

## Research article

# Estimating structure quality trends in the Protein Data Bank by equivalent resolution



Anurag Bagaria, Victor Jaravine, Peter Güntert\*

*Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute of Advanced Studies, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany*

## ARTICLE INFO

## Article history:

Received 27 February 2013

Accepted 29 April 2013

## Keywords:

Protein structure validation

Multiple linear regression

X-ray and NMR

PDB

Structure quality

Equivalent resolution

## ABSTRACT

The quality of protein structures obtained by different experimental and ab-initio calculation methods varies considerably. The methods have been evolving over time by improving both experimental designs and computational techniques, and since the primary aim of these developments is the procurement of reliable and high-quality data, better techniques resulted on average in an evolution toward higher quality structures in the Protein Data Bank (PDB). Each method leaves a specific quantitative and qualitative “trace” in the PDB entry. Certain information relevant to one method (e.g. dynamics for NMR) may be lacking for another method. Furthermore, some standard measures of quality for one method cannot be calculated for other experimental methods, e.g. crystal resolution or NMR bundle RMSD. Consequently, structures are classified in the PDB by the method used. Here we introduce a method to estimate a measure of equivalent X-ray resolution (e-resolution), expressed in units of Å, to assess the quality of any type of monomeric, single-chain protein structure, irrespective of the experimental structure determination method. We showed and compared the trends in the quality of structures in the Protein Data Bank over the last two decades for five different experimental techniques, excluding theoretical structure predictions. We observed that as new methods are introduced, they undergo a rapid method development evolution: within several years the e-resolution score becomes similar for structures obtained from the five methods and they improve from initially poor performance to acceptable quality, comparable with previously established methods, the performance of which is essentially stable.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

The accuracy and quality of a three-dimensional protein structure are important factors deciding its utility. Knowledge of the three-dimensional structure is important for studying a protein's biological role, molecular mechanism, and molecular interactions. The closer an experimentally determined or theoretically calculated structure is to its native structure, the more useful it is for research. For example, it would be nearly meaningless to use a target protein structure for structure-based drug design if we are unsure about the quality of the target protein model. The importance of protein structures and their association with biological regulation has been known for over half a century (Tomkins et al., 1963). This interest was the driving force behind developing and improving methods of structure calculation (Kuntz et al., 1976;

Floudas et al., 2006; Güntert P, 2009; Kelley and Sternberg, 2009). With structures coming from many different experimental and theoretical methods, several methods to assess protein structure quality have been developed, e.g. (Floudas et al., 2006; Sanchez and Sali, 1997; Bhattacharya et al., 2007; Rosato et al., 2012; Moulton et al., 2009; Janin et al., 2003; Laskowski et al., 1993, 1996; Chen et al., 2010; Davis et al., 2004, 2007; Sippl, 1993; Vriend G, 1990; Cristobal et al., 2001; Siew et al., 2000; Eisenberg et al., 1997; Lovell et al., 2003; Huang et al., 2005; Bagaria et al., 2012; Berjanskii et al., 2012). Validation methods using the experimental data generally utilize their own set of specially designed scores. In most cases, these scores cannot be computed for structures obtained by other structure determination methods because the required experimental data are not available. Here we attempt to solve this problem by extending the notion of X-ray resolution toward a more generalized definition based on the linear combination of different coordinate-based validation scores. Restricting the input scores to only those coming from molecular coordinates implies that the e-resolution can be computed for any given protein structure.

About 81,700 protein structures have been deposited in the Protein Data Bank as of February 25, 2013. Most of these structures were determined by the two most popular experimental

\* Corresponding author at: Institute of Biophysical Chemistry, Max-von-Laue Str. 9, 60438 Frankfurt am Main, Germany. Tel.: +49 69 79829621; fax: +49 69 79829632.

E-mail addresses: [anurag.bagaria@bpc.uni-frankfurt.de](mailto:anurag.bagaria@bpc.uni-frankfurt.de) (A. Bagaria), [zharavin@em.uni-frankfurt.de](mailto:zharavin@em.uni-frankfurt.de) (V. Jaravine), [guintert@em.uni-frankfurt.de](mailto:guintert@em.uni-frankfurt.de) (P. Güntert).

techniques: X-ray diffraction (88.8% of the structures) and solution NMR (10.5%). The deposition of structures calculated via “new methods” started only recently: electron crystallography (1991), fiber diffraction (1994), solid state NMR (1997), electron microscopy (1997), and solution scattering (1999). Experimental techniques have evolved over time for all these methods (DiMaio et al., 2011; Joosten et al., 2009; Chen et al., 2012). Sophistication of instruments has allowed performing more complicated experiments with improved efficiency and effectiveness. Comparative modeling of proteins including theoretical modeling have grown and improved rapidly (Sanchez and Sali, 1997; Pantazes et al., 2011). However, even when a structure is determined using the most accurate and established experimental method, we still have the obvious question of whether the structure is correct overall and in all its parts.

For X-ray structures, the most accepted criterion to assess the amount of experimental data is the crystal resolution. However, many other measures like R-factor, B-factors, stereo-chemical parameters etc. are also used for more detailed analyses of the structure. The resolution is formally defined as the smallest distance between structural features that still provide measurable X-ray diffraction and, as a result, can be distinguished from each other in electron density maps (Wlodawer et al., 2008). High-resolution structures have a resolution below 1.8 Å and the ones above 2.7 Å are considered to be of low resolution. The intermediary ones are classified as medium resolution (Minor, 2007). In case of structures obtained via NMR spectroscopy, structural quality is described by the bundle RMSD, the amount of experimental restraints, the number of distance and angle violations, RPF scores (Huang et al., 2012), etc. (Bhattacharya et al., 2007; Doreleijers et al., 2012).

Structure quality assessment criteria using experimental data are different for the major experimental methods. This makes it difficult to compare the quality of structures obtained by the two methods. Even comparing local and global features in a non-redundant dataset containing NMR and X-ray structure pairs of the same proteins revealed systematic (including method-related) differences (Sikic et al., 2010). Similar systematic differences were pointed out (Bagaria et al., 2012) while comparing the structure quality of proteins in the CASP8 (Moult et al., 2009) and CASD-NMR (Rosato et al., 2012, 2009) projects.

There exists a large range of software tools with their respective scores designed to evaluate different quality aspects of protein structures. These are based on the various elements and properties of molecular structure: torsion angles, bond lengths, atom clashes, van der Waals violations, stereo-chemical violations etc. Most of these tools do not provide an obvious scale to easily comprehend the overall goodness of a structure in question. For instance, there have been a number of approaches to convert various measures of NMR protein structure bundles into a single resolution score, using Ramachandran plot quality, ensemble precision, or numbers of NOEs per residue (Kwan et al., 2011). An attempt has been made to develop an intuitive score for the quality of a protein structure in terms of predicting its RMSD from the native structure (Bagaria et al., 2012). Several attempts were also made to develop “equivalent” X-ray resolution for structures based only on coordinate information (Laskowski et al., 1996; Chen et al., 2010), including a “resolution-by-proxy” measure (Berjanskii et al., 2012) incorporating 25 protein structure features.

Here, we propose another definition of “equivalent” resolution that is generally applicable to any protein structure regardless of the method that was used to determine it. We estimate and report the quality of structures obtained over the last two decades by five popular methods: X-ray, solution state NMR, neutron diffraction, solid state NMR, and hybrid methods.

## 2. Materials and methods

### 2.1. Molecular coordinate data sets

For this study we used all PDB entries that contained a single chain of a protein determined by the authors to be monomeric, irrespective of the experimental method they were obtained with. We grouped them by their experimental technique, and divided into sub-groups by year of submission to the PDB. This constituted 22,016 X-ray, 3777 NMR, 18 neutron diffraction, 12 solid-state NMR and 7 hybrid method protein structures. Non-monomeric biological units of proteins, complexes, and some structures for which certain validation scores could not be computed for technical reasons were excluded. Regarding structures solved by electron microscopy, it must be noted that though over 340 structures have been solved by this technique, we had to exclude this method because over 95% of these structures are either not single-chain or not monomeric. Although the restriction to monomeric, single chain proteins excluded many PDB entries, such a large scale study of protein structure quality trends across different experimental fields has not been performed so far. Structures from electron crystallography, solution scattering, and fiber diffraction were omitted because fewer than 4 single chained monomeric structures by these techniques were deposited in the last 5 years. More details regarding data composition may be found in the Discussion section. The time range selected for this study is from the year 1995 to 2012.

Based on earlier reports about the dependence of structure quality on protein size (Bagaria et al., 2012), this fact was presumed and the protein structures were divided into 3 size groups: “S” or “Small” (<100 amino acid residues), “M” or “Medium” (100–400 amino acid residues), and “L” or “Large” (>400 amino acid residues). This choice of segregating structures into size-dependent bins resulted in a clear-cut improvement of their resolution predictions manifested by a reduced mean absolute error (MAE) (see Section 3). For each protein structure, 17 score values from the software tools listed below were obtained. In the case of NMR structure bundles, the scores were calculated separately for each of the top 10 conformers sorted by increasing root-mean-squared-deviation (RMSD) to the mean of the structure bundle.

### 2.2. Validation scores

The following coordinate based validation scores were used to assess the quality of the protein structures. These scores were selected based on their popularity as indicated by the number of publication citations, and the possibility to implement and evaluate them for structures obtained by any method.

The ProsaII scores (Sippl, 1993) are based on the probability for two residues to be at a specific distance from each other. In this validation score the amino acid types, the distance, as well as the sequence separations are used. We used three of the ProsaII scores. Zp-Pair (Z score for pair potential energy), Ep-Pair (pair potential energy based on atom–atom interaction) and Ep-Surf (Surface energy based on atom-solvent interaction).

ProQ is a neural network based predictor (Wallner and Elofsson A, 2003). Based on a number of structural features, it predicts the quality of a protein model. ProQ is optimized to find correct models in contrast to methods which are optimized to find native structures. Two quality measures are predicted, the LGscore (Cristobal et al., 2001) and MaxSub (Siew et al., 2000).

The Procheck software (Laskowski et al., 1993, 1996) takes into account the number of residues in allowed/disallowed areas of Ramachandran plot, the number of unusual bond lengths or bond angles, and so forth. Here we choose two scores from the Procheck software: Core and Gener. The former represents the percentage of

**Table 1**  
The MLR coefficients for e-resolution prediction for three size groups based on 13 selected scores.

Score	S (<100 aa)	M (100–400 aa)	L (>400 aa)
Protein size	−1.39	−5.48	−5.08
ProsallZp-pair	−0.0135	−0.0167	−0.0229
ProsallEp-pair	0.00249	0.000742	0.00105
ProsallEp-surf	0.00332	0.00629	0.00603
LG-score	0.0399	−0.0541	−0.0857
MaxSub	−0.131	0.553	0.655
Procheck core	−0.00018	−0.00391	−0.00412
Procheck gener	−0.00494	−0.0262	−0.0445
Molprob clash	0.00151	0.00418	−0.00099
Molprob global	0.354	0.363	0.417
Verify3D quality	−0.000383	−0.000445	−0.000518
Whatcheck INO	0.59	0.191	0.366
Whatcheck structure Z-score	0.0111	−0.0353	−0.0727
Constant	0.775	2.49	2.33

See Section 2 for details.

residues present in the most favored regions of the Ramachandran plot while the latter indicates the percentage of residues present in the generously allowed regions of the Ramachandran plot for the protein.

The Molprobit program (Davis et al., 2007) calculates a score based on a number of validations including all-residue Ramachandran analysis, rotamer analysis, and all-atom clash analysis. We consider two scores from this program. The Molprobit Clash-score is the number of serious clashes per 1000 atoms and the Molprobit Global score is based on the global quality analysis.

Verify3D (Eisenberg et al., 1997; Lovell et al., 2003) is based on 3D–1D profiles and assigns an environmental class to each residue in a protein. The environments are divided into 18 classes based on the secondary structure, buried area, and the fraction of polar contacts. Next, the probability for each amino acid type to be assigned to each type of environment is calculated. During evaluation of a model, the sum of probabilities over a window, or over the entire protein, is calculated. If the probability is low, it is likely that the model is incorrect.

The program Whatcheck (Vriend G, 1990; Hooft et al., 1996) provides an Inside/Outside (INO) distribution normality RMS Z-score and a Structure Z-score. Hydrophobic residues are expected to be buried. Hydrophilic residues are expected to be exposed. The INO score tests whether a protein is normal in this aspect. It will report a normality RMS Z-score for the whole structure. Inside-out structures, membrane proteins and mis-threaded structures will trigger this check. For each residue the solvent accessibility is calculated. These values are divided by the “vacuum accessibility” of the residue type, resulting in an accessibility fraction. These numbers are sorted from low to high. Using the mean and standard deviation for the location in the array from the WHAT IF database, a Z-score is calculated for each residue. The Z-scores for the residues are used to calculate an RMS Z-score for the structure.

Molecular size shows a correlation with several existing structure quality assessment scores, as was noticed in an earlier work of ours (Bagaria et al., 2012). For all protein structures, irrespective of the five methods that we studied, we see a high dependence of the quality on the molecular size.

Further scores from Prosall (Zp-comb, Zp-surf, and Ep-comb) and the allowed score of Procheck (indicating the percentage of residues in the allowed region of the Ramachandran plot) were also considered but found to have no significant contribution to the predicted e-resolution. They were therefore dropped in the final calculation. The significance of contributions by each of the scores to the predicted resolution was calculated based on its *P*-value and the Akaike information criterion (AIC). A low *P*-value for an individual score indicates that the probability of this

score's contribution in the fit being random is low. AIC provides information about the relative goodness of fit of a statistical model. Scores below a *P*-value of 0.05 were considered for the calculation of the e-resolution. The approach of selecting the scores was similar to our earlier work (Bagaria et al., 2012). Following these steps was necessary to rule out the presence of mutually linearly dependent scores. This is an important consideration while performing a multiple linear regression analysis where linear correlations between two or more independent variables and a single dependent variable are examined. Only those scores were considered which had, according to the above criteria, a statistically significant contribution to the e-resolution prediction and that were also mutually independent.

### 2.3. MLR analysis

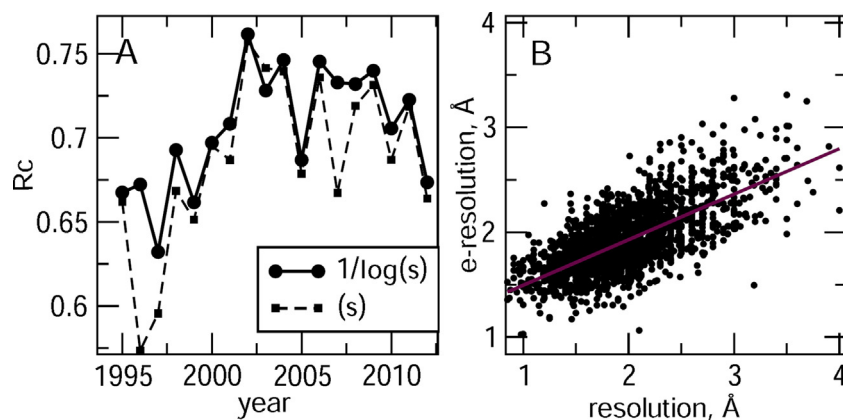
Multiple linear regression (MLR) is a multivariate statistical technique for examining the linear correlations between two or more independent variables and a single dependent variable. Here we consider a linear model by which the predicted equivalent resolution (e-resolution) value  $y_i^{LM}$  for the *i*-th protein structure depends linearly on *m* validation scores  $x_{i1}, \dots, x_{im}$ , each of which describes a particular aspect of structure quality. To calculate the e-resolution  $y_i^{LM}$  for the *i*th structure, the validation score values  $x_{ij}$  are multiplied with its corresponding size dependent coefficient  $b_j$  from Table 1 and the products are summed up, and a constant *a* is added.

$$y_i^{LM} = \sum_{j=1}^m b_j x_{ij} + a$$

The constants *a* and  $b_1, \dots, b_m$  are determined by a linear least-squares fit to the actual resolution values  $y_i$  from a training set of  $i = 1, \dots, n$  known X-ray structures. The fit is performed to minimize the  $\chi^2$  value,

$$\chi^2 = \sum_{i=1}^n \left( y_i - \sum_{j=1}^m b_j x_{ij} - a \right)^2$$

The e-resolution thus represents an approximation of the experimental resolution of the X-ray structures. In addition, here we extrapolate this approximation to other experimental techniques. MLR calculations were performed with the R software environment for statistical computing and graphics (<http://www.r-project.org/>).



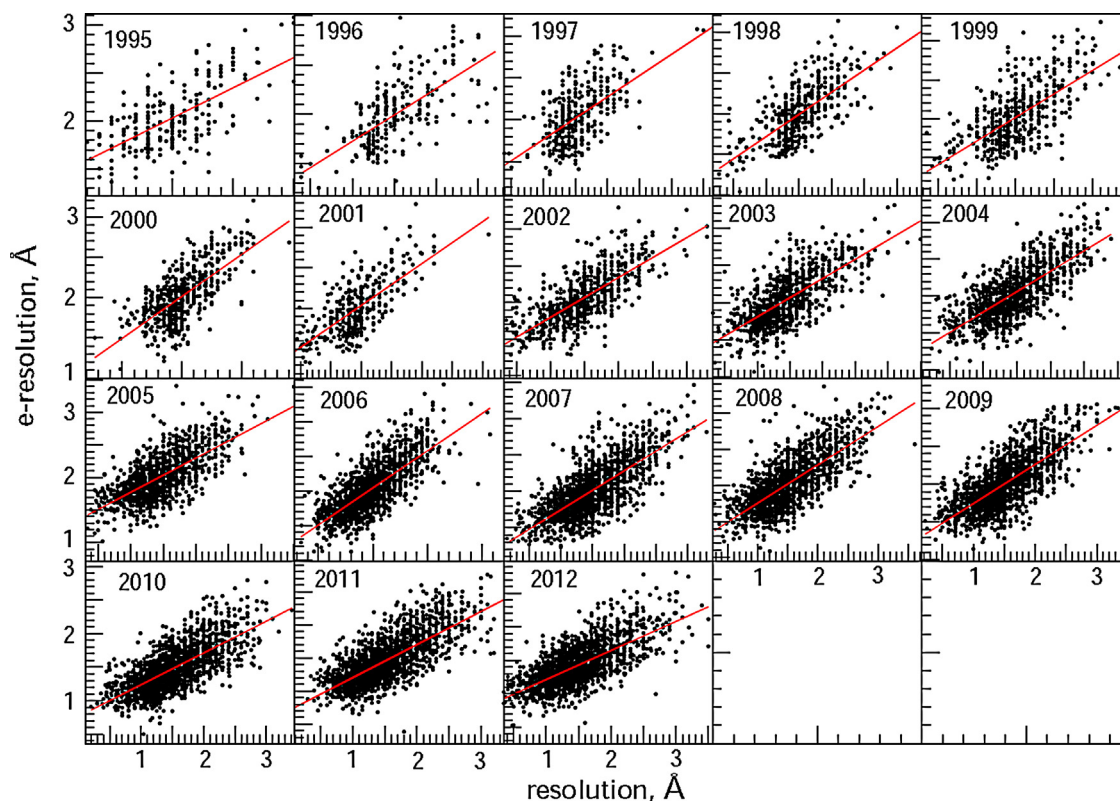
**Fig. 1.** (A) Correlation coefficients between predicted e-resolution and the actual resolution reported in the PDB for the X-ray yearly data from 1995 to 2012. For each predicted year the PDB data of the year itself was not included in the training. The solid lines and dotted lines represent predictions incorporating the set of coefficients when molecular size was trained as a function of “ $1/\log(\text{size})$ ” or a function of “size” terms, respectively. (B) Correlation graph of resolution and its prediction for X-ray structures of the year 2012, with training on structures from the three size-dependent bins (S, M, and L) for years 1995–2011. The plot shows the combined results from all the three bins (S, M, and L).

### 3. Results

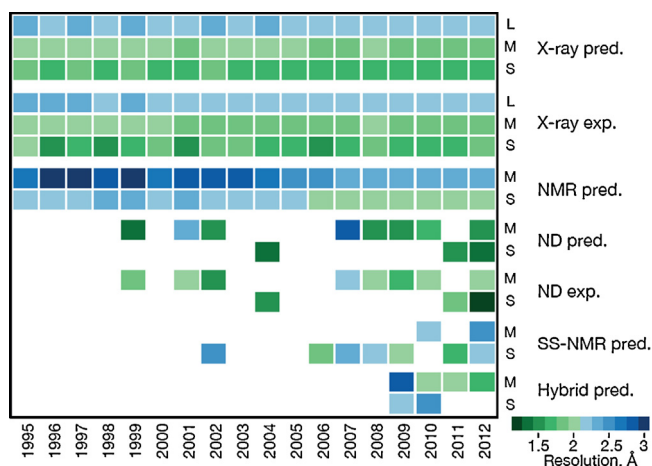
#### 3.1. Training and testing on X-ray protein structures

An initial training on X-ray datasets for different years was performed for each of the “S”, “M” and “L” molecular size groups. For predicting the e-resolution of the structures of a given year, all data of the corresponding molecular size group was used to obtain the set of coefficients, except the data of the year to be predicted (jack knife procedure). This set of coefficients obtained for each size group was then used to predict the resolution of its corresponding

size group (S, M, and L) for the year that was excluded from the training dataset. The procedure was done repeatedly for all years. The correlation coefficients between the actual resolutions of the X-ray structures reported in the PDB and the MLR e-resolution predictions are in the range  $0.70 \pm 0.05$  for the different years (Fig. 1A). Considering the number and variety of X-ray structures involved in each of the training and testing sets, this correlation is remarkable. As an example, Fig. 1B shows the correlation between the experimental and predicted resolution values of the X-ray structures for the year 2012. Distributions obtained for the other years are similar (Fig. 2).



**Fig. 2.** Correlation graph of resolution and its prediction for X-ray structures of the years 1995–2012, with training for all the years excluding the given year itself. The “ $1/\log(\text{size})$ ” parameter was used for these plots.



**Fig. 3.** Time evolution of average resolution of protein structures in the PDB, grouped by the five most used experimental methods (names on the right): X-ray, NMR, neutron diffraction, solid-state NMR, and hybrid methods. The methods are sorted by the total number of structures in PDB from the highest at the top to the lowest at the bottom. On the vertical axis, S, M and L stand for small (<100 amino acid residues), medium (100–400 amino acid residues) and large (>400 amino acid residues) proteins, respectively. The horizontal axis indicates the year when the corresponding structures were deposited in the PDB. The label “exp.” corresponds to experimental resolution (reported in the PDB entry), while “pred.” represents the predicted e-resolution. Training and testing was done for each of the size groups S, M and L separately. “X-ray pred” represents the test set predictions during the cross-validation process. The range of resolution values presented here (1.1–3.1 Å) are the averaged resolutions of the structures in that particular year.

### 3.2. Application of the MLR coefficients to structures from other methods

For each of the size groups S, M, and L a unique set of coefficients was obtained based on the X-ray structures from all years (1995–2012) belonging to the respective size group (Table 1). The set of coefficients for each size group (S, M, and L) was obtained in the following manner. Cross-validation was performed for each instance of training and testing sets. A decent level of prediction was achieved with an average correlation coefficient of  $0.70 \pm 0.05$ . This led into deciding that this linear fit method is suitable for the predictions. A unique set of coefficients for the corresponding size group was then obtained by training on the whole dataset of all years. These sets of coefficients were then used to predict the e-resolution of the structures from their respective size bins. The same set of coefficients for the S and M size groups were used to predict the e-resolution of the structures from other experimental methods in their respective size groups: NMR, neutron diffraction, solid state NMR, and hybrid methods (Fig. 3). More information on the selection of data and the cross-validation follows in Section 4.

Apart from enabling a comparison of the structural quality among these methods, the e-resolution graphs of Fig. 3 shows the evolution for the different methods. Overall, it is clear from the data that the PDB protein structures obtained by all methods in the study have, on average, improved in quality over time as indicated by the values of e-resolution which has constantly improved toward the recent years.

A more detailed analysis of the results shows the following. First, the predictions work correctly: the average experimental and predicted resolutions for X-ray and neutron diffraction methods in Fig. 3 are similar and visually comparable for all size groups. Secondly, for all of the five experimental methods compared in the study, there is a clear dependence of the quality on the protein size. All the structure determination methods show significantly better e-resolution for smaller sized proteins, while larger proteins show a lower e-resolution value. Thus the quality of structures obtained via

all the methods discussed here is dependent, to different degrees, on the molecular size of the protein. Besides, even though experimental resolution for NMR proteins is not defined and thus the estimates cannot be verified directly, the estimates show that over the period the average quality of NMR structures has improved to the levels comparable to X-ray, i.e. 2.3 Å for medium- and 2.0 Å for small-sized proteins. For all methods the yearly-average equivalent resolution rarely exceeds 3 Å. This can be partially explained by the PDB deposition acceptance threshold criteria or few low resolution structures there. It is worth noting that the range of resolution values (1.1–3.1 Å) presented in Figs. 3 and 4 are the average resolutions of all the structures deposited in a particular year, which should not be confused with the range of resolution for individual structures (e.g. 0.5–5.0 Å, or even higher for some structures). Solid state NMR is a relatively new method, which has been applied to a few small proteins and then to medium sized ones, and already shows good quality. Though there are a very few structures solved by neutron diffraction (ND), it shows the best resolution so far. These very few structures also explain the high fluctuations (for ND) and missing data-points (for ND and hybrid methods) in their resolutions presented in Fig. 4.

### 3.3. Dependence of the e-resolution on protein size

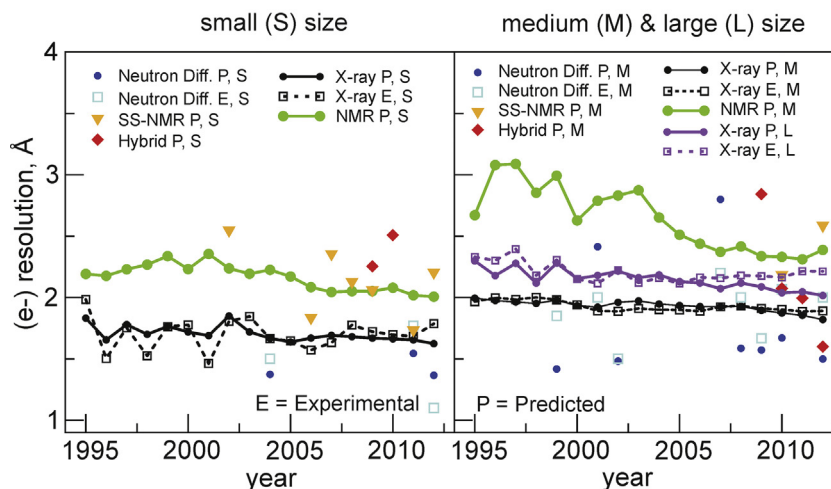
We found that by training and testing the data without dividing them into size dependent bins, the correlation coefficient of prediction (of e-resolution) was considerably lower. The sets of coefficients obtained for specific size bins were used for e-resolution prediction. Segregating the molecules into size dependent bins improved the prediction correlation coefficient from 0.66 to 0.68. As a second step of incremental improvement, the correlation coefficient of predicting the resolution was significantly higher when using “ $1/\log(\text{size})$ ” (solid line in Fig. 1A) instead of using the size directly (dotted line in Fig. 1A). Here the “size” is the molecular size expressed as the number of residues in a protein. The correlation coefficient became more stable (indicated by reduced mean absolute error) and further improved by 0.02 on average that is from  $0.68 \pm 0.08$  to  $0.70 \pm 0.05$ . The overall optimization of the size parameter, as explained in the two incremental steps above, improved the overall predictions by at least 5%, i.e. from a correlation coefficient of 0.66 to 0.70. An empirical relation between the mass of the protein, crystal size, and resolution has been suggested for X-ray structures (Holton and Frankel KA, 2010). Assuming the crystal size as a constant, the relation between protein size and resolution of the structure, suggested in Holton and Frankel KA (2010) holds good in our study too. This again emphasizes the molecular size dependence of the quality of protein structures.

## 4. Discussion

What can be a good measure of the quality of a three-dimensional protein structure, irrespective of the method it was obtained from? For each method, it depends on many parameters, e.g. instrument used, beam intensity, wavelength, crystal size, protein size, etc. Apart from that, certain protein classes, GPCRs for example, can be a challenge even for the state-of-the-art tools to determine protein structures. Considering that all these factors combined are resulting in an uncertainty of the atomic coordinates, the uncertainty gives us a numeric degree of significance of values of atomic coordinates, expressed in Å as the equivalent resolution of the structure.

For several decades X-ray was considered to be a more mature technique able to provide better quality structures compared to solution NMR. Experimental and computational methods for X-ray diffraction have been perfected already long time ago.

## Structure accuracy by different methods



**Fig. 4.** Line plot of the e-resolution for each data type. This is a more detailed representation of the results presented in Fig. 3. S, M, and L denote small, medium and large sized proteins. “exp” and “pred” denote respectively the experimentally reported values and the predicted e-resolutions.

Consequently, the structures obtained from this technique are on average of higher quality. Though X-ray and NMR methods are continuously being improved, the law of diminishing returns kicks in for X-ray structures. Additional time and work result only in small improvements. By now the quality of NMR structures has become good for small and medium sized proteins, for which NMR is able to provide quality comparable to that of X-ray structures. Solid state NMR and hybrid methods emerged recently and already showed noticeable improvement. We see these trends in the results. However, all these methods except X-ray still do not produce the large-size structures. Although neutron diffraction yields high-quality structures, few structures have been actually determined by it. Besides, we found that its performance varied over the years, and overall its quality is comparable to that of X-ray.

The most common measures of molecular structure quality use some sort of comparison of coordinates to a reference structure. Since we cannot have a reference structure for every existing structure, one rather needs an absolute measure that is derived from the structure itself. Thus, relative measures, such as RMSD from reference structure, cannot be obtained for all structures in the PDB. The second most used measure of quality is derived from evaluating the agreement between the experimental data and the resulting structure. Since the nature of the experimental data depends on the method used, we cannot apply these criteria to structures of all methods. Thus, after restricting the choice of scores and not using reference structure in the method, the “e-resolution” can be considered as an absolute quality measure, derived exclusively from structure itself.

It is worthwhile to consider again the systemic differences found in structures obtained via different experimental techniques, X-ray and NMR for example (Bagaria et al., 2012). The range of spreads for different validation scores is different for structures obtained by different methods. The number of protein structures used here is large (22,016), thus covering large ranges of distributions for each of the validation scores. We noticed that the respective validation scores for structures from individual experimental techniques fall within the large distribution ranges here. Therefore, we applied the same set of coefficients, trained on the scores for X-ray structures, to extrapolate the e-resolution predictions to other methods.

To corroborate this further, we performed the cross-validation of the prediction model. Structures deposited in a certain year (test set) were systematically left out and a new linear model was trained

each time by using all the remaining structures (training set). While testing a specific year, the number of protein structures used for each of the training sets was different and ranged between 19,738 and 21,765 structures as the number of structures deposited in the PDB varies over the years. The training sets comprised 90–99% of the whole dataset. Alternatively, Jack-knifing of the data using randomly chosen training and test sets of fixed size could be used. But year-based selection for the training set was preferred here over random subsets of structures, because we aimed at predicting a yearly trend for e-resolution and did not want to include any structures from the year which was to be predicted. Nonetheless, we performed the standard Jack-knifing tests. For example, 50 randomly chosen training datasets for each of 60%, 70%, 80% and 90% of the whole X-ray structure data was cross-validated by testing the prediction of e-resolution on the corresponding remainder test sets of 40%, 30%, 20% and 10% respectively. The overall correlation coefficients of prediction showed small changes. Thus we ruled out any bias caused by training/test set sizes.

Apart from MLR, its variants GLR (generalized LR) and RLR (Robust LR) were tested on the X-ray data, and several other clustering and regression techniques, in an attempt to obtain better accuracy in the predicted values for the e-resolution. These methods included: K-means and K-means++ clustering analysis, K-nearest-neighbor clustering (KNN), and regression trees. These unsupervised methods were tested in order to see if this large set of proteins from the PDB formed any specific number of clusters. The aim was to obtain a minimum number of clusters so that within all clusters the mean absolute error of prediction decreased, thereby decreasing the overall mean absolute error. In that the Euclidean distance of each of the validation scores for each protein structure (cluster point) was minimized from their respective cluster mean. No significant improvement in the predictions was observed, neither in terms of reduction of the mean absolute error of the prediction of the e-resolution, nor in the corresponding correlation coefficient. We also tried to use second order terms for all the scores in addition to the linear ones in MLR, but again no significant improvement was obtained compared to the simple multi-linear method. This consistency may be attributed to the sufficiently large dataset used for this study. Applying Occam’s razor as a heuristic to guide the development of theoretical models by choosing the simplest hypothesis among the competing ones (Myung and Pitt MA, 1997), we present here only the simplest of the tested methods, MLR.

A similar study was performed by Berjanskii et al. (2012) who applied a support vector regression (SVR) method on 25 coordinate-based scores to predict resolution-by-proxy using 2927 protein structures for their predictions. Here we use over 22,000 structures from the PDB, i.e. several times more than in the earlier study. Dividing the training sets into size-dependent bins makes the predictions more robust, as the unique set of coefficients (weights) obtained for each size bin make the predictions more specific with regard to the size. Additionally, we extrapolated the predictions to 3 more experimental methods.

With a rapidly increasing interest and amount of protein structures being solved over the last two decades, several quality assessment initiatives like CASP (Moult et al., 2009), CASD-NMR (Rosato et al., 2012, 2009), and CAPRI (Janin et al., 2003) evolved. Their aim is mainly to promote the calculation of accurate and good quality protein structures via different structure calculation techniques. CASD-NMR for example, has succeeded in revealing the best tools, techniques and practices prevailing in the field of NMR (Rosato et al., 2012). While evaluating protein structure quality using the GLM-RMSD, systematic differences were reported between structures obtained via theoretical and experimental methods (Bagaria et al., 2012). This makes structures obtained via different techniques partially or completely incomparable. Therefore, here we attempt to remedy this situation by moving to a quality measure for a structure that is independent of the experimental method used, which was tested here for five methods. Even though there is still relatively little amount of data for the newest methods, in comparison to X-ray and NMR, the results here show the undeniable trend that quality of protein structures obtained by all methods is improving over time and becoming comparable. Our similar linear regression based study using GLM-RMSD for structure quality evaluation was shown to work well with predicted structures from CASP8, especially by incorporating the DP-score which is calculated using experimental data from NMR (Bagaria et al., 2012). Extending the e-resolution score to theoretically predicted structures is a future scope for this work.

### Authors' contributions

AB and VJ conceived the project, performed the calculations, and analyzed the results. All authors wrote the paper, and read and approved the final manuscript.

### Funding

Funding was provided by the Deutsche Forschungsgemeinschaft (DFG grant JA1952/1-1 to V. Jaravine and P. Güntert) and the Lichtenberg program of the Volkswagen Foundation (P. Güntert). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Competing interests

The authors have declared that no competing interests exist.

### Acknowledgments

We thank Dr. John Ferebee and Donata Kirchner for helpful discussions.

### References

Bagaria, A., Jaravine, V., Huang, Y.P.J., Montelione, G.T., Güntert, P., 2012. Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Science* 21, 229–238.

- Berjanskii, M., Zhou, J., Liang, Y., Lin, G., Wishart, D.S., 2012. Resolution-by-proxy: a simple measure for assessing and comparing the overall quality of NMR protein structures. *Journal of Biomolecular NMR* 53, 167–180.
- Bhattacharya, A., Tejero, R., Montelione, G.T., 2007. Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure, Function, and Bioinformatics* 66, 778–795.
- Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., et al., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* 66, 12–21.
- Chen, H.Y., Rogalski, M.M., Anker, J.N., 2012. Advances in functional X-ray imaging techniques and contrast agents. *Physical Chemistry Chemical Physics* 14, 13469–13486.
- Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., Elofsson, A., 2001. Article No.: 5. A study of quality measures for protein threading models. *BMC Bioinformatics* 2.
- Davis, I.W., Murray, L.W., Richardson, J.S., Richardson, D.C., 2004. MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research* 32, W615–W619.
- Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., et al., 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research* 35, W375–W383.
- DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., et al., 2011. Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* 473, 540–U149.
- Dorelijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A.E.M., et al., 2012. CING: an integrated residue-based structure validation program suite. *Journal of Biomolecular NMR* 54, 267–283.
- Eisenberg, D., Lüthy, R., Bowie, J.U., 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods in Enzymology* 277, 396–404.
- Floudas, C.A., Fung, H.K., McAllister, S.R., Mönningmann, M., Rajgaria, R., 2006. Advances in protein structure prediction and de novo protein design: a review. *Chemical Engineering Science* 61, 966–988.
- Güntert, P., 2009. Automated structure determination from NMR spectra. *European Biophysics Journal* 38, 129–143.
- Holton, J.M., Frankel, K.A., 2010. The minimum crystal size needed for a complete diffraction data set. *Acta Crystallographica Section D: Biological Crystallography* 66, 393–408.
- Hoof, R.W.W., Vriend, G., Sander, C., Abola, E.E., 1996. Errors in protein structures. *Nature* 381, 272.
- Huang, Y.J., Powers, R., Montelione, G.T., 2005. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society* 127, 1665–1674.
- Huang, Y.J., Rosato, A., Singh, G., Montelione, G.T., 2012. RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Research* 40, W542–W546.
- Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M.J.E., et al., 2003. CAPRI: a critical assessment of predicted interactions. *Proteins-Structure Function and Bioinformatics* 52, 2–9.
- Joosten, R.P., Womack, T., Vriend, G., Bricogne, G., 2009. Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallographica Section D: Biological Crystallography* 65, 176–185.
- Kelley, L.A., Sternberg, M.J.E., 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* 4, 363–371.
- Kuntz, I.D., Crippen, G.M., Kollman, P.A., Kimelman, D., 1976. Calculation of protein tertiary structure. *Journal of Molecular Biology* 106, 983–994.
- Kwan, A.H., Mobli, M., Gooley, P.R., King, G.F., Mackay, J.P., 2011. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal* 278, 687–703.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26, 283–291.
- Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R., Thornton, J.M., 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR* 8, 477–486.
- Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., et al., 2003. Structure validation by  $\alpha$  geometry:  $\phi$ ,  $\psi$  and  $C\beta$  deviation. *Proteins: Structure, Function, and Genetics* 50, 437–450.
- Minor Jr., D.L., 2007. The neurobiologist's guide to structural biology: a primer on why macromolecular structure matters and how to evaluate structural data. *Neuron* 54, 511–533.
- Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Tramontano, A., 2009. Critical assessment of methods of protein structure prediction-Round VIII. *Proteins: Structure, Function, and Bioinformatics* 77, 1–4.
- Myung, I.J., Pitt, M.A., 1997. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin & Review* 4, 79–95.
- Pantazes, R.J., Grisewood, M.J., Maranas, C.D., 2011. Recent advances in computational protein design. *Current Opinion in Structural Biology* 21, 467–472.
- Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., et al., 2009. CASD-NMR: critical assessment of automated structure determination by NMR. *Nature Methods* 6, 625–626.
- Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., et al., 2012. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20, 227–236.

- Sanchez, R., Sali, A., 1997. Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology* 7, 206–214.
- Siew, N., Elofsson, A., Rychiewski, L., Fischer, D., 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* 16, 776–785.
- Sikic, K., Tomic, S., Carugo, O., 2010. Systematic comparison of crystal and NMR protein structures deposited in the Protein Data Bank. *Open Biochemistry Journal* 4, 83–95.
- Sippl, M.J., 1993. Recognition of errors in 3-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics* 17, 355–362.
- Tomkins, G.M., Yielding, K.L., Talal, N., Curran, J.F., 1963. Protein structure and biological regulation. *Cold Spring Harbor Symposia on Quantitative Biology* 28, 461–471.
- Vriend G, 1990. WHAT IF: a molecular modeling and drug design program. *Journal of Molecular Graphics* 8, 52–56.
- Wallner, B., Elofsson A, 2003. Can correct protein models be identified? *Protein Science* 12, 1073–1086.
- Wlodawer, A., Minor, W., Dauter, Z., Jaskolski M, 2008. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *Febs Journal* 275, 1–21.