifically will be very helpful in the scenarios where the interface is not correctly modelled (low PI-score) but is still well fitted in the map (high iCCC) as well as the cases where the interface is correctly modelled (high PI-score) but the fit-in-map is poor (low iCCC).

We acknowledge that there is scope for further improvement of the combined score as one should be wary of the limitations. As we have discussed above, CCC is intrinsically dependent on map resolution and there is no recommended/absolute cutoff to determine the quality of fit. This directly influences the performance of the combined score.

Potential usage of PI-score. PI-score has two key uses: to validate and to aid the modelling of interfaces in cryo-EM derived assemblies. We demonstrated its use as a validation score on the CASP and EM model challenge targets. In addition to these, PI-score can also be implemented as part of the model building/refinement process in software packages, such as CCP-EM[43] and Scipion[44], to guide the process of model building and model validation (e.g. as part of the CCP-EM validation suite at https://www.ccpem.ac.uk/download.php). Furthermore, PI-score can also be used to filter solutions and identify 'native-like' interfaces from protein–protein docking software, such as ZDOCK, and from software that use multi-component assembly fitting approaches, such as PRISM-EM[45], IMP[46] and gamma-TEMPy[47].

Summary and future directions. In this work we have introduced an interface-centric metric, PI-score, and systematically assessed the PDB cryo-EM derived assemblies for their interface quality. We are working towards expanding the features' set (e.g. coevolution scores) to calculate PI-score and implementing deep-learning approaches. We believe that PI-score will be a crucial

addition to the set of validation scores currently used in the cryo-EM community as part of structure modelling tools. It is likely that many protein–protein interfaces in future-deposited cryo-EM structures will contain errors, especially in low-to-intermediate resolution structures. Scores such as PI-score, which provide insights into interface modelling, have the potential to be extremely beneficial if included in the EMDB validation report.

Methods

Dataset of high-resolution complexes (positive dataset 1—PD1, 'native' interfaces). High-resolution crystal structures of complexes were obtained from PDB with following filters:

1. Minimum number of chains = 2
2. Experimental method = X-ray
3. Resolution between 0.0 and 2.5 Å
4. R factor (all) between 0 and 0.25
5. R-free between 0 and 0.3
6. Length of each chain >= 30 amino acids

Using these filters, we fetched the non-redundant PDB structures at 40% sequence identity, resulting in a total of 3926 complexes.

Interface assignment. Interface residues between two chains were defined using the distance-based threshold of Cα–Cα distance[48] of 7 Å. An interface was only included in the datasets, if it contained at least ten residues from each of the interacting chains.

The complexes were processed to remove symmetric interfaces present in the same structure using *iAlign*[33] to structurally align the protein–protein interfaces between different chains of the same PDB structure. At the recommended cut-off of interface similarity score of 0.7, non-identical or identical monomers forming similar interfaces were filtered out and the final set contained 2315 complexes with 2858 interfaces.

Dataset of 'near-native' complexes (positive dataset 2—PD2). 'Near-native' complexes were derived from the native complexes (PD1). The pair of interacting chains from PD1 dataset were subjected to protein–protein docking using ZDOCK[34] and the poses (interfaces) with f_{Nal} (fraction of aligned native interface residues) >= 0.7 and iRMSD (interface root mean square deviation)[33] < =3 Å were selected.

Dataset of 'non-biological' complexes (negative dataset—ND). The pairs of interacting chains from PD1 were subjected to docking using ZDOCK and the docked poses with $f_{Nal} < 0.3$ or iRMSD >4 Å were selected.

Calculation of interface features. The following interface parameters were computed:

1. Number of interface residues (num_intf_residues)
   This was calculated using an *in-house* python script to assign the interface as explained above and count the number of residues from each chain of the complex.
2. Conserved residues at the interface (conserved_interface)
   For each chain in a given protein–protein complex, the homologues were collected using *PSI-BLAST*[49] (number of iterations = 3, e-value = 10^{-5}, query coverage = 80%) against Swiss-Prot[50] database. The homologues were further clustered at 90% sequence identity using *usearch*[51], and subsequently aligned using MUSCLE (v3.8.31)[52]. The conservation scores were calculated using the generated multiple sequence alignment as input to the maximum likelihood-based method *Rate4Site*[53], which measures the evolution of amino acids residues and identifies functionally important sites. The intersection of conserved residues and interface residues (as assigned above) were selected as a set of conserved residues at the interface.
3. Charged residues at the interface (charged)
   The charged amino acids (Asp, Glu, Lys, Arg) were counted at the interface and this was normalised by the total number of interface residues.
4. Polar residues at the interface (polar)
   The polar amino acids (Ser, Thr, Asn, Gln, His and Tyr) were counted at the interface and this was normalised by the total number of interface residues.
5. Hydrophobic residues at the interface (hydrophobic)
   Hydrophobic amino acids (Ala, Leu, Ile, Val, Phe, Trp, Cys, Met) were counted at the interface and this was normalised by the total number of interface residues.
6. Number of contact pairs (contact_pairs)
   The contact pairs were defined as the number of atomic contacts between the interface residues from the interacting chains.
7. Shape complementarity (sc)
   Geometric shape complementarity of protein–protein interfaces were computed using the program-SC[17] from the CCP4[54] software suite. The value of the calculated statistic sc (shape correlation) describes the extent of interactions of the two chains with respect to each other and varies between 0–1. Protein–protein interface with sc = 1 suggests that the two protein subunits mesh precisely, whereas with sc closer to zero implies an interface with uncorrelated topography.
   The following features were calculated using PISA[55] (via CCP4) (Protein interfaces, surfaces and assemblies):
8. Hydrogen bonds (hb)
   The number of potential hydrogen bonds at the interface
9. Salt Bridges (sb)
   The number of potential salt bridges at the interface
10. Interface solvation energy (int_solv_en)
    The difference in energy between the bound and unbound monomers due to the solvation effect.
11. Hydrophobic *p*-value (*p* value)
    Probability measure of the specificity of a given interface. The lower the probability is, the more specific the interface is.
12. Interface surface area (int_area)
    Surface area, which becomes inaccessible to the solvent upon interface formation, measured in $Å^2$.

Importance of interface features. Five methods were used to rank the importance of each of the interface features: Ridge, Random forest, Recursive feature elimination, Linear Regression, and Lasso. We used the *sklearn* Python package with default parameter settings. The mean scores from each of these methods were used to rank the derived features.

Performance assessment metrics. True positive (TP), true negative (TN), false positive (FP) and false negative (FN) were used to assess the performance of the model using the following definitions:

$$\text{TPR (True positive rate)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{FPR (False positive rate)} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

$$\text{F1} = 2 * \frac{(\text{Precision} * \text{TPR})}{(\text{Precision} + \text{TPR})} \quad (5)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (6)$$

$$\text{Matthews correlation coefficient(MCC)} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (7)$$

Benchmark datasets.

1. The structures of fitted models and targets, and corresponding target cryo-EM maps were downloaded from the CASP13 website (https://predictioncenter.org/casp13/).
2. The targets' structure, map and submitted models for EM model challenge 2016 and 2019 were downloaded from EM model challenge website (https://model-compare.emdataresource.org/).
3. The entries with fitted models were obtained from EMDB (https://www.ebi.ac.uk/pdbe/emdb/).

Density-based scores. The goodness-of-fit between the model and cryo-EM map was estimated using global and local cross-correlation scores. The global cross-correlation (CCC) was calculated using Fit-in-Map function in UCSF Chimera[56] and the local scores (TEMPy SMOC-Segment-based Manders' Overlap Coefficient[6]) were calculated using the CCP-EM[43] GUI interface. CCC score per residue was calculated using the SCCC formula (Eq. 3 in Joseph et al.[5]). iCCC (interface CCC) is calculated as average CCC over the interface residues.

Comparison with CASP13 oligomeric scores. The models for CASP13 cryo-EM targets which were scored using our classifier were compared with the protein assembly scores used in CASP13. The machine learning-based classifier score (PI-score) for multiple interfaces within a model structure were averaged by the

number of interfaces and this was compared with the CASP13 scores (F1, Jaccard index, lDDT(oligo) and GDT(o))[39]. Interface contact similarity (F1) and interface patch scores (Jaccard coefficient) range from 0 (worst) to 1(best). GDTo and lDDT (oligo) (local distance difference test) consider the whole oligomeric assembly and range from 0 (different quaternary structure) to 1(similar quaternary structure). The latter are computed after mapping the equivalent chains between the target and the model using QS algorithm[39]. If at least one of these CASP13 multimeric scores was > = 0.5 and the model was scored positive using our classifier, it was treated as true positive (TP). True negatives (TN) are the set of model structures that do not have any of the CASP13 scores > = 0.5 and are scored negative by the classifier. False positives (FP) are the models which were scored > = 0.5 by at least one of the four CASP13 scores and negative using our classifier score whereas false negatives (FN) are the models scored negative using the classifier and have at least one of the CASP13 score > = 0.5.

Weighted combined score. As seen from Table 1, |PI-score| >= 2.5, achieves best TPR and is highly reliable, therefore 2.5 was used to normalise the absolute PI-scores. The following conditions were used to obtain the normalised value for PI-scores:

if PI-score >0:

$$normalised\ PIscore = min(PIscore/2.5, 1.0) \quad (8)$$

if PI-score <0:

$$normalised\ PIscore = max(PIscore/2.5, -1.0) \quad (9)$$

if PI-score = 0:

$$normalised\ PIscore = PIscore \quad (10)$$

The value for iCCC score is between −1 and +1 (higher the score, better is the fit to data).

The normalised PI-scores were binned in the range of [−1.0, +1.0] using a step size of 0.1. The equivalent weight matrix was designed with the values ranging between −10 and +10. The weights of normalised PI-scores ($w1$) were the respective indices for the score. Weighted combined score is defined as follows:

$$\text{Weighted combined score} = (w1 * \text{normalised PIscore} + w2 * \text{iCCC})/(w1 + w2) \quad (11)$$

where $w2 = 10$. The weighted combined score varies between −1 and 1 (the higher value the better).

Data availability

The structures and models assessed in this study are deposited in the freely accessible public database PDB and EMDB. The PDB IDs and EMDB IDs are appropriately cited throughout the text. All the data generated during this study are within the article and its Supplementary files.

Code availability

The software to calculate the PI-score is freely available for academic use through: https://gitlab.com/topf-lab/pi_score.

Received: 17 November 2020; Accepted: 14 April 2021;
Published online: 07 June 2021

References

1. Bai, X., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015).
2. Saibil, H. R. Blob-ology and biology of cryo-EM: an interview with Helen Saibil. *BMC Biol.* **15**, 77 (2017).
3. Patwardhan, A. Trends in the electron microscopy data bank (EMDB). *Acta Crystallogr D. Struct. Biol.* **73**, 503–508 (2017).
4. Malhotra, S., Träger, S., Dal Peraro, M. & Topf, M. Modelling structures in cryo-EM maps. *Curr. Opin. Struct. Biol.* **58**, 105–114 (2019).
5. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.* **199**, 12–26 (2017).
6. Joseph, A. P. et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100**, 42–49 (2016).
7. Farabella, I. et al. TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J. Appl. Crystallogr.* **48**, 1314–1323 (2015).
8. Pintilie, G. et al. Measurement of atom resolvability in cryo-EM maps with Q -scores. *Nat. Methods* **17**, 328–334 (2020).
9. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
10. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 12–21 (2010).
11. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink 'waters,' and NGL viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315–329 (2020).
12. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**, 190–202 (2004).
13. Malhotra, S., Sankar, K. & Sowdhamini, R. Structural interface parameters are discriminatory in recognising near-native poses of protein-protein interactions. *PLoS ONE* **9**, e80255 (2014).
14. Valdar, W. S. J. & Thornton, J. M. Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Bioinforma.* **42**, 108–124 (2001).
15. Guharoy, M. & Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl Acad. Sci. USA* **102**, 15447–15452 (2005).
16. Joseph, A. P., Swapna, L. S., Rakesh, R. & Srinivasan, N. Use of evolutionary information in the fitting of atomic level protein models in low resolution cryo-EM map of a protein assembly improves the accuracy of the fitting. *J. Struct. Biol.* **195**, 294–305 (2016).
17. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950 (1993).
18. Norel, R., Lin, S. L., Wolfson, H. J. & Nussinov, R. Shape complementarity at protein-protein interfaces. *Biopolymers* **34**, 933–940 (1994).
19. Tsuchiya, Y., Kinoshita, K. & Nakamura, H. Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Protein Eng. Des. Sel.* **19**, 421–429 (2006).
20. McCoy, A. J., Chandana Epa, V. & Colman, P. M. Electrostatic complementarity at protein/protein interfaces11Edited by B. Honig. *J. Mol. Biol.* **268**, 570–584 (1997).
21. Ofran, Y. & Rost, B. Analysing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387 (2003).
22. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins: Struct., Funct., Bioinforma.* **43**, 89–102 (2001).
23. Conte, L. L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198 (1999).
24. Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159 (2000).
25. Chakrabarti, P. & Janin, J. Dissecting protein–protein recognition sites. *Proteins: Struct. Funct. Bioinforma.* **47**, 334–343 (2002).
26. Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9 (1998).
27. Nooren, I. M. A. & Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492 (2003).
28. Yan, C., Wu, F., Jernigan, R. L., Dobbs, D. & Honavar, V. Characterization of protein–protein interfaces. *Protein J.* **27**, 59–70 (2008).
29. Jones, S., Marin, A. & Thornton, J. M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng. Des. Sel.* **13**, 77–82 (2000).
30. Liu, S., Liu, C. & Deng, L. Machine learning approaches for protein–protein interaction hot spot prediction: progress and comparative assessment. *Molecules* **23**, 2535 (2018).
31. Zhang, J. & Kurgan, L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief. Bioinform* **19**, 821–837 (2018).
32. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
33. Gao, M. & Skolnick, J. iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics* **26**, 2259–2265 (2010).
34. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–87 (2003).
35. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
36. Xia, S., Xiong, Z., Luo, Y. & Dong, L. A method to improve support vector machine based on distance to hyperplane. *Optik* **126**, 2405–2410 (2015).
37. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Struct. Funct. Bioinforma.* **87**, 1011–1020 (2019).
38. Kryshtafovych, A. et al. Cryo-electron microscopy targets in CASP13: overview and evaluation of results. *Proteins* **87**, 1128–1140 (2019).
39. Guzenko, D., Lafita, A., Monastyrskyy, B., Kryshtafovych, A. & Duarte, J. M. Assessment of protein assembly prediction in CASP13. *Proteins: Struct. Funct. Bioinforma.* **87**, 1190–1199 (2019).
40. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 486–501 (2010).

41. Dintyala, R. & Elber, R. PIE—Efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins* **78**, 400–419 (2010).
42. Viswanath, S., Ravikant, D. V. S. & Elber, R. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins: Struct. Funct. Bioinforma.* **81**, 592–606 (2013).
43. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Cryst. D.* **73**, 469–477 (2017).
44. de la Rosa-Trevín, J. M. et al. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).
45. Kuzu, G., Keskin, O., Nussinov, R. & Gursoy, A. PRISM-EM: template interface-based modelling of multi-protein complexes guided by cryo-electron microscopy density maps. *Acta Cryst. D.* **72**, 1137–1148 (2016).
46. Russel, D. et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).
47. Pandurangan, A. P., Vasishtan, D., Alber, F. & Topf, M. γ-TEMPy: simultaneous fitting of components in 3D-EM maps of their assembly using a genetic algorithm. *Structure* **23**, 2365–2376 (2015).
48. Cukuroglu, E., Gursoy, A., Nussinov, R. & Keskin, O. Non-redundant unique interface structures as templates for modeling protein interactions. *PLOS ONE* **9**, e86738 (2014).
49. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
50. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
51. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
53. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**, S71–S77 (2002).
54. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 235–242 (2011).
55. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).
56. Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

Acknowledgements

We thank Prof. Adrian Shepherd (Birkbeck) for the useful discussions and Dr. David Houldershaw (Birkbeck) for the computer support. We also thank Dr. Andriy Kryshtafovych (UC Davis) for the help and advice with CASP and EM model challenge targets. We are grateful for funding from the Wellcome Trust (209250/Z/17/Z and 208398/Z/17/Z). This work was also partially supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the 'AI for Science' theme within that grant & The Alan Turing Institute.

Author contributions

S.M., A.P.J., J.T. and M.T. designed the research. S.M. and A.P.J. wrote the software. S.M. performed the experiments, data analysis and wrote the first draft of the paper. S.M., A.P.J., J.T. and M.T. contributed to paper writing and revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23692-x.

Correspondence and requests for materials should be addressed to S.M. or M.T.

Peer review information *Nature Communications* thanks Lin Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.