STRUCTURAL BIOLOGY

ISSN 2059-7983

Predicting the performance of automated crystallographic model-building pipelines

Emad Alharbi,[a,b]* Paul Bond,[c] Radu Calinescu[a] and Kevin Cowtan[c]

[a]Department of Computer Science, University of York, Heslington, York YO10 5GH, United Kingdom, [b]Department of Information Technology, University of Tabuk, Tabuk, Saudi Arabia, and [c]Department of Chemistry, University of York, Heslington, York YO10 5DD, United Kingdom. *Correspondence e-mail: emad.alharbi@york.ac.uk,emalharbi@ut.edu.sa

Received 23 June 2021
Accepted 10 October 2021

Edited by K. Diederichs, University of Konstanz, Germany

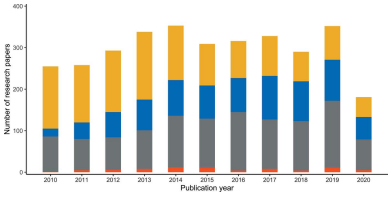
Keywords: structure solution; model building; software; prediction; automated pipelines.

Proteins are macromolecules that perform essential biological functions which depend on their three-dimensional structure. Determining this structure involves complex laboratory and computational work. For the computational work, multiple software pipelines have been developed to build models of the protein structure from crystallographic data. Each of these pipelines performs differently depending on the characteristics of the electron-density map received as input. Identifying the best pipeline to use for a protein structure is difficult, as the pipeline performance differs significantly from one protein structure to another. As such, researchers often select pipelines that do not produce the best possible protein models from the available data. Here, a software tool is introduced which predicts key quality measures of the protein structures that a range of pipelines would generate if supplied with a given crystallographic data set. These measures are crystallographic quality-of-fit indicators based on included and withheld observations, and structure completeness. Extensive experiments carried out using over 2500 data sets show that the tool yields accurate predictions for both experimental phasing data sets (at resolutions between 1.2 and 4.0 Å) and molecular-replacement data sets (at resolutions between 1.0 and 3.5 Å). The tool can therefore provide a recommendation to the user concerning the pipelines that should be run in order to proceed most efficiently to a depositable model.

1. Introduction

The first protein structures were determined in the 1950s using X-ray crystallography (Kendrew *et al.*, 1958). By 2020, the number of solved protein structures deposited in the Protein Data Bank (PDB) exceeded 154 000 (Berman *et al.*, 2000; https://www.rcsb.org/stats/summary). To enable this progress, researchers have automated the computational work of determining the protein structure from X-ray crystallographic data sets. Multiple protein model-building pipelines have been developed within the last three decades: *ARP*/*wARP* (Perrakis *et al.*, 1999; Lamzin & Wilson, 1993; Morris *et al.*, 2003; Langer *et al.*, 2008, 2013), *Buccaneer* (Cowtan, 2006, 2008), *Phenix AutoBuild* (Terwilliger *et al.*, 2008; Liebschner *et al.*, 2019) and *SHELXE* (Sheldrick, 2008, 2010; Thorn & Sheldrick, 2013; Usón & Sheldrick, 2018). In recent studies, we have shown that the performance of these pipelines differs significantly from one protein structure to another (Alharbi *et al.*, 2019), which makes selecting a particular pipeline difficult, and that using a pair of pipelines is sometimes the best option (Alharbi *et al.*, 2020), which greatly increases the number of options that crystallographers can choose from.

An important step in building the protein structure involves solving the phase problem. The phase problem may be solved



using either molecular replacement or experimental phasing methods; see, for example, McCoy & Read (2010) and Evans & McCoy (2008). These methods lead to electron-density maps with rather different properties: in the case of experimental phasing the maps usually contain noise due to ambiguity in the experimental phasing, whereas in the molecular-replacement case errors in the map can arise from possible bias towards the molecular-replacement model. The resolution of the experimental observations, the quality of experimental phasing or the similarity of the molecular-replacement model, and many other features such as ice rings may also affect the quality of the data. Each of these factors impact the performance of different model-building algorithms in different ways (Vollmar *et al.*, 2020; Alharbi *et al.*, 2019; Morris *et al.*, 2004).

The model-building process also contains stochastic elements. The placement of the first atom or residue in a chain will in turn influence the placement of all subsequent elements, and so substantially different model-building results may be obtained from very slight perturbations of the initial conditions. This is addressed in one model-building pipeline by building multiple models at each stage of the process (Terwilliger *et al.*, 2008).

We examined a selection of 3273 research papers cited in the PDB to evaluate how crystallographers currently choose which model-building software pipeline to use, by searching for occurrences of the pipeline names in the text of each paper and excluding papers where the search results were ambiguous or where multiple tools were mentioned. The results are plotted against year, journal and the country of the first author in Fig. 1. The most striking feature of this analysis is the correlation between the first author's country and the country where each pipeline has been developed, with US researchers more likely to use *Phenix Autobuild*, UK researchers more likely to use *Buccaneer* and German researchers more likely to use *ARP/wARP*. While there are practical reasons which might explain this correlation (for example access to developers and workshops), it would be surprising if cognitive biases such as affinity bias (Ashforth & Mael, 1989), to which we are all subject, did not play a role.

To help to eliminate this bias, we have developed a software tool that uses a machine-learning (ML) model to predict the performance of a wide range of model-building pipelines and pipeline combinations for a given crystallographic data set. Our prediction tool serves three purposes.

(i) To provide users with a more efficient route to a higher-quality depositable structure for their specific data set.

(ii) To challenge users to try different pipelines, and multiple combinations of pipelines, on the basis of likely performance rather than on the basis of familiarity or affinity

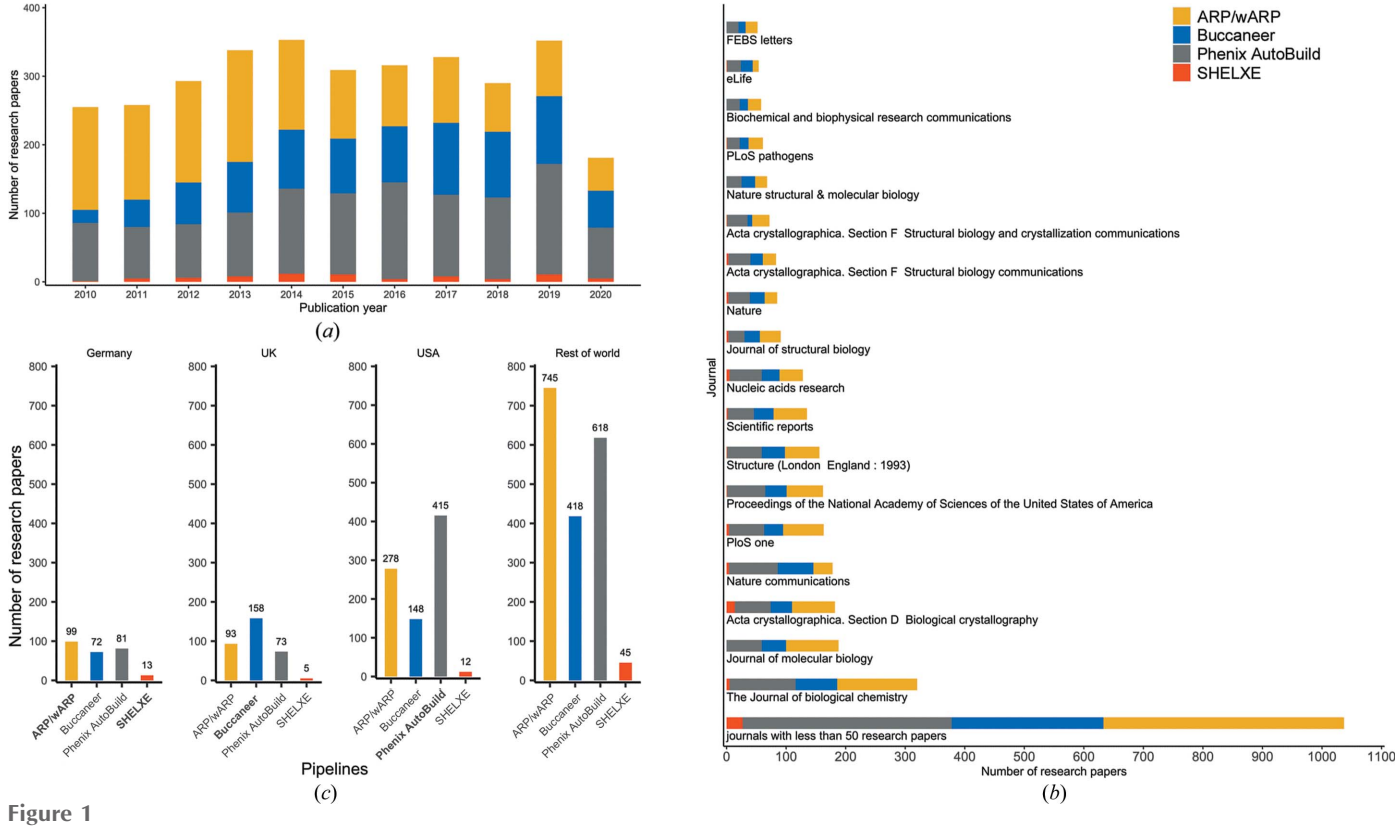


Figure 1
Analysis of the crystallographic model-building pipelines used in 3273 PDB protein-structure research papers published between 2010 and 2020. The papers were identified using either their PubMed identifier or DOI obtained from the PDB. We omitted research papers that used multiple pipelines. We compared the number of uses of each pipeline in its base country, depending on the home country of the first author's organization. (*a*) The number of research papers by publication year for each pipeline. (*b*) The journals in which the research papers were published; journals with fewer than 50 research papers are combined into one group. (*c*) The number of uses of each pipeline in its base country and across the rest of the world; the pipeline names are shown in bold in their base-country plot.

to the pipeline developers. Given that all pipelines provide very convenient user interfaces, the overhead of trying a new pipeline will cost less than the effort of model completion from a suboptimal starting point.

(iii) To assist future developers in the development of meta-tools which make use of multiple pipelines to further automate the process of structure solution and to obtain more complete models.

To the best of our knowledge, this is the first ML solution that guides the user in selection of the model-building pipelines that are most suitable for a given crystallographic data set. While a predictive model that employs similar ML techniques was recently proposed by Vollmar *et al.* (2020), that model addresses the complementary problem of predicting the usefulness of collected crystallographic data sets.

2. Predictive model

2.1. Data sets

We used data sets from three sources to train and evaluate our ML predictive model: 1203 experimental phasing data sets from the Joint Center for Structural Genomics (JCSG; van den Bedem *et al.*, 2011; Alharbi *et al.*, 2019), 32 newer experimental phasing data sets deposited between 2015 and 2021 and taken from the PDB, and 1332 molecular-replacement (MR) data sets from Bond *et al.* (2020). These data sets correspond to two techniques that can be used to build a protein structure. Experimental phasing is when the phases are determined from the observed data using the features of special atoms, such as a large number of electrons; see, for example, Dauter & Dauter (2017). In contrast, MR obtains initial phases from a known protein structure that is similar to the protein structure that we want to build; see, for example, Evans & McCoy (2008).

The resolution of the JCSG experimental phasing data sets ranges from 1.2 to 4.0 Å, with the low-resolution data sets augmented by simulation as in Alharbi *et al.* (2019), the resolution of the PDB experimental phasing data sets ranges from 1.1 to 5.8 Å, and the resolution of the MR data sets ranges from 1.0 to 3.5 Å. Lower resolution data sets have fewer experimental observations, which decreases the performance of the protein-building pipelines.

The way in which we partitioned these data sets into data for training and data for evaluation of our ML model is described in Section 2.5.

2.2. Crystallographic model-building pipelines

The four pipeline versions used in our work are *Phenix AutoBuild* version 1.14, *Buccaneer* in *CCP*4*i* version 7.0.066, *ARP*/*wARP* version 8 and *SHELXE* version 2019/1. These pipelines were run using the default parameters, both individually and in pairwise combinations where the protein model produced by a first pipeline *x* was supplied as input to a second pipeline *y*.

2.3. Protein structure evaluation

We focused on predicting three protein structure evaluation measures, namely $R_{\rm free}$, $R_{\rm work}$ and structure completeness. $R_{\rm free}$ and $R_{\rm work}$ measure the fit of the protein structure against the observed data, with $R_{\rm free}$ only using observations which are not used in the refinement calculation: typically 5% of the data (Brünger, 1992). Structure completeness is the percentage of residues in the deposited protein model with a matching residue in the built model. Residues are considered to match if they have the same type and the distance between their C^{α} atoms is less than 1 Å.

2.4. Electron-density map features

We trained our ML prediction model using the resolution of the crystallographic data set and the following measures of the quality of the electron-density map as input features.

(i) R.m.s.d.: the root-mean-square deviation of the electron density from the mean of the map.

(ii) Skew: the third moment of the electron density about the mean, which measures the asymmetry of the electron-density histogram (Terwilliger *et al.*, 2009).

(iii) Maximum density: the highest density of the electron-density map.

(iv) Minimum density: the lowest density of the electron-density map.

(v) Sequence identity: the sequence identity calculated by superposition of the homologue chain onto the target chain using *GESAMT* (Krissinel, 2012; Bond *et al.*, 2020).

2.5. Predictive model training

The individual pipelines were run on all data sets listed in Section 2.1. The pipeline combinations were only run on the experimental phasing data sets, as building protein models from such 'raw data' can often be improved by using combinations of pipelines (Alharbi *et al.*, 2020). The results of these runs are described in detail in our recent work (Alharbi *et al.*, 2019, 2020). The data sets and the protein structures obtained from these runs were used to train and evaluate the predictive ML model as follows.

(i) 80% of the JCSG experimental phasing data sets and 80% of the MR data sets were used to train the predictive model.

(ii) The remaining 20% of the JCSG experimental phasing and MR data sets, and all 32 PDB experimental phasing data sets, were used to evaluate the trained model.

We used random forests (Breiman, 2001) as implemented in the Weka framework (Hall *et al.*, 2009; Frank *et al.*, 2016) for the predictive model, as this approach showed the lowest error rate across the ML algorithms that we tested, which included a support vector machine (Cortes & Vapnik, 1995) and the *RepTree* decision-tree algorithm. We varied the number of trees in the random forest from 1 to 5000 in geometric sequence, and 1024 was chosen for the final training as this showed the lowest error rate. The depth of the trees was set to unlimited, and bagging (Breiman, 1996) was used to reduce the variance. We trained the predictive model using a 173-node

high-performance cluster with 7024 Intel Xeon Gold/Platinum cores and a total memory of 42 TB.

A separate regression ML model (random forest model) was trained for each of the 24 pipeline variants (*i.e.* individual pipelines or pipeline combinations) in Fig. 2 and for each of the three structure evaluation measures in Section 2.3 relevant to the considered pipeline variant. For instance, $R_{\rm free}$ is not relevant for *ARP/wARP* and *SHELXE* with and without *Parrot* used on their own, so no ML model was built for these individual pipelines and $R_{\rm free}$. We obtained a total of 69 and ten regression ML models for experimental phasing and for MR, respectively. Our predictive model consists of these regression ML models taken together.

We used the root-mean-square error (RMSE) and mean absolute error (MAE) measures to compare the accuracy of our predictive model with that of a 'baseline' predictive model. In line with the standard practice for the evaluation of regression models, we used the *Zero-R* algorithm as a baseline predictive model (Choudhary & Gianey, 2017). Given a pipeline variant and any evaluation data set, the *Zero-R* algorithm predicts that the $R_{\rm free}/R_{\rm work}$ and structure completeness for the structure built by the pipeline would be the same as the median $R_{\rm free}/R_{\rm work}$ and structure completeness for the training data sets, respectively.

To evaluate the accuracy of the predictive model for data sets of different resolutions, we partitioned the evaluation data sets into classes based on their resolutions, and we examined the prediction errors for each such class. Finally, to show the time saved by running only the pipeline variant predicted to build the best protein structure for a data set, we compared the execution time of this pipeline with the time required to run all of the pipeline variants for that data set.

To quantify the uncertainty of the ML prediction, we calculated prediction intervals using the kernel estimator method from Frank & Bouckaert (2009). The width of these intervals reflects the prediction uncertainty. As such, we sort and report the pipelines in increasing prediction interval width order, with pipelines of similar prediction uncertainty (*i.e.* with no more than 5% difference in prediction interval width) grouped together.

Finally, we generate a script for each pipeline and pipeline combination, ensuring that the users of our tool can run the individual pipelines and pipeline combinations in the manner used to obtain the training data sets for our ML prediction model. Furthermore, these ready-to-run scripts are customized based on data provided by the tool users.

3. Predictive model evaluation

3.1. Evaluation of the crystallographic data-set features used for model training

We evaluated the importance of the features used to train our predictive model by removing one feature at a time and comparing the accuracy of the model trained without that feature with the accuracy of the predictive model when trained on all of the features. Fig. 3 shows the difference in MAE and RMSE when one feature is removed compared with when all of the features are used in training for each of the four individual pipelines, with separate MAE and RMSE presented for the JCSG experimental phasing and MR data sets.

This analysis indicates that *Phenix AutoBuild* and *ARP/wARP* are more dependent on the data-set resolution than *Buccaneer*, which is in line with previous results (Alharbi *et al.*, 2019). However, *Phenix AutoBuild* and *ARP/wARP* are less sensitive to the resolution for MR data sets compared with experimental phasing data sets. R.m.s.d. and skew have different effects on the performance of the pipelines. For example, *Buccaneer* is affected by these two features more than *Phenix AutoBuild* for the experimental phasing data set, indicating a greater dependence on the noise level in the starting map. For MR data sets, the sequence identity affected the performance of all pipelines, with the highest effect for *Buccaneer*.

3.2. Evaluation of predictive model performance

Fig. 2 shows the MAE and RMSE for both types of data sets (experimental phasing and MR) for each of the three protein structure evaluation measures. For the JCSG experimental phasing data sets, both the MAE (0.04–0.19) and RMSE (0.08–

	Experimental phasing (JCSG)												Experimental phasing (recently deposited data sets)												MR											
	MAE						RMSE						MAE						RMSE						MAE						RMSE					
Pipeline variant	Completeness		$R_{\rm free}$		$R_{\rm work}$		Completeness		$R_{\rm free}$		$R_{\rm work}$		Completeness		$R_{\rm free}$		$R_{\rm work}$		Completeness		$R_{\rm free}$		$R_{\rm work}$		Completeness		$R_{\rm free}$		$R_{\rm work}$		Completeness		$R_{\rm free}$		$R_{\rm work}$	
	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M
ARP/wARP	0.06	0.15	-	-	0.03	0.03	0.14	0.34	-	-	0.05	0.05	0.27	0.57	-	-	0.04	0.05	0.38	0.72	-	-	0.06	0.07	0.2	0.39	-	-	0.04	0.06	0.29	0.58	-	-	0.05	0.07
ARP/wARP → Buccaneer	0.19	0.3	0.05	0.08	0.05	0.07	0.25	0.34	0.07	0.1	0.06	0.08	0.26	0.42	0.08	0.12	0.07	0.1	0.32	0.44	0.09	0.13	0.09	0.11	-	-	-	-	-	-	-	-	-	-	-	-
ARP/wARP → Phenix AutoBuild(Parrot)	0.11	0.24	0.06	0.07	0.02	0.03	0.15	0.34	0.08	0.09	0.03	0.03	0.16	0.61	0.05	0.1	0.03	0.05	0.21	0.65	0.07	0.12	0.04	0.06	-	-	-	-	-	-	-	-	-	-	-	-
ARP/wARP → Phenix AutoBuild	0.1	0.23	0.06	0.07	0.02	0.03	0.15	0.33	0.07	0.09	0.03	0.03	0.23	0.6	0.07	0.1	0.04	0.05	0.32	0.64	0.09	0.12	0.05	0.05	-	-	-	-	-	-	-	-	-	-	-	-
Buccaneer	0.18	0.3	0.05	0.08	0.05	0.07	0.23	0.33	0.07	0.09	0.06	0.08	0.24	0.4	0.07	0.12	0.07	0.1	0.31	0.42	0.09	0.13	0.09	0.1	0.15	0.29	0.04	0.07	0.04	0.07	0.2	0.37	0.06	0.11	0.05	0.09
Buccaneer → ARP/wARP	0.06	0.17	0.05	0.08	0.02	0.03	0.15	0.37	0.07	0.11	0.03	0.03	0.27	0.62	0.09	0.19	0.04	0.05	0.38	0.75	0.12	0.21	0.05	0.06	-	-	-	-	-	-	-	-	-	-	-	-
Buccaneer → Phenix AutoBuild(Parrot)	0.16	0.28	0.05	0.07	0.03	0.05	0.21	0.32	0.06	0.08	0.04	0.06	0.1	0.33	0.04	0.1	0.04	0.07	0.14	0.35	0.06	0.11	0.05	0.08	-	-	-	-	-	-	-	-	-	-	-	-
Buccaneer → Phenix AutoBuild	0.15	0.28	0.05	0.07	0.04	0.05	0.2	0.31	0.06	0.08	0.05	0.06	0.2	0.33	0.06	0.1	0.05	0.07	0.28	0.35	0.07	0.11	0.07	0.08	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild(Parrot)	0.09	0.21	0.03	0.06	0.02	0.04	0.12	0.3	0.05	0.08	0.03	0.06	0.11	0.56	0.04	0.12	0.03	0.09	0.13	0.59	0.05	0.13	0.04	0.1	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild(Parrot) → ARP/wARP	0.04	0.16	0.04	0.08	0.02	0.02	0.09	0.37	0.06	0.11	0.03	0.03	0.18	0.71	0.07	0.2	0.03	0.04	0.29	0.79	0.1	0.22	0.05	0.05	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild(Parrot) → Buccaneer	0.17	0.27	0.05	0.07	0.05	0.06	0.22	0.32	0.07	0.09	0.06	0.07	0.13	0.31	0.05	0.1	0.04	0.07	0.17	0.33	0.07	0.11	0.06	0.08	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild → ARP/wARP	0.04	0.16	0.04	0.07	0.02	0.02	0.08	0.37	0.06	0.11	0.03	0.03	0.21	0.68	0.08	0.19	0.04	0.05	0.31	0.78	0.1	0.21	0.07	0.08	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild → Buccaneer	0.18	0.26	0.05	0.07	0.05	0.06	0.23	0.31	0.07	0.09	0.06	0.07	0.15	0.29	0.05	0.1	0.04	0.07	0.19	0.32	0.07	0.11	0.06	0.08	-	-	-	-	-	-	-	-	-	-	-	-
Phenix AutoBuild	0.09	0.21	0.03	0.05	0.03	0.04	0.12	0.3	0.04	0.08	0.03	0.05	0.16	0.55	0.05	0.12	0.05	0.09	0.26	0.58	0.08	0.14	0.09	0.1	0.21	0.28	0.07	0.09	0.06	0.08	0.27	0.35	0.09	0.11	0.08	0.1
SHELXE	0.14	0.18	-	-	0.02	0.03	0.2	0.26	-	-	0.03	0.03	0.18	0.28	-	-	0.03	0.04	0.23	0.4	-	-	0.03	0.06	0.17	0.36	-	-	0.02	0.04	0.23	0.39	-	-	0.04	0.05
SHELXE → ARP/wARP	0.17	0.23	0.06	0.08	0.03	0.03	0.26	0.41	0.08	0.11	0.04	0.04	0.22	0.37	0.08	0.12	0.05	0.06	0.32	0.55	0.11	0.17	0.06	0.07	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE → Buccaneer	0.12	0.1	0.04	0.05	0.04	0.04	0.19	0.2	0.06	0.06	0.05	0.06	0.27	0.26	0.07	0.09	0.07	0.08	0.35	0.43	0.1	0.13	0.09	0.11	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE → Phenix AutoBuild(Parrot)	0.06	0.06	0.03	0.03	0.02	0.03	0.08	0.09	0.03	0.04	0.03	0.03	0.13	0.15	0.05	0.06	0.04	0.05	0.19	0.28	0.06	0.08	0.06	0.07	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE → Phenix AutoBuild	0.07	0.07	0.03	0.04	0.02	0.03	0.11	0.12	0.03	0.04	0.03	0.04	0.22	0.18	0.06	0.07	0.06	0.06	0.32	0.34	0.09	0.11	0.09	0.09	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE(Parrot) → ARP/wARP	0.17	0.2	0.06	0.07	0.03	0.03	0.26	0.38	0.07	0.1	0.03	0.03	0.23	0.33	0.08	0.12	0.04	0.05	0.32	0.51	0.11	0.17	0.06	0.07	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE(Parrot) → Buccaneer	0.11	0.11	0.04	0.05	0.03	0.04	0.17	0.21	0.05	0.07	0.05	0.06	0.21	0.26	0.06	0.09	0.05	0.08	0.3	0.42	0.08	0.13	0.07	0.11	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE(Parrot) → Phenix AutoBuild(Parrot)	0.06	0.06	0.03	0.03	0.02	0.03	0.09	0.1	0.03	0.04	0.03	0.03	0.11	0.14	0.05	0.06	0.04	0.04	0.17	0.27	0.07	0.09	0.06	0.07	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE(Parrot) → Phenix AutoBuild	0.06	0.07	0.02	0.03	0.02	0.03	0.11	0.12	0.03	0.04	0.03	0.04	0.21	0.13	0.06	0.05	0.06	0.04	0.29	0.26	0.08	0.08	0.08	0.07	-	-	-	-	-	-	-	-	-	-	-	-
SHELXE(Parrot)	0.11	0.14	-	-	0.02	0.02	0.16	0.21	-	-	0.02	0.03	0.17	0.25	-	-	0.03	0.04	0.22	0.37	-	-	0.03	0.05	-	-	-	-	-	-	-	-	-	-	-	-

Figure 2
Mean absolute error (MAE) and root-mean-squared error (RMSE) of structure completeness and $R_{\rm free}/R_{\rm work}$ for two types of experimental phasing data sets and for molecular-replacement (MR) data sets. *ARP/wARP* and *SHELXE* are not used for $R_{\rm free}$. For the MR data sets, only individual pipelines were run. MAE and RMSE were calculated for the ML predictive model (P) and median predictor (M) used as a baseline (*Zero-R*) model.

0.26) for predicting the protein structure completeness are higher than the MAE and RMSE for the other measures. The values of MAE (0.02–0.06) and RMSE (0.02–0.08) decreased when predicting R_{free}/R_{work}. For MR data sets, the MAE of structure completeness increased to 0.15–0.21 and the RMSE to 0.20–0.29. The MAE of R_{free}/R_{work} was between 0.02 and 0.07, compared with the RMSE, which is between 0.04 and 0.09.

Different levels of predictability were achieved for different pipeline variants. For the experimental phasing data sets and *ARP/wARP* after *Phenix AutoBuild*, the predictive model achieved the lowest MAE for structure completeness (0.04), with a similar RMSE, which indicates a small number of large error predictions. On the other hand, for MR data sets, the MAE for structure completeness for *ARP/wARP* and *Phenix AutoBuild* run individually increased to 0.20 and 0.21, respectively. *Buccaneer* run individually and after *ARP/wARP* or *Phenix AutoBuild* showed the lowest predictability, with MAE and RMSE values above 0.17.

R_{free}/R_{work} are more predictable across all pipeline variants and for both types of data sets, with lower MAE and RMSE values than those achieved for structure completeness. For the JCSG experimental phasing data sets, the predictive model achieved a low MAE for R_{work} (0.02–0.03) and only a slightly larger MAE for R_{free} (0.03–0.05) for all of the individual pipelines. The MAE obtained for pipeline combinations and R_{work} ranged between 0.02 and 0.05, and that for R_{free} varied between 0.04 and 0.06. RMSE is slightly higher than MAE for both the individual and the combined pipelines. For the MR data sets, the MAE of R_{work} is between 0.02 and 0.06, with the lowest value being obtained for *SHELXE*, and the MAE for R_{free} is between 0.04 and 0.07. Finally, the RMSEs of R_{free} and R_{work} are between 0.06 and 0.09 and between 0.04 and 0.08, respectively.

Compared with the baseline *Zero-R* predictive model (see Section 2.5), our predictive model achieved lower or much lower MAE and RMSE prediction errors for almost all of the pipeline variants, types of data sets and protein structure evaluation measures, *i.e.* for 288 of the 296 entries in Fig. 2. Notably, the predictions for recently PDB-deposited experimental phasing data sets (which we did not use in the training of the predictive model) also have a much lower error for our predictive model than for the *Zero-R* predictive model (Fig. 4), with the exception of the predictions for *SHELXE* before *Buccaneer* and *Phenix AutoBuild*, for which the *Zero-R* baseline model predictions achieve similar or marginally lower errors.

To evaluate the fitting of our predictive model, Fig. 5 shows the difference in MAE and RMSE between training and testing for the JCSG experimental phasing and the MR data sets. The difference in MAE and RMSE between training and testing data sets for structure completeness is higher than that in R_{work}/R_{free} for the JCSG experimental phasing and the MR data sets. When comparing the pipelines by structure completeness, *Phenix AutoBuild* and *Buccaneer* have the lowest error difference for the JCSG experimental phasing and the MR data sets, respectively. For R_{work}/R_{free}, the pipelines have a smaller difference in MAE and RMSE between the training and testing data sets compared with the structure completeness.

To further evaluate the accuracy of our predictive model, we analysed the mean and standard deviation (SD) of the predicted and actual protein structure evaluation measures for the crystallographic data sets grouped based on their

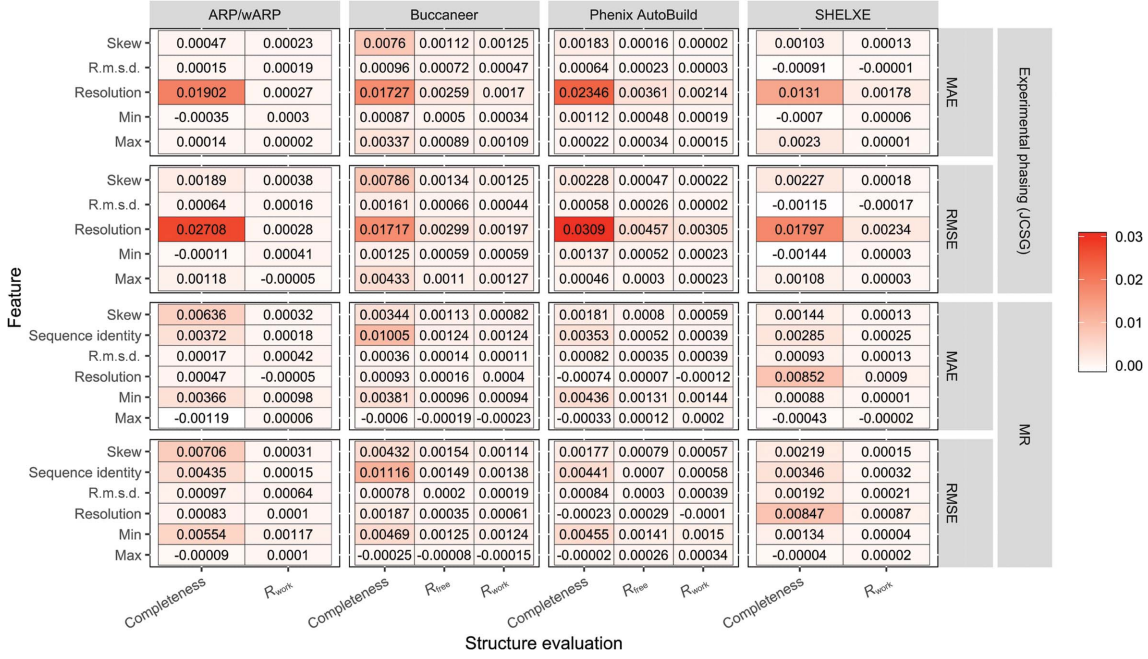


Figure 3
Ablation studies showing the difference in MAE and RMSE when the ML model was trained on all features and when one feature is removed at a time. Higher values indicate more important features.

resolutions. Figs. 6 and 7 show the results of this analysis for JCSG experimental phasing data sets for the pipeline variants without *SHELXE* and with *SHELXE*, respectively. For resolutions between 1.2 and 3.1 Å, the predicted and actual mean and SD values are very close for most pipeline variants. The spread of the predicted structure completeness for *ARP/wARP* run alone and run after *SHELXE* has a higher SD compared with the completeness achieved when the pipelines were run in reality. At worse than 3.2 Å, the predicted R_{free}/R_{work} have mean and SD values close to the real results, while the predicted structure completeness has a larger difference in the SD and a smaller difference in the mean than the actual results.

Fig. 8 shows the results of the same analysis as above for the MR data sets. The mean of all the predicted structure evaluation measures as well as the SD values for the predicted R_{free}/R_{work} are close to the actual results. However, at resolutions better than 3.0 Å the difference between the SD for

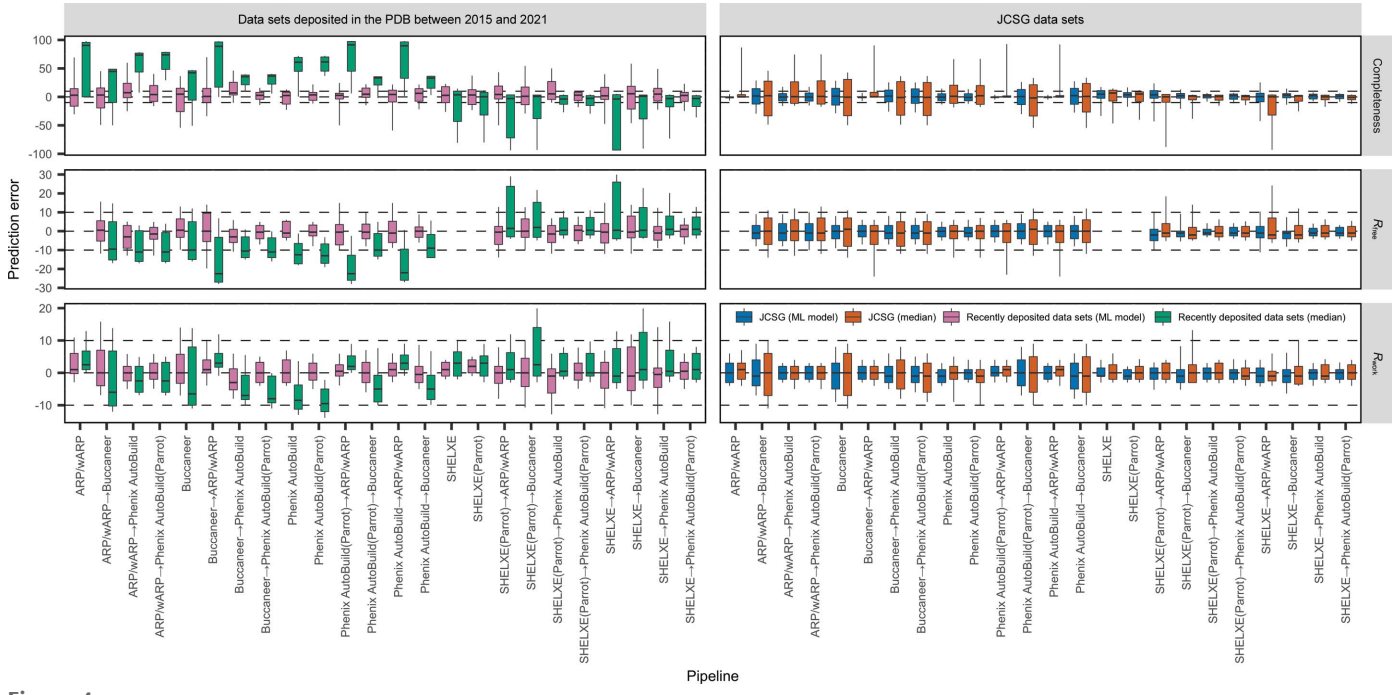


Figure 4
Prediction error for the ML predictive model and the median predictor for recently deposited and JCSG experimental phasing data sets.

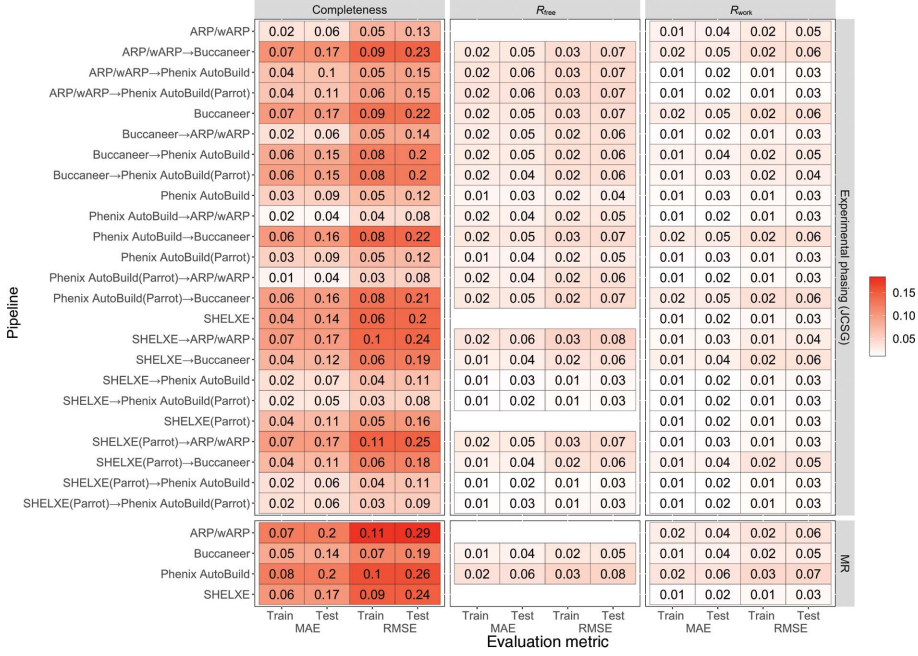


Figure 5
MAE and RMSE of structure completeness and R_{free}/R_{work} for training and testing for the JCSG experimental phasing data sets and the MR data sets. The entries are shaded based on the magnitude of the difference in MAE and RMSE between the training and testing data sets.

the predicted and actual structure completeness is larger than that for R_{free}/R_{work}. At resolutions of 3.1 Å or worse, this difference decreases significantly.

To evaluate the predictive model uncertainty, we grouped the pipelines using the method described in Section 2.5. We evaluated this by checking whether the pipeline with the lowest prediction error was classified in the first group for each protein structure in our testing data set. For the JCSG experimental phasing data set, 85%, 94% and 91% of the pipelines with the lowest prediction error were classified in the

Pipeline variant	Structure evaluation	Resolution																					
		1.2 - 3.1(39)				3.2(45)				3.4(41)				3.6(31)				3.8(43)				4.0+(42)			
		mean		SD		mean		SD		mean		SD		mean		SD		mean		SD		mean		SD	
		P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
ARP/wARP →Buccaneer	Completeness	0.87	0.89	0.15	0.19	0.64	0.6	0.17	0.34	0.56	0.54	0.17	0.29	0.46	0.49	0.14	0.29	0.28	0.34	0.12	0.24	0.15	0.16	0.08	0.18
	R_{free}	0.32	0.31	0.04	0.07	0.40	0.41	0.05	0.09	0.42	0.42	0.04	0.07	0.44	0.43	0.04	0.08	0.48	0.47	0.03	0.07	0.50	0.50	0.03	0.06
	R_{work}	0.29	0.28	0.04	0.06	0.33	0.34	0.05	0.09	0.35	0.35	0.04	0.07	0.37	0.35	0.04	0.07	0.40	0.39	0.03	0.07	0.42	0.42	0.02	0.05
ARP/wARP → PHENIX AutoBuild(Parrot)	Completeness	0.9	0.91	0.11	0.11	0.35	0.36	0.08	0.22	0.28	0.29	0.06	0.18	0.19	0.17	0.06	0.16	0.11	0.12	0.04	0.1	0.07	0.07	0.03	0.06
	R_{free}	0.27	0.26	0.04	0.05	0.39	0.39	0.02	0.07	0.39	0.38	0.03	0.06	0.42	0.42	0.02	0.08	0.43	0.43	0.03	0.09	0.44	0.43	0.03	0.08
	R_{work}	0.23	0.22	0.02	0.03	0.25	0.25	0.01	0.02	0.25	0.24	0.01	0.02	0.27	0.28	0.01	0.02	0.27	0.27	0.01	0.03	0.27	0.28	0.01	0.03
ARP/wARP → PHENIX AutoBuild	Completeness	0.91	0.91	0.09	0.12	0.38	0.35	0.06	0.22	0.28	0.29	0.08	0.16	0.19	0.2	0.06	0.15	0.14	0.14	0.05	0.11	0.09	0.1	0.03	0.07
	R_{free}	0.26	0.26	0.04	0.06	0.38	0.39	0.02	0.07	0.39	0.38	0.03	0.06	0.42	0.42	0.03	0.07	0.43	0.44	0.02	0.10	0.43	0.43	0.03	0.07
	R_{work}	0.22	0.22	0.02	0.03	0.25	0.25	0.01	0.02	0.25	0.24	0.01	0.02	0.27	0.28	0.01	0.03	0.27	0.26	0.01	0.03	0.26	0.27	0.01	0.03
ARP/wARP	Completeness	0.83	0.77	0.21	0.32	0.08	0.09	0.07	0.16	0.03	0.03	0.03	0.05	0.01	0.01	0.01	0.03	0.01	0	0.01	0	0	0	0	0
	R_{work}	0.23	0.23	0.02	0.03	0.22	0.23	0.01	0.05	0.22	0.22	0.02	0.04	0.22	0.21	0.02	0.04	0.21	0.21	0.02	0.05	0.21	0.22	0.02	0.05
Buccaneer → ARP/wARP	Completeness	0.86	0.83	0.2	0.26	0.12	0.14	0.08	0.21	0.04	0.04	0.05	0.08	0.01	0.01	0.01	0.04	0	0	0	0	0	0	0	0
	R_{free}	0.29	0.29	0.06	0.08	0.49	0.49	0.03	0.08	0.51	0.50	0.03	0.06	0.52	0.51	0.02	0.06	0.54	0.54	0.02	0.07	0.55	0.55	0.02	0.04
	R_{work}	0.23	0.23	0.01	0.03	0.20	0.20	0.01	0.03	0.20	0.20	0.01	0.02	0.19	0.19	0.02	0.03	0.19	0.19	0.02	0.04	0.19	0.20	0.02	0.04
Buccaneer → PHENIX AutoBuild(Parrot)	Completeness	0.91	0.93	0.07	0.09	0.69	0.67	0.12	0.28	0.61	0.62	0.12	0.24	0.52	0.56	0.11	0.27	0.4	0.41	0.13	0.25	0.23	0.2	0.08	0.16
	R_{free}	0.26	0.25	0.03	0.04	0.34	0.35	0.03	0.08	0.36	0.35	0.03	0.07	0.38	0.36	0.03	0.06	0.40	0.40	0.03	0.06	0.43	0.43	0.02	0.05
	R_{work}	0.23	0.23	0.02	0.03	0.27	0.27	0.02	0.06	0.28	0.27	0.02	0.04	0.31	0.29	0.02	0.05	0.32	0.32	0.02	0.05	0.34	0.34	0.02	0.03
Buccaneer → PHENIX AutoBuild	Completeness	0.91	0.93	0.07	0.08	0.72	0.67	0.11	0.27	0.6	0.62	0.14	0.23	0.54	0.56	0.14	0.28	0.41	0.43	0.14	0.25	0.23	0.22	0.08	0.16
	R_{free}	0.26	0.25	0.03	0.04	0.34	0.34	0.02	0.08	0.37	0.35	0.04	0.06	0.38	0.36	0.03	0.07	0.40	0.40	0.03	0.06	0.43	0.44	0.02	0.04
	R_{work}	0.23	0.23	0.02	0.03	0.27	0.27	0.02	0.06	0.28	0.28	0.03	0.05	0.31	0.29	0.02	0.05	0.32	0.32	0.02	0.05	0.34	0.35	0.01	0.03
Buccaneer	Completeness	0.83	0.86	0.15	0.19	0.63	0.61	0.16	0.32	0.56	0.53	0.15	0.28	0.47	0.48	0.11	0.28	0.33	0.37	0.12	0.26	0.19	0.17	0.08	0.17
	R_{free}	0.33	0.32	0.05	0.07	0.40	0.41	0.04	0.09	0.42	0.43	0.04	0.08	0.45	0.43	0.02	0.07	0.47	0.46	0.03	0.08	0.50	0.50	0.03	0.05
	R_{work}	0.30	0.29	0.04	0.06	0.34	0.34	0.04	0.09	0.35	0.36	0.04	0.07	0.37	0.36	0.03	0.07	0.39	0.38	0.03	0.07	0.41	0.42	0.02	0.05
PHENIX AutoBuild(Parrot) → ARP/wARP	Completeness	0.9	0.89	0.15	0.15	0.05	0.08	0.05	0.16	0.03	0.02	0.03	0.04	0.01	0.01	0.01	0.03	0	0	0	0	0	0	0	0
	R_{free}	0.28	0.28	0.04	0.05	0.50	0.49	0.03	0.07	0.50	0.50	0.03	0.05	0.52	0.51	0.02	0.07	0.53	0.53	0.02	0.07	0.53	0.53	0.03	0.06
	R_{work}	0.22	0.22	0.01	0.03	0.21	0.21	0.01	0.02	0.21	0.20	0.01	0.02	0.19	0.20	0.02	0.02	0.19	0.19	0.01	0.04	0.19	0.21	0.02	0.03
PHENIX AutoBuild(Parrot) →Buccaneer	Completeness	0.92	0.94	0.09	0.09	0.7	0.66	0.13	0.28	0.63	0.61	0.13	0.23	0.54	0.53	0.12	0.28	0.42	0.44	0.13	0.28	0.28	0.29	0.1	0.22
	R_{free}	0.31	0.30	0.03	0.04	0.38	0.40	0.04	0.08	0.40	0.41	0.03	0.07	0.42	0.42	0.03	0.09	0.45	0.46	0.03	0.08	0.48	0.48	0.03	0.05
	R_{work}	0.28	0.27	0.03	0.04	0.32	0.33	0.04	0.08	0.33	0.34	0.03	0.06	0.35	0.35	0.03	0.08	0.37	0.36	0.03	0.08	0.39	0.39	0.03	0.05
PHENIX AutoBuild(Parrot)	Completeness	0.91	0.92	0.08	0.07	0.37	0.38	0.06	0.15	0.32	0.32	0.06	0.13	0.26	0.26	0.06	0.13	0.18	0.18	0.04	0.12	0.12	0.12	0.03	0.07
	R_{free}	0.27	0.26	0.04	0.04	0.41	0.41	0.01	0.04	0.41	0.41	0.01	0.04	0.42	0.42	0.01	0.05	0.43	0.44	0.01	0.06	0.44	0.44	0.02	0.04
	R_{work}	0.23	0.23	0.02	0.03	0.33	0.33	0.01	0.03	0.33	0.33	0.01	0.03	0.34	0.34	0.01	0.03	0.35	0.35	0.01	0.04	0.36	0.35	0.01	0.03
PHENIX AutoBuild → ARP/wARP	Completeness	0.87	0.89	0.14	0.14	0.07	0.07	0.05	0.15	0.02	0.02	0.02	0.04	0.01	0.01	0.01	0.02	0	0	0	0.01	0	0	0	0
	R_{free}	0.29	0.28	0.04	0.06	0.50	0.50	0.03	0.07	0.51	0.51	0.03	0.04	0.52	0.50	0.03	0.06	0.53	0.53	0.02	0.06	0.52	0.53	0.03	0.06
	R_{work}	0.23	0.22	0.02	0.03	0.21	0.21	0.01	0.03	0.21	0.20	0.01	0.02	0.19	0.19	0.02	0.03	0.19	0.19	0.01	0.04	0.19	0.20	0.02	0.03
PHENIX AutoBuild →Buccaneer	Completeness	0.9	0.93	0.09	0.09	0.7	0.69	0.13	0.26	0.59	0.61	0.13	0.27	0.54	0.56	0.13	0.25	0.42	0.45	0.14	0.27	0.29	0.29	0.11	0.24
	R_{free}	0.31	0.30	0.03	0.05	0.39	0.39	0.04	0.08	0.41	0.40	0.03	0.07	0.43	0.42	0.03	0.07	0.45	0.46	0.03	0.08	0.48	0.48	0.03	0.06
	R_{work}	0.29	0.27	0.03	0.04	0.32	0.32	0.03	0.07	0.34	0.34	0.03	0.07	0.35	0.34	0.03	0.07	0.37	0.36	0.03	0.07	0.39	0.39	0.03	0.05
PHENIX AutoBuild	Completeness	0.9	0.91	0.07	0.09	0.38	0.39	0.05	0.15	0.32	0.32	0.07	0.12	0.26	0.25	0.06	0.12	0.19	0.19	0.05	0.12	0.12	0.12	0.04	0.07
	R_{free}	0.27	0.26	0.03	0.04	0.41	0.41	0.01	0.05	0.41	0.41	0.02	0.03	0.42	0.41	0.02	0.05	0.43	0.43	0.01	0.05	0.44	0.44	0.02	0.03
	R_{work}	0.24	0.23	0.02	0.03	0.33	0.33	0.01	0.03	0.33	0.33	0.01	0.02	0.35	0.34	0.01	0.04	0.35	0.35	0.01	0.03	0.36	0.36	0.01	0.02

0.0 0.2 0.0 0.06

Figure 6
Mean and standard deviation (SD) of the real and predicted structure evaluation measures for the JCSG experimental phasing data sets grouped based on resolution, with the number of data sets in each group shown in parentheses. The entries are shaded based on the magnitude of the difference between the real (R) and predicted (P) results.

Pipeline variant	Completeness				R_{free}				R_{work}			
	mean		SD		mean		SD		mean		SD	
	P	R	P	R	P	R	P	R	P	R	P	R
SHELXE→ ARP/wARP	0.77	0.73	0.25	0.36	0.32	0.32	0.09	0.11	0.24	0.24	0.03	0.04
SHELXE→Buccaneer	0.88	0.88	0.16	0.19	0.32	0.31	0.05	0.06	0.29	0.29	0.04	0.06
SHELXE→PHENIX AutoBuild(Parrot)	0.92	0.91	0.05	0.09	0.26	0.26	0.03	0.04	0.23	0.23	0.02	0.03
SHELXE→PHENIX AutoBuild	0.91	0.9	0.07	0.12	0.26	0.26	0.03	0.05	0.23	0.23	0.02	0.04
SHELXE	0.78	0.78	0.17	0.26	-	-	-	-	0.46	0.46	0.02	0.03
SHELXE(Parrot)→ARP/wARP	0.77	0.78	0.25	0.35	0.32	0.31	0.08	0.10	0.24	0.23	0.03	0.03
SHELXE(Parrot)→Buccaneer	0.88	0.88	0.16	0.2	0.32	0.31	0.05	0.07	0.29	0.28	0.04	0.06
SHELXE(Parrot)→PHENIX AutoBuild(Parrot)	0.92	0.91	0.04	0.1	0.26	0.26	0.03	0.04	0.23	0.23	0.02	0.03
SHELXE(Parrot)→PHENIX AutoBuild	0.92	0.91	0.05	0.12	0.26	0.26	0.03	0.04	0.23	0.23	0.02	0.04
SHELXE(Parrot)	0.79	0.81	0.17	0.21	-	-	-	-	0.46	0.45	0.02	0.03

0.0 0.2 0.0 0.06

Figure 7
Mean and SD of the real and predicted structure evaluation measures for the JCSG experimental phasing data sets for *SHELXE* and its combinations. The resolutions of the data sets are between 1.2 and 3.1 Å. The results are shaded based on the difference between the real (R) and predicted (P) results.

Pipeline variant	Structure evaluation	Resolution																					
		1.0 – 1.5 (65)				1.6 – 2.0 (65)				2.1 – 2.5 (50)				2.6 – 3.0 (55)				3.1+ (31)					
		mean		SD		mean		SD		mean		SD		mean		SD		mean		SD			
		P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R		
ARP/wARP	Completeness	0.65	0.63	0.31	0.46	0.47	0.51	0.31	0.46	0.32	0.37	0.28	0.43	0.16	0.18	0.18	0.34	0.04	0.02	0.03	0.04		
	R_{work}	0.25	0.25	0.03	0.05	0.27	0.27	0.03	0.06	0.30	0.29	0.05	0.08	0.29	0.29	0.04	0.07	0.31	0.31	0.05	0.08		
Buccaneer	Completeness	0.46	0.47	0.26	0.38	0.43	0.45	0.26	0.37	0.33	0.35	0.25	0.32	0.27	0.3	0.19	0.29	0.18	0.15	0.13	0.16		
	R_{free}	0.45	0.44	0.08	0.11	0.46	0.46	0.07	0.10	0.50	0.49	0.06	0.08	0.51	0.51	0.04	0.06	0.53	0.53	0.02	0.04		
	R_{work}	0.42	0.41	0.08	0.10	0.42	0.41	0.07	0.10	0.43	0.43	0.06	0.08	0.43	0.42	0.04	0.06	0.44	0.44	0.02	0.04		
PHENIX AutoBuild	Completeness	0.72	0.72	0.13	0.3	0.71	0.74	0.14	0.29	0.58	0.56	0.19	0.38	0.52	0.55	0.18	0.33	0.38	0.42	0.17	0.27		
	R_{free}	0.29	0.29	0.04	0.10	0.32	0.31	0.05	0.10	0.38	0.38	0.06	0.12	0.39	0.39	0.05	0.09	0.43	0.42	0.05	0.08		
	R_{work}	0.27	0.27	0.03	0.09	0.28	0.28	0.05	0.09	0.33	0.33	0.06	0.10	0.32	0.32	0.04	0.08	0.36	0.35	0.05	0.06		
SHELXE	Completeness	0.84	0.85	0.11	0.25	0.71	0.67	0.15	0.35	0.38	0.34	0.19	0.31	0.2	0.2	0.09	0.18	0.12	0.13	0.07	0.13		
	R_{work}	0.46	0.45	0.01	0.03	0.47	0.48	0.02	0.04	0.51	0.52	0.02	0.06	0.53	0.52	0.01	0.03	0.53	0.52	0.01	0.02		

0.0 0.2 0.0 0.06

Figure 8
Mean and SD of the real and predicted structure evaluation measures for the MR data sets grouped based on resolution, with the number of data sets in each group shown in parentheses. The entries are shaded based on the difference between the real (R) and predicted (P) results.

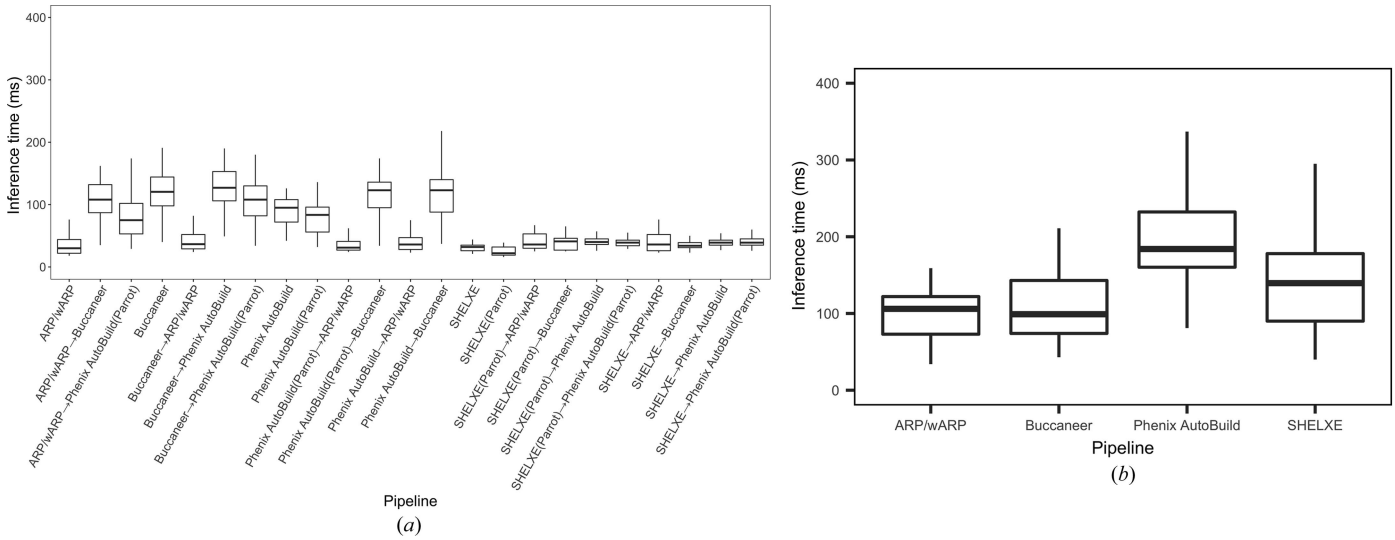


Figure 9
Inference time for the predictive model for individual pipelines and pipeline combinations. For each data set in the JCSG experimental phasing and MR data sets, the inference time is the total time taken to predict the structure completeness, R_{free} and R_{work}. (*a*) Inference time for the JCSG experimental phasing data sets and (*b*) inference tine for the MR data sets.

first group for structure completeness, R_{free} and R_{work}, respectively. For the MR data set the percentages were 60%, 69% and 87%, respectively.

Fig. 9 shows the inference time of the predictive model for individual pipelines and pipeline combinations for the JCSG experimental phasing and MR data sets. The inference time is the total time taken to predict the structure completeness and R_{free}/R_{work}. The *SHELXE* variants for the JCSG experimental phasing data set and *ARP/wARP* and *Buccaneer* for the MR data set have the lowest inference times.

3.3. Evaluation of the recommended pipeline variant

To further evaluate our predictive model, we analysed the potential benefits of using the pipeline variant recommended by the model, *i.e.* the pipeline variant predicted to achieve the best completeness or R_{free}/R_{work} for each of the data sets.

To this end, we first analysed the time savings that can be achieved by using the recommended pipeline variant instead of running all of the pipeline variants in order to obtain the best possible structure. Fig. 10 shows the total execution time when running all of the pipeline variants and when only the pipeline recommended by our predictive model was run. The time saved (on the powerful high-performance cluster mentioned in Section 2.5) was up to 20 h for a small protein structure and up to 60 h for large structures. When these pipeline variants were ran in parallel on our high-performance cluster, this time saving was reduced; however, running the recommended pipeline still saved up to 30 h when building large structures.

Next, we analysed how close the completeness and R_{free}/R_{work} of the protein structure built by the recommended pipeline variant was to the best completeness and R_{free}/R_{work} values achievable by running all of the pipeline variants. Figs. 11 and 12 present the results of this analysis for the JCSG experimental phasing and MR data sets, respectively. These results show that the recommended pipeline variant built protein structures with a completeness, R_{free} and R_{work} within only 1% of those of the best pipeline for 32%, 50% and 59% of the JCSG experimental phasing data sets and 70%, 99% and 71% of the MR data sets, respectively, and within only 5% of those of the best pipeline for 52%, 78% and 93% of the JCSG experimental phasing data sets and 83%, 100% and 87% of the MR data sets, respectively.

Finally, for each of the 15 research papers that we could find for our testing MR data sets that mentioned the pipeline used to build the protein structure, we compared the pipeline used in the paper with the pipeline variant recommended by our predictive model. To ensure a fair comparison, we ran the

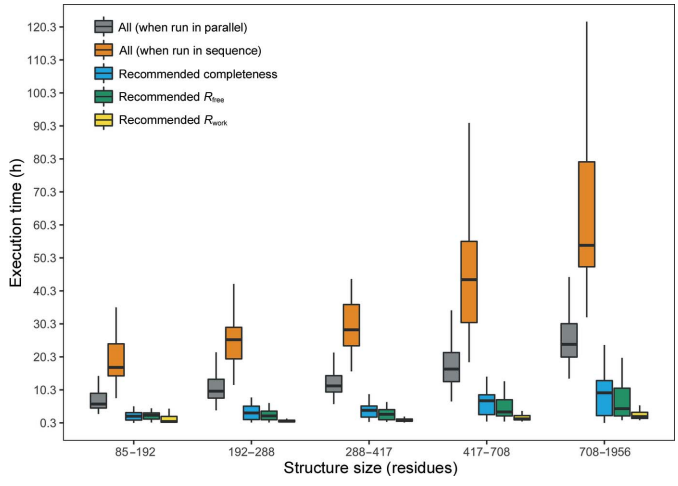


Figure 10
Execution time required to run all of the pipeline variants (in parallel and in sequence) versus the execution time required to run the pipeline recommended by the predictive model (for best completeness, best R_{free} and best R_{work}) for the JCSG experimental phasing data sets.

pipeline used in the paper and the pipeline recommended by our predictive model using the same search model to obtain initial phases for each structure. This search model could not be the same as that used for the PDB-deposited structure, which is unavailable.

Fig. 13 presents the structure completeness achieved by the pipeline that was chosen to solve the protein structure when deposited in the PDB compared with the completeness achieved by our recommended pipeline for each of these MR data sets. As shown in this figure, our recommended pipeline

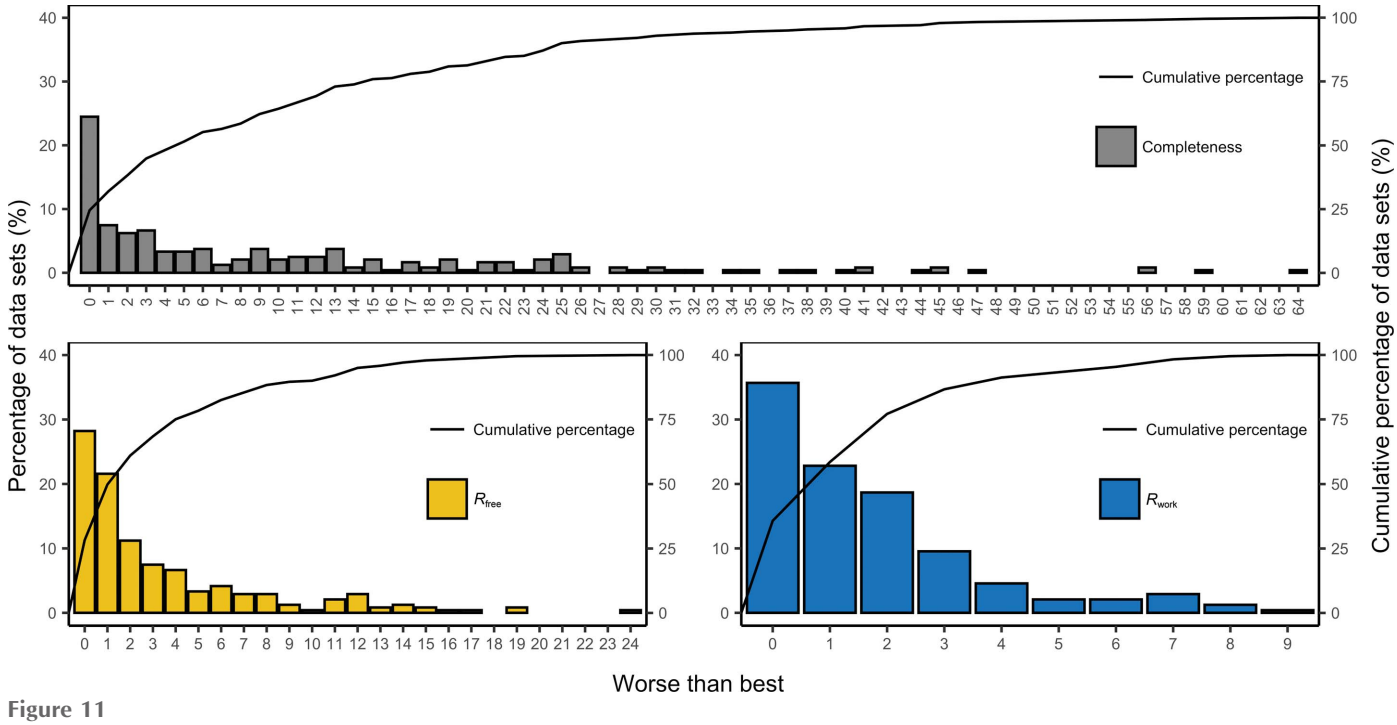


Figure 11
Difference between the best completeness, R_{free} and R_{work} achieved by running all of the pipeline variants and running the recommended pipeline variant for the JCSG experimental phasing data sets. The percentage of the data sets for each difference group is shown on the left and the cumulative percentage is shown on the right.

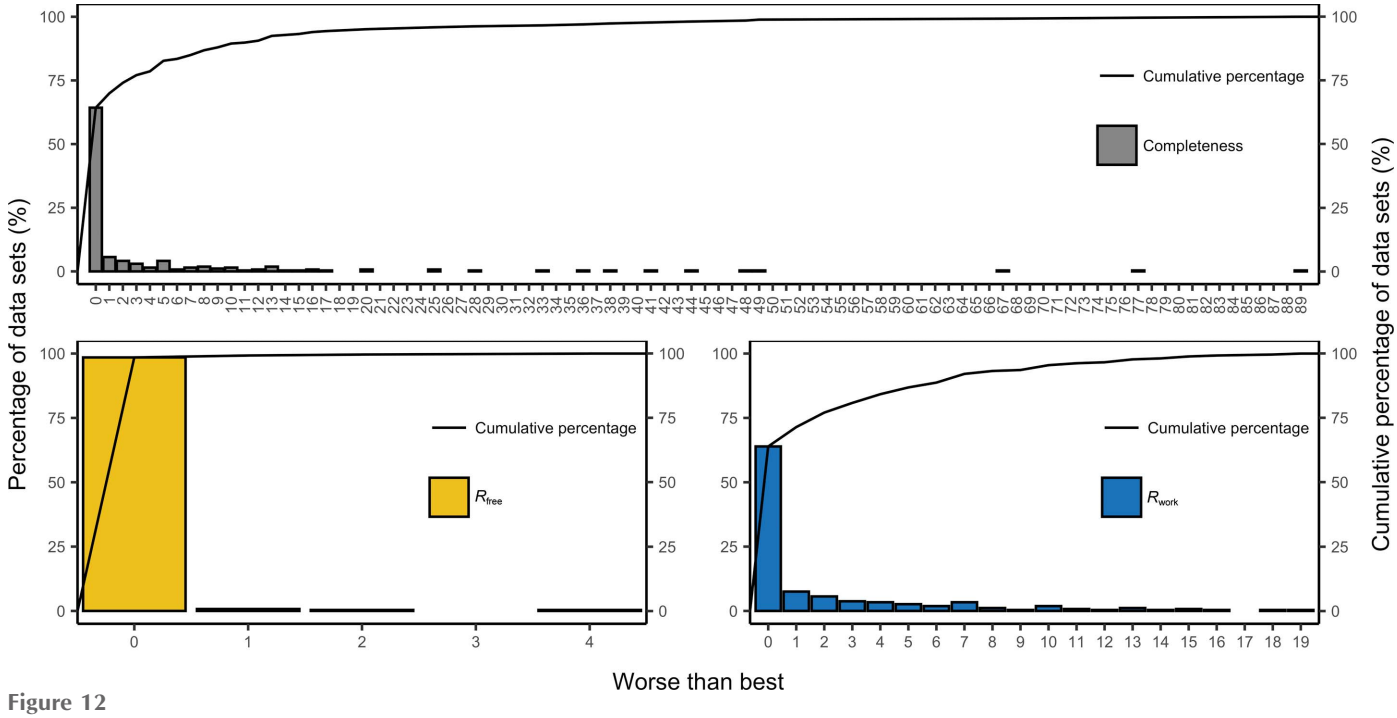


Figure 12
Difference between the best completeness, R_{free} and R_{work} achieved by running all of the pipeline variants and running the recommended pipeline variant for the MR data sets. The percentage of the data sets for each difference group is shown on the left and the cumulative percentage is shown on the right

PDB ID	Used		Recommended		
	Pipeline variant	Real completeness	Pipeline variant	Real completeness	Predicted completeness
4YVO	ARP/wARP	98.32	SHELXE	97.48	91.45
3MOK	ARP/wARP	97.8	ARP/wARP	97.8	93.75
2VMH	ARP/wARP	97.28	SHELXE	97.96	91.57
5N13	ARP/wARP	97.22	SHELXE	100.0	90.07
5WRT	Phenix AutoBuild	93.58	Phenix AutoBuild	93.58	60.62
5Y3D	Phenix AutoBuild	83.03	Phenix AutoBuild	83.03	70.88
6GP5	ARP/wARP	64.8	Phenix AutoBuild	88.48	84.55
3TZE	Buccaneer	57.63	Phenix AutoBuild	85.16	75.43
4D70	Buccaneer	40.62	Phenix AutoBuild	80.31	83.62
5IXG	Phenix AutoBuild	33.63	SHELXE	76.59	79.82
2XSJ	Buccaneer	10.22	Phenix AutoBuild	0.11	44.58
6FDH	ARP/wARP	0.45	SHELXE	91.82	56.39
2CQT	ARP/wARP	0.06	Phenix AutoBuild	81.2	34.73
5ZWP	ARP/wARP	0	SHELXE	94.69	76.69
3P7Y	ARP/wARP	0	SHELXE	94.21	81.96

Figure 13
Real structure completeness achieved by the pipeline that was used to solve the protein structure when deposited in the PDB and by the pipeline recommended by the predictive model for the MR data sets.

achieved better completeness than the other pipeline for ten of the 15 protein structures, and an identical completeness for three additional structures for which the predictive model recommended the same pipeline as that used to build the PDB structure. The recommended pipeline achieved worse completeness for only two of the 15 protein structures (with a decrease in completeness of less than 1% for one of these).

4. Discussion

We have presented a predictive model of the performance of four widely used protein model-building pipelines and of their pairwise combinations. We have separately trained this predictive model for both experimental phasing and molecular-replacement data sets and for three commonly used structure evaluation measures. Using this predictive model, we aim to help users choose the best pipeline for solving their protein structure based on the features of their starting data, to encourage them to use pipelines which may be less familiar to them and to increase the joint use of multiple pipelines, as doing so is likely to yield a more complete and more refined structure.

The features were calculated in scale-dependent measures; however, scale-independent measures are more natural in the crystallographic context. The scale-dependent measures were implemented first, yielding almost indistinguishable results. We assume that this is due to the machine-learning model effectively factoring out scale internally.

The MAE and RMSE analysis showed that $R_{\rm free}$ and $R_{\rm work}$ are more predictable than structure completeness in both experimental phasing and MR data sets. This unpredictability differs between the pipeline variants, suggesting that the electron-density map features have different effects on the performance of the pipelines. The predictability of pipelines involving *Phenix Autobuild* tends to be higher, which is likely to be due to the use of multiple models to offset stochastic effects. Both the MAE and RMSE for our predictive model are significantly lower than the MAE and RMSE for the training data set median used by the baseline, *Zero-R* predictive model.

When comparing the individual data sets by using the mean and SD for the real and predicted structure evaluation measures at high resolution, which is considered to be an easier case, the performance of the pipelines is more predictable than at low resolution. When the data sets become worse in terms of resolution (which typically also means that the phases become worse), the difference in SD between the real and predicted results becomes larger.

The pipeline variant predicted to build the best protein structure frequently produced structures with the same or similar completeness and/or $R_{\rm free}/R_{\rm work}$ as the best pipeline variant. Moreover, using the pipeline variant recommended by our predictive model save days of pipeline execution time on high-specification computers, and the time saved increases when the protein structure is larger. Finally, the predictive model can be used to try massive search models in MR cases, enabling the selection of good initial phases (Simpkin *et al.*, 2018; Bibby *et al.*, 2012).

Future work will consider a multi-task method for predicting structure completeness, $R_{\rm free}$ and $R_{\rm work}$, and will combine the ML models into a single model. We envisage that this could lead to more accurate predictions and to better pipeline ranking. Moreover, we will explore additional ML algorithms, for example *XGBoost* (Chen & Guestrin, 2016), as this may improve our predictive model.

5. Availability

We implemented the predictive model described in the paper as a web application that is publicly available and free to use at http://www.robin-predictor.org. The source code for the application is available at https://doi.org/10.15124/ee9d169f-c34b-44f2-8c75-3b68e7cd68a8.

Acknowledgements

This project was undertaken on the Viking Cluster, a high-performance computing facility provided by the University of York. We are grateful for the computational support received from the University of York's High Performance Computing service, Viking Cluster and Research Computing team. This work used advanced computing resources from the IN2P3-IRES resource centre of the EGI federation for hosting the

predictive model web application. The services are co-funded by the EGI-ACE project (grant number 101017567).

Funding information

Funding for this research was provided by: University of Tabuk (scholarship to Emad Alharbi); Biotechnology and Biological Sciences Research Council (grant No. BB/S005099/1 to Paul Bond and Kevin Cowtan).

References

Alharbi, E., Bond, P. S., Calinescu, R. & Cowtan, K. (2019). *Acta Cryst.* D**75**, 1119–1128.

Alharbi, E., Calinescu, R. & Cowtan, K. (2020). *Acta Cryst.* D**76**, 814–823.

Ashforth, B. E. & Mael, F. (1989). *Acad. Manag. Rev.* **14**, 20–39.

Bedem, H. van den, Wolf, G., Xu, Q. & Deacon, A. M. (2011). *Acta Cryst.* D**67**, 368–375.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* D**68**, 1622–1631.

Bond, P. S., Wilson, K. S. & Cowtan, K. D. (2020). *Acta Cryst.* D**76**, 713–723.

Breiman, L. (1996). *Mach. Learn.* **24**, 123–140.

Breiman, L. (2001). *Mach. Learn.* **45**, 5–32.

Brünger, A. T. (1992). *Nature*, **355**, 472–475.

Chen, T. & Guestrin, C. (2016). *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. New York: ACM.

Choudhary, R. & Gianey, H. K. (2017). *2017 International Conference on Machine Learning and Data Science (MLDS)*, pp. 37–43. Piscataway: IEEE.

Cortes, C. & Vapnik, V. (1995). *Mach. Learn.* **20**, 273–297.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Cowtan, K. (2008). *Acta Cryst.* D**64**, 83–89.

Dauter, M. & Dauter, Z. (2017). *Methods Mol. Biol.* **1607**, 349–356.

Evans, P. & McCoy, A. (2008). *Acta Cryst.* D**64**, 1–10.

Frank, E. & Bouckaert, R. R. (2009). *Advances in Machine Learning*, edited by Z.-H. Zhou & T. Washio, pp. 65–81. Berlin, Heidelberg: Springer-Verlag.

Frank, E., Hall, M. A. & Witten, I. H. (2016). *The Weka Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques'*. Burlington: Morgan Kaufmann.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). *ACM SIGKDD Explor. Newsl.* **11**, 10–18.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R., Wyckoff, H. & Phillips, D. C. (1958). *Nature*, **181**, 662–666.

Krissinel, E. (2012). *J. Mol. Biochem.* **1**, 76–85.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.

Langer, G. G., Hazledine, S., Wiegels, T., Carolan, C. & Lamzin, V. S. (2013). *Acta Cryst.* D**69**, 635–641.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Cryst.* D**75**, 861–877.

McCoy, A. J. & Read, R. J. (2010). *Acta Cryst.* D**66**, 458–469.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229–244.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nat. Struct. Biol.* **6**, 458–463.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Simpkin, A. J., Simkovic, F., Thomas, J. M. H., Savko, M., Lebedev, A., Uski, V., Ballard, C., Wojdyr, M., Wu, R., Sanishvili, R., Xu, Y., Lisa, M.-N., Buschiazzo, A., Shepard, W., Rigden, D. J. & Keegan, R. M. (2018). *Acta Cryst.* D**74**, 595–605.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst.* D**65**, 582–601.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.

Thorn, A. & Sheldrick, G. M. (2013). *Acta Cryst.* D**69**, 2251–2256.

Usón, I. & Sheldrick, G. M. (2018). *Acta Cryst.* D**74**, 106–116.

Vollmar, M., Parkhurst, J. M., Jaques, D., Baslé, A., Murshudov, G. N., Waterman, D. G. & Evans, G. (2020). *IUCrJ*, **7**, 342–354.