New method to compute $R_{complete}$ enables maximum likelihood refinement for small datasets

Jens Luebben[a] and Tim Gruene[b,1]

[a]Institute for Inorganic and Applied Chemistry, D-20146 Hamburg, Germany; and [b]Department of Structural Chemistry, Georg-August-University Göttingen, D-37077 Göttingen, Germany

Edited by Axel T. Brunger, Stanford University, Stanford, CA, and approved June 9, 2015 (received for review February 1, 2015)

The crystallographic reliability index $R_{complete}$ is based on a method proposed more than two decades ago. Because its calculation is computationally expensive its use did not spread into the crystallographic community in favor of the cross-validation method known as R_{free}. The importance of R_{free} has grown beyond a pure validation tool. However, its application requires a sufficiently large dataset. In this work we assess the reliability of $R_{complete}$ and we compare it with *k*-fold cross-validation, bootstrapping, and jackknifing. As opposed to proper cross-validation as realized with R_{free}, $R_{complete}$ relies on a method of reducing bias from the structural model. We compare two different methods reducing model bias and question the widely spread notion that random parameter shifts are required for this purpose. We show that $R_{complete}$ has as little statistical bias as R_{free} with the benefit of a much smaller variance. Because the calculation of $R_{complete}$ is based on the entire dataset instead of a small subset, it allows the estimation of maximum likelihood parameters even for small datasets. $R_{complete}$ enables maximum likelihood-based refinement to be extended to virtually all areas of crystallographic structure determination including high-pressure studies, neutron diffraction studies, and datasets from free electron lasers.

structure determination | reliability index | maximum likelihood refinement | overfitting | model bias

The quality of crystallographic models is described by several quality indicators. Both for small and macromolecular structure deposition, the crystallographic reliability index $R1$ must be provided (1, 2). It is calculated for the dataset H of observations and a structural model as

$$R1 = \frac{\sum_{\mathbf{h}\in H} \left| |F_{obs}(\mathbf{h})| - |F_{calc}(\mathbf{h})| \right|}{\sum_{\mathbf{h}\in H} |F_{obs}(\mathbf{h})|}. \quad [1]$$

Depending on the data-to-parameter ratio, $R1$ is affected by more or less severe overfitting (3, 4). To overcome this problem, cross-validation was introduced into crystallography (5–9). For cross-validation in crystallography, a certain fraction of the observations, typically 5–10%, are withheld as test set T and never used for model building and refinement. They are only used to calculate the reliability index R_{free}:

$$R_{free} = \frac{\sum_{\mathbf{h}\in T} \left| |F_{obs}(\mathbf{h})| - |F_{calc}(\mathbf{h})| \right|}{\sum_{\mathbf{h}\in T} |F_{obs}(\mathbf{h})|}. \quad [2]$$

R_{free} is much less affected by overfitting and since its introduction it has gained importance beyond validation of the structural model. It is used to optimize weights for restrained refinement (4, 10–13). The concept of R_{free} paved the way for maximum likelihood methods in crystallography. It was shown that the estimation of maximum likelihood parameters based on the test set T provides much better accuracy than that based on the data used during refinement (14–16).

Cross-validation reduces the bias of a statistic (17, 18) but can show large variance, especially when T is small (8, 17). The relative error of the crystallographic R_{free} was established as $\sigma(R_{free}) = R_{free}/\sqrt{2|T|}$ (19). The test set should hold at least 500 data points so that $\sigma(R_{free})/R_{free} \leq 0.032$. Maximum likelihood methods estimate parameters in resolution bins, and a total of $|T| = 2{,}000$ may be required for robust estimation. To assess the accuracy of a statistic such as R_{free} one could apply *k*-fold cross-validation, the bootstrap method, and the jackknife method (7, 8, 17, 20). *k*-fold cross-validation divides the dataset into k approximately equally sized and pairwise disjoint subsets $H = \cup_{i=1}^{k} T_i$ and cross-validation is carried out for each of the parts separately. $\langle R_{free} \rangle$ and $\sigma(R_{free})$ are calculated from the k resulting R_{free}. As mentioned above, for small test sets, that is, $k \to |H| \Leftrightarrow |T_i| \to 1$, $\sigma(R_{free})$ becomes very large. Both the bootstrap and the jackknife method reduce the variance of an estimator like R_{free}. The jackknife artificially creates $|H|$ datasets $H_i := H \setminus \{\mathbf{h}_i\}$, that is, with the ith data point removed, so that

$$R^i_{jack} = \frac{\sum_{\mathbf{h}\in H_i} \left| |F_{obs}(\mathbf{h})| - |F_{calc}(\mathbf{h})| \right|}{\sum_{\mathbf{h}\in H_i} |F_{obs}(\mathbf{h})|}. \quad [3]$$

The estimator is calculated as arithmetic mean

$$R_{jack} = \frac{1}{|H|} \sum_i R^i_{jack} \quad [4]$$

with the jackknife estimate of variance (8)

Significance

Modern crystallographic structure determination uses maximum likelihood methods. They rely on error estimates between the work model and the unknown target based on a small fraction of the data. This can introduce a large uncertainty and, even worse, restricts the method to projects where sufficient data are available. We investigate the $R_{complete}$ method. It enables the use of all data for error estimation. It reduces the uncertainty associated with the conventional R_{free} approach for small datasets. We show that our approach reduces the effect of overfitting. This enables maximum likelihood methods to be extended to a much wider field of applications, including free electron laser experiments, high-pressure crystallography, and low-resolution structures.

Author contributions: T.G. designed research; J.L. and T.G. performed research; J.L. and T.G. contributed new reagents/analytic tools; J.L. and T.G. analyzed data; and T.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: tg@shelx.uni-ac.gwdg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1502136112/-/DCSupplemental.

$$\sigma^2_{\text{jack}} = \frac{|H|-1}{|H|}\sum_i \left(R^i_{\text{jack}} - R_{\text{jack}}\right)^2. \quad [5]$$

Bootstrapping differs from jackknifing in that the bootstrap datasets H_i are generated from H by random sampling with replacement with $|H_i| = |H|\ \forall i$. Thus, one could calculate R^i_{boot} up to $(2|\text{H}|-1)!/(|\text{H}|!(|\text{H}|-1)!)$ times, although a few thousand samples are usually sufficient. Let b the number of bootstrap samples. The bootstrap R value and its estimate of variance are defined as (8)

$$R_{\text{boot}} = \frac{1}{b}\sum_i R^i_{\text{boot}} \quad [6]$$

$$\sigma^2_{\text{boot}} = \frac{1}{b-1}\sum_i \left(R^i_{\text{boot}} - R_{\text{boot}}\right)^2. \quad [7]$$

None of these methods avoids the deficiency that the variance of the respective R value is large when the test sets T_i are very small. This was already shown in ref. 6 and can be seen in *SI Appendix*, Fig. S1: Because the R value has a lower bound of 0, large outliers will drag any mean up from its real value. Our interest in alternative ways to calculate R_{free} arose during the project in ref. 21. Macromolecular neutron datasets are often small with low data completeness. Leaving out 500 or more data points during model building and refinement would destabilize these processes and thus impede the quality of the final model. In high-pressure crystallography the situation is even worse because the incompleteness of the data owing to shadows from the experimental setup is systematic and leads to data-to-parameter ratios too low to rely on $R1$ alone. The entire dataset may have fewer than 500 observations (22).

To circumvent these difficulties, Brünger (6) suggested the method of R_{complete} validation: Instead of creating the test sets required for k-fold cross-validation at the very beginning after data collection, they are created only when the calculation of a reliable R value is needed. Strictly speaking, the R_{complete} method is not cross-validation because the statistic of interest, the R value, is not calculated as mean from a number of refinement runs, but in analogy to Eq. **1** from the entire dataset, as will be detailed below. The critical point for using R_{complete} is the question of how to reduce the effect of overfitting from the structural model after it was refined against all data points. Proper cross-validation as realized with R_{free} does not share this problem because the data from the test set are never used during refinement and model building throughout the entire process from data acquisition to publication. Brünger (6) suggested simulated annealing. Others apply random parameter perturbation (13, 23, 24). A third option that has been discussed in the crystallographic community was suggested by Tickle (25): Refinement of a structural model to convergence should reduce the effect of overfitting against any observation not used during such a refinement run. Here we concentrate on "Tickle's conjecture" for its obvious advantage: Both simulated annealing and parameter perturbation introduce random shifts into the structural model. In the worst case this may result in a nonchemical structure shown in *SI Appendix*, Fig. S4. It may result in several structures that differ significantly, that is, so that a biologist or chemist would speak of different structures. Hence, one could no longer speak of the structural model and its R value. In this article we present a series of experimental approaches that show that the R_{complete} method results in as little bias as R_{free}. We show that R_{complete} varies much less than R_{free} in the case of very small datasets. We confirm Tickle's conjecture and, thus, in the light of ref. 15, our work enables maximum likelihood-based refinement of crystallographic models against small datasets as in neutron diffraction, high-pressure crystallography, low-resolution macromolecular studies, supramolecular chemistry, and, at its current state, structural data from free electron lasers.

This manuscript is structured as follows. The *Methods* section first describes how we calculate R_{complete}. The following subsections describe the experiments we carried out. The *Results* section repeats all of the subsections with the respective results. The description is held as general as possible. The details about programs and parameters are given in *SI Appendix*.

Methods

The datasets used in this work are summarized in Table 1 including their IDs used throughout this manuscript. Throughout this manuscript we use the terms "working set" and "test set," defined below. These terms are commonly used in crystallography. In other contexts the equivalent terms "training set" and "validation set" are used, respectively. In the presence of a test set, the reliability index $R1$ defined in Eq. **1** is calculated only from the observations used in refinement, that is, only for $\mathbf{h} \in H \backslash T$.

Data Preparation and Calculation of R_{complete}. The starting point is a merged dataset H and the structural model P (i.e., the set of parameters for which R_{complete} is to be calculated). The model should have been refined against the entire dataset until convergence. The dataset is randomly partitioned into k test sets T_i so that $H = \cup T_i$ and $T_i \cap T_j = \emptyset\ \forall i, j$. If k does not divide $|H|$, the last test set is smaller than the other test sets. For better readability of the manuscript we generally do not point out this fact when abbreviating the test set size as $|T_i|$. The structural model is refined until convergence against each of the working sets $W_i := H \backslash T_i$, resulting in the structural models P_i. Then, related to equation 16 in ref. 6,

$$R_{\text{complete}} := \frac{\sum_i \sum_{\mathbf{h}\in T_i} \left||F_{obs}(\mathbf{h})| - |F_{calc}(\mathbf{h})|\right|}{\sum_i \sum_{\mathbf{h}\in T_i} |F_{obs}(\mathbf{h})|}. \quad [8]$$

By construction, R_{complete} is calculated from the entire dataset. In the numerator of Eq. **8** $|F_{calc}(\mathbf{h})|$ is calculated from from the model P_i for an observation $\mathbf{h}$, which was not used in the refinement of model P_i.

Note the difference to k-fold cross-validation:

$$\langle R_{\text{free}} \rangle := \frac{1}{k}\sum_{i=1}^{k} \frac{\sum_{\mathbf{h}\in T_i} \left||F_{obs}(\mathbf{h})| - |F_{calc}(\mathbf{h})|\right|}{\sum_{\mathbf{h}\in T_i} |F_{obs}(\mathbf{h})|}. \quad [9]$$

The following subsections describe the experiments we carried out.

Stability with Respect to the Test Set Size. Both R_{complete} and $\langle R_{\text{free}} \rangle$ were calculated for datasets 5 and 6′. The partition size was varied between 1 and 500 (see *SI Appendix*, Fig. S1 and Tables S1 and S2).

Stability of R_{complete} with Partition. Unless $k = |\text{H}|$, R_{complete} may depend on the partitioning of the dataset. We randomly partitioned datasets 6′, 4, and 7 20 times and calculated R_{complete} for each partition to assess how much it varies with the partitioning. Results are listed in *SI Appendix*, Tables S3–S5.

Validation I: How "free" Is R_{complete}? Dataset 8 was partitioned into 90 test sets T_i. The test sets and the working sets $W_i = H \backslash T_i$ were separated to ensure the test sets were not used in any of the subsequent steps. For each working set, the structure was automatically solved with standard single-wavelength anomalous dispersion of S atoms (S-SAD) and expanded to a poly-Alanine model. Each poly-Alanine model was subsequently further completed by automated model building with the amino acid sequence as input. These models were finally refined with 200 cycles conjugate gradient least-squares refinement. R^i_{free} was calculated with each model against its test set. Because the test set was never used during the creation of the structural model, R^i_{free} is free from overfitting. For each structural model, R^i_{complete} was calculated as described above. Results are listed in *SI Appendix*, Tables S7–S9.

As a second type of experiment the small-molecule datasets 2 and 3 were each partitioned into 20 test sets T_i and solved by standard direct methods against the working sets $W_i = H \backslash T_i$. Dataset 3 is similar to dataset 2 except for a disordered solvent molecule, resulting in greater fluctuations. Each of the 20 resulting structural models was refined against its respective work set W_i with 10 cycles of least-squares minimization. The R_{free} values were calculated from each structural model against its test set. Results are listed in *SI Appendix*, Tables S10 and S11.

Table 1. List of datasets

Dataset	1	2	3	4	5	6	7	8
Name	n/a	n/a	n/a	Ciprofloxacin	Hormaomycin	Insulin	Insulin	Elastase
SG	$P\bar{1}$	$P\bar{1}$	$P2_1/c$	$P\bar{1}$	$P2_1$	$I2_13$	$I2_13$	$P2_12_12_1$
d_{min}, Å	0.44	1.00	1.00	0.70	1.02	1.10	2.30	1.37
No. of atoms	56	60	60	60	215	436	802	2,163
No. of data	42,997	5,117	5,069	6,227	7,800	32,598	3,747	44,784
Source	31	32	32	33	34	SI	35	SI

Dataset 6′ is identical to dataset 6 but with d_{min} reduced to 1.9 Å with 6,533 observations. Dataset 7 is the neutron dataset. n/a, not applicable. No. of atoms, number of non-H atoms per asymmetric unit; for dataset 7, number of all atoms. No. of data, number of unique observations; SG, space group; SI, see *SI Appendix*.

Validation II: Comparison with Calculated Data. Diffraction data were calculated from the structural model of dataset 6′ to $d_{\min} = 1.9$ Å and from the structural model of dataset 4 to $d_{\min} = 0.7$ Å. Hydrogen atoms were not included for the calculations. We checked that in both cases $R1 = 0.0$ against the calculated data without refinement. For the structural model of dataset 6′, the oxygen atoms of four water molecules were removed and two oxygen atoms were replaced as sodium atoms. For the structural model of dataset 4, the oxygen atom of one water molecule was replaced as sodium atom (i.e., the model contains three electrons too many compared with the data). The $R1$ values were calculated without refinement, thus representing the real $R1$ value. The small molecule from dataset 4 was refined with 50 cycles of least-squares minimization, and the macromolecule from dataset 6′ was refined with 30 cycles of conjugate gradient least-squares minimization. $R_{complete}$ was calculated with $|T_i| = 10$ for dataset 4 and $|T_i| = 30$ for dataset 6′. Whereas the experiments of the previous subsection address the resistance of $R_{complete}$ against overfitting, the experiments of this subsection also address the effects of structural model bias. The R values are listed in *SI Appendix*, Table S12.

Effect of Parameter Perturbation. We use the symbol X for the amount of random perturbation of coordinates and atomic displacement parameters of the structural models P_i. Coordinates of atoms not on special positions were displaced by an average distance X Å in a random direction. When applicable hydrogen atoms were generated after the application of shifts. No shifts were applied to fixed coordinates (e.g., for special positions). Isotropic atomic displacement parameters and the main diagonal elements U_{ii} were multiplied by a random factor so that they change by an average of X Å^2. Off-diagonal atomic displacement parameters U_{12}, U_{13}, and U_{23} for anisotropic atoms were not modified to avoid the generation of matrices with physically impossible nonpositive eigenvalues.

To investigate how random parameter perturbation reduces the effect of overfitting from the structural model, we created a regular grid of dummy atoms. We used the cell from dataset 6 as an example of a noncentrosymmetric space group and from dataset 1 as an example of a centrosymmetric space group. The number of grid points corresponds roughly to the number of atoms for the respective structure. This ensures realistic data-to-parameter ratios. To introduce overfitting the set of parameters was refined to convergence without restraints against the respective data at various resolution cut-offs (see *SI Appendix*, Tables S13 and S14, respectively). The parameters of both overfitted structural models were randomly perturbed with an amplitude X varying from 0.1 to 1.0 and their $R1$ values was calculated against all data up to the given resolution. The perturbation was repeated 500 times and the $R1$ values averaged.

The numerical results are listed per resolution cut-off in *SI Appendix*, Tables S15–S20 for dataset 6 and in *SI Appendix*, Tables S21–26 for dataset 1.

Influence of Parameter Perturbation on Convergence Rate. The value $R_{complete}$ was monitored for the structural model of dataset 6′ with varying amplitudes $X \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$ of perturbation. The number of least-squares refinement cycles is listed in *SI Appendix*, Table S27; 4,000 and 10,000 cycles were calculated only for $X = 0.0$ and $X = 0.3$.

Results

Stability with Respect to the Test Set Size. Cross-validation and especially k-fold cross-validation are known to produce values with theoretically little bias, yet with small test sets they suffer from large variance (17). In addition to the large variance, the averaged mean of any value with a lower bound but no upper bound such as the crystallographic reliability index will probably be pushed up by very large outliers. We compared the behavior of $\langle R_{free} \rangle$ with that of $R_{complete}$ for small test set sizes. For this purpose we calculated both values for the structural models of datasets 5 and 6′ in dependence of the test set size. Our results show that $R_{complete}$ is independent of the test set size. The mean value averaged over all tested set sizes is 0.1653 ± 0.0003 for dataset 5 and 0.2239 ± 0.0006 for dataset 6′. $\langle R_{free} \rangle$, on the contrary, shows the expected extremely large variance. More importantly, its value rises when the test set size is below 20, a behavior known since the introduction of R_{free} (6). The bootstrapped values R_{boot} are listed in *SI Appendix*, Tables S1 and S2, respectively. They replicate the values of $\langle R_{free} \rangle$ with an SE one order of magnitude smaller. Hence, bootstrapping does not avoid the instability of $\langle R_{free} \rangle$ for small test sets. However, $R_{complete}$ is reliable even when the entire dataset except a single observation is used for refinement. At the suggested lower limit for the test set size $|T_i| = 500$ (6), R_{free} has a reasonably narrow range within $15.5\% < R_{free} < 17.7\%$ for dataset 5 and $19.3\% < R_{free} < 24.5\%$ for dataset 6′. However, with $|T_i| = 100$, the range increases to $13.2\% < R_{free} < 20.5\%$ for dataset 5 and $15.5\% < R_{free} < 33.2\%$ for dataset 6′. Note that these are the ranges for one particular partition. They do not cover all possible test sets except for $|T_i| = 1$. Because for the conventional R_{free} the free set is chosen randomly, one might have ended up with any such value for the same model. This illustrates why we describe R_{free} as unstable. $R_{complete}$ can be calculated from any convenient test set size to optimally balance between computation time and data completeness used for refinement.

Stability of $R_{complete}$ with Partition. Except for $|T_i| = 1$ there are a large number of possible partitions for a dataset, and $R_{complete}$ might vary depending on which partition is used. We computed $\langle R_{complete} \rangle$ and $\sigma(R_{complete})$ from 20 different partitionings. We find $\langle R_{complete} \rangle = 21.92\% \pm 0.02\%$ for dataset 6′, $32.64\% \pm 0.09\%$ for dataset 7, and $4.88\% \pm 0.01\%$ for dataset 4 (i.e., $R_{complete}$ does not depend on the choice of partition). We conclude that $R_{complete}$ can be calculated from a single partitioning. In combination with the previous subsection, the size of the subsets of the partitioning of the dataset can be chosen as convenient and only a single partition needs to be considered.

Validation I: How "free" Is $R_{complete}$? One of the basic questions for the relevance of our work is whether the procedure described above really reduces the effect of overfitting (i.e., whether $R_{complete}$ is really "free"). We carried out proper k-fold cross-validation in the sense that we calculated $\langle R_{free} \rangle$ from test sets that were never used for model building or refinement throughout the entire process.

Dataset 8 was solved from 90 different working sets by SAD phasing, density modification, and model completion by autobuilding. Each of the resulting 90 structural models was refined to convergence. For each structural model a proper R_{free} was

calculated against its respective test set and $R_{complete}$ was calculated as described above.

Our calculations resulted in

$$\left\langle \frac{R_{complete}}{R_{free}} \right\rangle = 0.9866 \quad \sigma\left(\frac{R_{complete}}{R_{free}}\right) = 0.0427 \qquad [10]$$

$$\left\langle \frac{R_{complete}}{R1} \right\rangle = 1.1195 \quad \sigma\left(\frac{R_{complete}}{R1}\right) = 0.0040. \qquad [11]$$

Note that 90 structural models can have significant differences in the number of amino acids, the orientation of side chains, and so on. Therefore, the calculation of $\langle R_{complete}\rangle/\langle R_{free}\rangle$ is not meaningful. Within less than half an SD, $R_{complete} = R_{free}$ so that we consider $R_{complete}$ as free from overfitting as R_{free}. The ratio 1.12 between $R_{complete}$ and $R1$ indicates the effect of overfitting present in $R1$, as one would expect. When bootstrapping is applied to the ratio $R_{complete}/R_{free}$, the average value remains at 0.9866 with $\sigma_{boot} = 0.00439$ and $R_{complete} = R_{free}$ only within $3.1\sigma_{boot}$ (see *SI Appendix*). With bootstrapping as criterion we can consider $R_{complete}$ to slightly suffer from overfitting compared with proper cross-validation, but still much less than $R1$, underlining the value of $R_{complete}$ for validation.

To assess whether $R_{complete}$ correlates with the quality of the respective structural models, we calculated the average phase difference between each structural model and the fully refined structure. The correlation between $\langle\Delta\Phi\rangle_i$ and $R^i_{complete}$ for all 90 models is 99.1%, compared with only 74.1% between $\langle\Delta\Phi\rangle_i$ and R^i_{free}. The correlation between $\langle\Delta\Phi\rangle_i$ and $R1^i$ is 98.9% (i.e., for these high-quality data it compares with $R_{complete}$). We conclude that $R_{complete}$ is a good estimator for the quality of a structural model.

We repeated a similar experiment with the small molecule dataset 2. Despite two independent approaches the results are remarkably similar:

$$\left\langle \frac{R_{complete}}{R_{free}} \right\rangle = 0.9900 \quad \sigma\left(\frac{R_{complete}}{R_{free}}\right) = 0.0815 \qquad [12]$$

$$\left\langle \frac{R_{complete}}{R1} \right\rangle = 1.1064 \quad \sigma\left(\frac{R_{complete}}{R1}\right) = 0.0049. \qquad [13]$$

The large variation for the ratio with R_{free} once more underlines the greater stability of $R_{complete}$ compared with R_{free} for small test sets, $|T_i| = 256$ in this case. Bootstrapping the ratio between $R_{complete}$ and R_{free} with 20,000-fold resampling results in the same average ratio 0.9900 with $\sigma_{boot} = 0.0178$ (i.e., in this case $R_{complete} = R_{free}$ within $0.6\sigma_{boot}$).

Similarly, for dataset 3:

$$\left\langle \frac{R_{complete}}{R_{free}} \right\rangle = 0.9983 \quad \sigma\left(\frac{R_{complete}}{R_{free}}\right) = 0.0875 \qquad [14]$$

$$\left\langle \frac{R_{complete}}{R1} \right\rangle = 1.1149 \quad \sigma\left(\frac{R_{complete}}{R1}\right) = 0.0067. \qquad [15]$$

In this case, bootstrapping provides $\sigma_{boot} = 0.0190$ and thus $R_{complete} = R_{free}$ within only 0.09σ.

The $R_{complete}$ values for dataset 2, listed in *SI Appendix*, Table S10, clearly cluster about two values, 12.04% and 12.12%. Inspection of the structural models revealed that the structure solution step wrongly assigned one particular carbon atom, having six electrons, as a nitrogen atom, having seven electrons, in exactly those models with $R_{complete} = 12.12$ %. Neither R_{free} nor $R1$ reveal the same. This is an example where $R_{complete}$ is superior to both R_{free} and $R1$.

For dataset 3, $R_{complete}$ displays a similar sensitivity. It points at two outlier runs that neither $R1$ nor R_{free} make obvious. The disordered solvent molecule is a tetrahydrofuran, a five-membered ring with four carbon atoms and one oxygen atom. In all cases with $R_{complete} = 12.73\%$ as well as the run with $R_{complete} = 12.91\%$, an incorrect six-membered all-carbon ring was modeled. In the run with $R_{complete} = 13.23\%$, a five-membered all-carbon ring was modeled. The decreased value of $R_{complete}$ for the six-membered ring might be due to a better modeling of the disorder, but it may also be due to the addition of four parameters by the extra carbon atom. The run with $R_{complete} = 12.91\%$ contains another error: The oxygen of a second, ordered tetrahydrofuran molecule was assigned as nitrogen. Hence, in this case, $R_{complete}$ is capable of distinguishing two types of structures different by only one electron out of 385 in total.

Validation II: Comparison with Calculated Data. The computation of $R_{complete}$ provides a set of calculated structure factor amplitudes $|F_{calc}(\mathbf{h})|$ for the entire dataset H. With the $R_{complete}$ method $|F_{calc}(\mathbf{h})|$ is computed from a structural model that was not refined against the particular observation $\mathbf{h}$. We were interested in whether the structure factor amplitudes from the computation of $R_{complete}$ result in better electron density maps. Electron density maps are difficult to compare, the differences may be very subtle, and the map quality is affected by Fourier truncation errors as well as noise from missing low-resolution observations. For this reason we used calculated data from the structural models for datasets 4 and 6′, modified as described above.

The $R_{complete}$-based electron density map from dataset 4 has a stronger signal for the wrongly placed sodium than the conventional electron density map (see *SI Appendix*, Fig. S3*A*). Similar results are shown for dataset 6′ in *SI Appendix*, Fig. S3*B*. In both cases the $R_{complete}$-based map is less biased toward the structural model.

Because we were interested in whether parameter perturbation might have a different effect, we produced *SI Appendix*, Fig. S4, a nonchemical structure resulting from a perturbation amplitude of only $X = 0.6$. It illustrates why we do not recommend applying random parameter perturbation if one wishes to calculate the reliability index of one particular structural model. The next two sections illustrate this further.

Effect of Parameter Perturbation. The previous examples show that $R_{complete}$ is as good a quality indicator as R_{free} with the benefit that it can be computed for datasets with very few data at constant reliability. The $R_{complete}$ method we propose and assessed in this work uses refinement to convergence to reduce the effect of overfitting from the structural model. As mentioned in the introduction, alternatives have been suggested such as simulated annealing and random parameter perturbation (6, 13, 23). We addressed the question as to what extent random parameter perturbation affects the reduction of the effect of overfitting. For this purpose we created one set of parameters aligned on a regular grid for the centrosymmetric space group $P\bar{1}$ and similarly for the noncentrosymmetric space group $I2_13$. Neither of these sets of parameters contains any chemical information and refinement of these parameters is purely based on overfitting. Even $R1 = 0$ can be reached when the data-to-parameter ratio is well below 1.

The effect on the reduction of overfitting was checked by calculating $R1$ after random parameter perturbation. When only the coordinates are perturbed, the Wilson limits, 82.8% for centric and 58.6% for noncentric space groups (26), are hardly reached even with very large amplitudes $X = 1.0$, and only for very high data-to-parameter ratios. When both coordinates and atomic displacement parameters are perturbed, the situation is a little better, although even then the Wilson limit is only reached at high amplitudes X (see *SI Appendix*, Fig. S5). However, random parameter perturbation can severely compromise the

structural integrity of a model (see *SI Appendix*, Fig. S4). We do not recommend the use of random parameter perturbation for the computation of $R_{complete}$.

Influence of Parameter Perturbation on Convergence Rate. Although we already came to the recommendation not to use parameter perturbation for the calculation of $R_{complete}$, we were interested in the effect of random parameter perturbation on the rate of convergence. We used dataset 6′ and the corresponding structural model. The input structural model was refined to convergence with $R1 = 20.56\%$. The value of $R_{complete}$ was monitored with an increasing number of refinement cycles with five perturbation amplitudes $X = 0.0 \ldots 0.5$ applied both to the coordinates and the atomic displacement parameters. Parameter perturbation has no beneficial effect on the rate of convergence (see *SI Appendix*, Fig. S6 and Table S27). After 100 cycles of refinement with and without parameter perturbation $R_{complete}$ has reached the same value 23.8%, then fluctuates about this value. Graphs such as *SI Appendix*, Fig. S6 could be used to determine the number of refinement cycles needed to achieve the desired precision for $R_{complete}$.

Conclusions

Crystallographic studies make intensive use of the R_{free} concept: A structural model is cross-validated against a small test set. The data of the test set are never used for refinement or model building. Therefore, cross-validation with R_{free} is unaffected by overfitting. R_{free} and the "free" set of observations are not only used for validation purposes. Weights for restrained refinement are optimized by minimizing R_{free}, and the test set is used for estimating maximum likelihood parameters (15, 16, 27–29). The calculation of R_{free} should be based on at least 500 data points. For reliable parameter estimation, at least 2,000 data points are usually set aside. There are many types of crystallographic studies that cannot afford excluding the required data points from refinement because the entire dataset is too small. Such studies include low-resolution macromolecular studies, high-pressure studies, neutron studies, and some of the latest data from free electron lasers (21, 22, 30).

In this work we assessed an alternative to R_{free}, namely the method of $R_{complete}$. Its calculation was first suggested along with R_{free} (6). In contrast to R_{free}, $R_{complete}$ is calculated from the entire dataset with observations that, at some point, were previously used during refinement. Therefore, the $R_{complete}$ method relies on the reduction of the effect of overfitting from the structural model. Several methods have been suggested to reduce the effect of overfitting including simulated annealing (6), random parameter perturbation (13, 21, 23), and refinement until convergence. We show here that refinement until convergence is sufficient. We show that $R_{complete}$ has at least the same low statistical bias as R_{free}. Unlike R_{free}, the value of $R_{complete}$ does not vary even when only a single observation is left out from each refinement run. Therefore, the $R_{complete}$ method enables the estimation of maximum likelihood parameters for small datasets.

To carry out model building, the structural model used as input for the calculation of $R_{complete}$ should also be refined against all data (cf. ref. 14). The bias reduced electron density map is a byproduct of the calculation of $R_{complete}$. Analyzing fluctuations of specific atoms in the structural models P_i, that are produced by the $R_{complete}$ method, point at parts of the model that deserve special attention, such as weak electron density. See *SI Appendix*, *section 4* for an example.

ACKNOWLEDGMENTS. We thank our referees and the expert editor for their supportive and constructive criticism. We thank R. Pannu and P. Skubak for discussion about the expansion of the presented ideas to maximum likelihood-based refinement; P. Lafond for discussion about cross-validation, bootstrap, and jackknife methods; G. Murshudov for information about implementation details in Refmac5; G. M. Sheldrick for critical reading of the manuscript; and A. Paesch for crystal growth and data collection of cubic insulin. T.G. was partially supported by the Volkswagen Stiftung via the Niedersachsenprofessur awarded to Prof. G. M. Sheldrick.

1. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
2. Kennard O (1981) Cambridge Crystallographic Database. *Acta Crystallogr A* 37:C343.
3. Kleywegt GJ, Jones TA (1995) Where freedom is given, liberties are taken. *Structure* 3(6):535–540.
4. Kleywegt GJ, Brünger AT (1996) Checking your imagination: Applications of the free R value. *Structure* 4(8):897–904.
5. Brünger AT (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355(6359):472–475.
6. Brünger AT (1997) Free R value: Cross-validation in crystallography. *Methods Enzymol* 277:366–396.
7. Geisser S (1993) *Predictive Inference: An Introduction*. Monographs on Statistics and Applied Probability (Chapman & Hall, London), Vol 55.
8. Efron B, Tibshirani R (1994) *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability (Chapman & Hall, London), Vol 57.
9. Stone M (1974) Cross-validatory choice and assessment of stratistical predictions. *J R Stat Soc B* 36:111–147.
10. Schröder GF, Levitt M, Brunger AT (2010) Super-resolution biomolecular crystallography with low-resolution data. *Nature* 464(7292):1218–1222.
11. Brunger AT, et al. (2012) Improving the accuracy of macromolecular structure refinement at 7 Å resolution. *Structure* 20(6):957–966.
12. Pražnikar J, Afonine PV, Gunčar G, Adams PD, Turk D (2009) Averaged kick maps: less noise, more signal... and probably less bias... .. *Acta Crystallogr D Biol Crystallogr* 65(Pt 9):921–931.
13. Joosten RP, Long F, Murshudov GN, Perrakis A (2014) The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* 1(Pt 4):213–220.
14. Brünger AT (1993) Assessment of phase accuracy by cross validation: The free R value. Methods and applications. *Acta Crystallogr D Biol Crystallogr* 49(Pt 1):24–36.
15. Lunin V, Skovoroda T (1995) R-free likelihood-based estimates of errors for phases calculated from atomic models. *Acta Crystallogr A* 51(Pt 6):880–887.
16. Pannu NS, Murshudov GN, Dodson EJ, Read RJ (1998) Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr D Biol Crystallogr* 54(Pt 6 Pt 2):1285–1294.
17. Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat* 37:6–48.
18. Kohavi R (1995) *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection* (Morgan Kaufmann, Burlington, MA), pp 1137–1145.
19. Tickle IJ, Laskowski RA, Moss DS (2000) Rfree and the rfree ratio. II. Calculation Of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. *Acta Crystallogr D Biol Crystallogr* 56(Pt 4):442–450.
20. Gong G (1986) Cross–validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *J Am Stat Assoc* 81:108–113.
21. Gruene T, Hahn HW, Luebben AV, Meilleur F, Sheldrick GM (2014) Refinement of macromolecular structures against neutron data with SHELXL2013. *J Appl Cryst* 47(Pt 1):462–466.
22. Fabbiani FP, Buth G, Levendis DC, Cruz-Cabeza AJ (2014) Pharmaceutical hydrates under ambient conditions from high-pressure seeds: A case study of GABA monohydrate. *Chem Commun (Camb)* 50(15):1817–1819.
23. Pražnikar J, Turk D (2014) Free kick instead of cross-validation in maximum-likelihood refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 70(Pt 12):3124–3134.
24. Turk D (2011) MAIN 2011: Refining against all diffraction data – free of R-free. *Acta Crystallogr A* 67:C598.
25. Tickle I (2015) CCP4 Bulletin Board. Available at www.ccp4.ac.uk/ccp4bb.php. Accessed November 25, 2014.
26. Wilson A (1950) Largest likely value for the reliability index. *Acta Crystallogr* 3:397–398.
27. Tronrud DE (2004) Introduction to macromolecular refinement. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2156–2168.
28. Murshudov GN, et al. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67(Pt 4):355–367.
29. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.
30. Barends TR, et al. (2014) De novo protein crystal structure determination from X-ray free-electron laser data. *Nature* 505(7482):244–247.
31. Herbst-Irmer R (2014) Experimental charge density studies: Discard valid data and overfit? *Acta Crystallogr A* 70:C282.
32. Mondal KC, et al. (2014) One-electron-mediated rearrangements of 2,3-disiladicarbene. *J Am Chem Soc* 136(25):8919–8922.
33. Holstein J, Hubschle C, Dittrich B (2012) Electrostatic properties of nine fluoroquinolone antibiotics derived directly from their crystal structure refinements. *CrystEngComm* 14:2520–2531.
34. Gruene T, Sheldrick G, Zlatopolskiy B, Kozhushkov S, de Meijere A (2014) Structure of hormaomycin, a naturally occurring cyclic octadepsipeptide, in the crystal. *Z Naturforsch B* 69b:945–949.
35. Ishikawa T, et al. (2008) An abnormal pK(a) value of internal histidine of the insulin molecule revealed by neutron crystallographic analysis. *Biochem Biophys Res Commun* 376(1):32–35.