### Title
Universal parameters of bulk-solvent masks




### Journal
Acta Crystallographica Section A: Foundations and advances, 80(Pt 2)

### ISSN
0108-7673



### Authors
Urzhumtsev, Alexandre
Adams, Paul
Afonine, Pavel



### Publication Date
2024-03-01

### DOI
10.1107/s2053273324000299

Peer reviewed

Acta Cryst A FOUNDATIONS ADVANCES

ISSN 2053-2733

Received 23 September 2023
Accepted 8 January 2024

Edited by P. M. Dominiak, University of Warsaw, Poland

**Keywords:** bulk-solvent modelling; flat models; solvent masks; grid steps; mask parameters.

# Universal parameters of bulk-solvent masks

**Alexandre Urzhumtsev,[a,b] Paul Adams[c,d] and Pavel Afonine[c]***

[a]Centre for Integrative Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS–INSERM-UdS, 1 rue Laurent Fries, BP 10142, 67404 Illkirch, France, [b]Faculté des Sciences et Technologies, Université de Lorraine, BP 239, 54506 Vandoeuvre-les-Nancy, France, [c]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA, and [d]Department of Bioengineering, University of California, Berkeley, Berkeley, California, USA. *Correspondence e-mail: pafonine@lbl.gov

The bulk solvent is a major component of biomacromolecular crystals that contributes significantly to the observed diffraction intensities. Accurate modelling of the bulk solvent has been recognized as important for many crystallographic calculations. Owing to its simplicity and modelling power, the flat (mask-based) bulk-solvent model is used by most modern crystallographic software packages to account for disordered solvent. In this model, the bulk-solvent contribution is defined by a binary mask and a scale (scattering) function. The mask is calculated on a regular grid using the atomic model coordinates and their chemical types. The grid step and two radii, solvent and shrinkage, are the three parameters that govern the mask calculation. They are highly correlated and their choice is a compromise between the computer time needed to calculate the mask and the accuracy of the mask. It is demonstrated here that this choice can be optimized using a unique value of 0.6 Å for the grid step irrespective of the data resolution, and the radii values adjusted correspondingly. The improved values were tested on a large sample of Protein Data Bank entries derived from X-ray diffraction data and are now used in the computational crystallography toolbox (*CCTBX*) and in *Phenix* as the default choice.

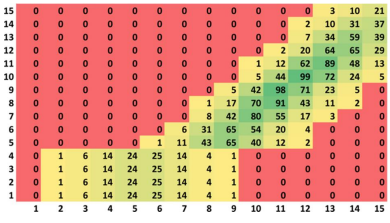
## 1. Introduction

Bulk solvent (or disordered solvent) on average occupies about half the volume of a macromolecular crystal and noticeably contributes to the medium- and low-resolution structure factor intensities [see *e.g.* Weichenberger *et al.* (2015) for a recent review]. It is therefore important to include its contribution into the model-calculated structure factors to account for the entire unit-cell contents adequately. The procedure needs to be fast and accurate because this calculation is repeated many times during atomic model refinement.

The flat (or mask-based) bulk-solvent model (Jiang & Brünger, 1994) is currently the option of choice in most crystallographic software packages. The model first requires the definition of a solvent mask in the unit cell. This mask is a binary function calculated on a regular grid with values of zero inside the molecular region and one outside. The Fourier coefficients $\mathbf{F}_{\rm mask}(\mathbf{s})$ of this binary mask are then calculated and scaled together with the structure factors $\mathbf{F}_{\rm calc}(\mathbf{s})$ calculated from the atomic model,

$$\mathbf{F}_{\rm model}(\mathbf{s}) = k_{\rm total}(\mathbf{s})\big[\mathbf{F}_{\rm calc}(\mathbf{s}) + k_{\rm mask}(\mathbf{s})\mathbf{F}_{\rm mask}(\mathbf{s})\big]. \qquad (1)$$

The resolution-dependent scales $k_{\rm mask}(\mathbf{s})$ and $k_{\rm total}(\mathbf{s})$ are obtained by fitting $\mathbf{F}_{\rm model}(\mathbf{s})$ to the experimental data [see *e.g.* Afonine *et al.* (2013)].



The mask calculation as introduced by Jiang & Brünger (1994) uses the following parameters:

(i) The size of the grid step $d_{\rm grid}$.

(ii) The solvent probe radius $r_{\rm probe}$ or $r_{\rm solv}$.

(iii) The shrinking radius $r_{\rm shrink}$.

(iv) Tabulated atomic van der Waals radii.

The mask calculation procedure involves augmenting the atomic van der Waals radius with the solvent probe radius to create a sphere of combined radius around each atom. Grid points falling outside of these spheres, which define the expanded macromolecular region, are designated as the solvent-accessible region surrounding the macromolecule. Subsequently, all grid points within a distance $r_{\rm shrink}$ from any point of the tentative solvent-accessible region defined above are assigned to the bulk-solvent region. The resulting mask is referred to as the bulk-solvent mask.

An optimal choice for these parameters should balance structure factor accuracy and the time required to compute the mask and its Fourier coefficients by Fourier transformation. Based on two cases at 2.2 and 1.8 Å resolution, Jiang & Brünger (1994) determined optimal choices for $r_{\rm probe}$, $r_{\rm shrink}$ and $d_{\rm step}$ to be 1.0, 1.1 and ($d_{\rm min}$/4) Å, respectively, where $d_{\rm min}$ is the resolution of the data set. Later, Rees *et al.* (2005) showed that for low-resolution data sets a step size of ($d_{\rm min}$/4) Å is too coarse, leading to inaccurate masks, and that a step size somewhere between $d_{\rm min}$/5 and $d_{\rm min}$/10 is more appropriate. While this solves the problem of mask accuracy at low resolution, at high resolution such a fine grid step will result in a significant (and unnecessary) increase in computational time.

In this work, we suggest that the grid step for mask calculation should not depend on the resolution. We demonstrate that using a step size of 0.6 Å, along with values of $r_{\rm solv}$ and $r_{\rm shrink}$ set to 1.1 and 0.9 Å, respectively, does not compromise the accuracy of the mask or the calculation time. Therefore, we recommend this combination of parameters for calculating the bulk-solvent mask for structures of any resolution.

## 2. Method

### 2.1. Why is a common resolution-independent grid step expected?

The electron-density distribution of a macromolecule in a crystal is a peaky function, while the function that describes the bulk solvent is a flat function with a smooth border [see *e.g.* Fenn *et al.* (2010)]. Consequently, the Fourier coefficients that describe the bulk-solvent distribution decrease sharply, and usually become negligibly small, at resolutions better than $d_{\rm solv} \simeq$ 3.5–4.0 Å [see *e.g.* Phillips (1980), Jiang & Brünger (1994) or Afonine *et al.* (2013)]. To understand the consequences of this on the choice of the grid step, we refer to a one-dimensional example below.

For a periodic function of a single variable with period $a$, the integral Fourier transform gives an infinite number of Fourier coefficients $\mathbf{F}(h)$, where $h$ is an integer, both positive and negative. When such a function is sampled on a regular grid with $N$ points per interval, the discrete Fourier transform yields only $N$ independent values $\mathbf{F}_{\rm grid}(h)$, *e.g.*

$$\mathbf{F}_{\rm grid}(h) = \mathbf{F}(h) + \sum_{m=1}^{\infty} [\mathbf{F}(h + mN) + \mathbf{F}(h - mN)], \qquad (2)$$

for $-N/2 < h \leq N/2$ [see, for example, formula (4) in Ten Eyck (1977)]. $\mathbf{F}_{\rm grid}(h)$ differ from the respective $\mathbf{F}(h)$ by a convergent infinite series of correcting terms (Appendix *A*) where $\lim_{|h|\to\infty} \mathbf{F}(h) = 0$. Let us suppose that $\mathbf{F}(h)$ values are equal exactly to zero for $|h| > H = a/d_{\rm solv}$ with some resolution limit $d_{\rm solv}$. If $N$ in (2) is sufficiently large, for example, $N > 2H$, all correcting terms with indices $h \pm mN$ are also zero, resulting in $\mathbf{F}_{\rm grid}(h) = \mathbf{F}(h)$ as desired. Taking $N$ larger than $2H$, *i.e.* taking the grid step $d_{\rm grid} = a/N$ smaller than

$$\frac{a}{N} < \frac{a}{2H} = \frac{1}{2}\, d_{\rm solv}, \qquad (3)$$

has no effect. Conversely, taking smaller $N$ makes $\mathbf{F}_{\rm grid}(h)$ different from $\mathbf{F}(h)$ by at least one non-zero term $\mathbf{F}(h \pm mN)$, $m \neq 0$. The analogue of (2) for three-dimensional functions can be found in Sayre (1951), Lunin *et al.* (2002), Navaza (2002) and Afonine & Urzhumtsev (2004).

This suggests the potential existence of a universally optimal grid step $d^0_{\rm grid}$ for the problem under study, which is related to $d_{\rm solv} \simeq 3.5$ Å in a manner similar to (3), albeit with a scale factor that may not be equal to $\frac{1}{2}$; the latter arises from the facts that these high-resolution structure factors may be different from exactly zero and the Fourier analysis is carried out in three-dimensional space.

### 2.2. Models and data

The search for the optimal bulk-solvent mask parameters was conducted using 277 quality-filtered models and X-ray diffraction data obtained from the Protein Data Bank (PDB; Burley *et al.*, 2021). The quality filters included a crystallographic $R$ factor better than 0.25, overall and per-resolution shell data completeness above 95%, no data pathologies such as twinning, and a relatively high upper data resolution limit ($d_{\rm min} \leq$ 3.0 Å). To accelerate the calculations, we excluded very large models. The results obtained with these 277 models were then validated with a larger set of 2077 structures used in a previous bulk-solvent study (Afonine *et al.*, 2013) and representing a broad range of model sizes and data resolutions, from subatomic to low.

Among the 277 models, 54% were refined using *Refmac* (Murshudov *et al.*, 2011), 25% using *Phenix* (Afonine *et al.*, 2012), 14% using *CNS/X-plor* (Brünger *et al.*, 1998), 0.04% using *BUSTER*/*TNT* (Tronrud *et al.*, 1987; Roversi *et al.*, 2000; Blanc *et al.*, 2004) and 0.03% using *SHELX* (Sheldrick, 2015). These programs may potentially employ different types of bulk-solvent models, for example the Babinet-based model (Langridge *et al.*, 1960; Moews & Kretsinger, 1975) in *SHELX*, *Refmac* and *BUSTER*/*TNT*, as well as different parameters for mask calculations when the mask-based solvent model was used, such as in *Phenix*, *CNS/X-plor* and again *Refmac*. We believe that this diversity in refinement programs, each with its

distinct formulation of bulk-solvent modelling, helps mitigate any potential model bias related to the solvent parameters used in these programs.

### 2.3. Finding optimal values for $r_{solv}$, $r_{shrink}$ and $d_{grid}$

For each selected PDB entry, the values of $d_{grid}$ were systematically sampled in the range between 0.2 and 1 Å in steps of 0.1 Å, and both $r_{solv}$ and $r_{shrink}$ were sampled between 0.5 and 1.5 Å, also in steps of 0.1 Å. For each trial triplet of values ($r_{solv}$, $r_{shrink}$, $d_{grid}$), the scales $k_{total}(\mathbf{s})$ and $k_{mask}(\mathbf{s})$ in (1) were re-calculated as detailed by Afonine *et al.* (2013), and the *R*-factor values, referred to as $R^{(4)}$, were calculated using all reflections up to 4 Å resolution. In what follows, $R_n^{(4)}$ stands for $R^{(4)}$ calculated for the structure numbered $n$. These values were the principal information to identify potential universal values for $r^0_{solv}$, $r^0_{shrink}$ and $d^0_{grid}$. The details follow in Section 3.

## 3. Results

### 3.1. Variation of the optimal *R* factor with the mask grid step

The search for common parameters was based on the hypothesis that there is a common behaviour of $R_n^{(4)}$ with parameter values for different structures. First, we tried to decouple the search for optimal $d_{grid}$ and ($r_{solv}$, $r_{shrink}$). This was achieved by finding the combination of ($r_{solv}$, $r_{shrink}$) that minimizes $R_n^{(4)}$ for each trial $d_{grid}$ and for each structure $n$,

$$\bar{R}_n^{(4)}(d_{grid}) = \min_{r_{solv}, r_{shrink}} \{R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid})\}. \quad (4)$$

Obviously, these values are different for each structure, but they vary with $d_{grid}$ very similarly. In particular, this is true for their variation around the average of $\bar{R}_n^{(4)}(d_{grid})$ over grid steps

$$\Delta\bar{R}_n^{(4)}(d_{grid}) = \bar{R}_n^{(4)}(d_{grid}) - \left\langle \bar{R}_n^{(4)}(d_{grid}) \right\rangle_{grid}. \quad (5)$$

Subtracting the average in (5) makes the dependency of $\Delta\bar{R}_n^{(4)}(d_{grid})$ on $d_{grid}$ similar for all structures, which in turn makes it possible to analyse the average of these dependencies over all structures (Fig. 1).

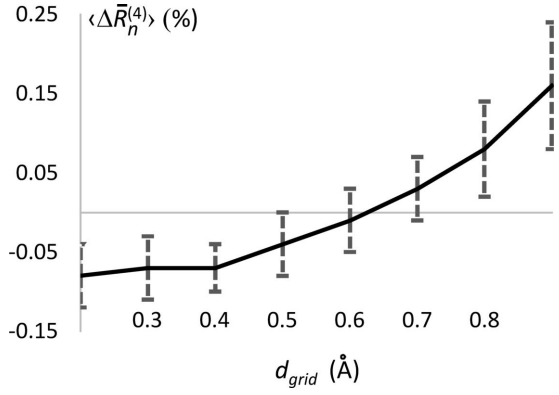



**Figure 1**
The variation $\Delta\bar{R}_n^{(4)}$ in the $\bar{R}_n^{(4)}$ factor, as defined in the text, as a function of the grid step $d_{grid}$. Each data point is the average of $\Delta\bar{R}_n^{(4)}$ across all structures. Intervals of 1$\sigma$ are given for each grid step.

It is to be expected that $\Delta\bar{R}_n^{(4)}(d_{grid})$ should increase with step size. However, the value does not change significantly for $d_{grid}$ in the range of 0.2–0.4 Å, suggesting that steps smaller than 0.4 Å are unnecessarily small. Above this step value, $\Delta\bar{R}_n^{(4)}(d_{grid})$ starts to increase, and the goal is to find a compromise between the introduced errors and the gain in computation time. Increasing the step from 0.4 Å to 0.6 Å or 0.8 Å increases the grid size, and therefore the number of computing operations, by about four times or eight times, respectively.

A step size of 1.0 Å resulted in very large errors and was excluded from further analysis. Calculations with a step size of 0.9 Å resulted in a large number of outliers with large $\Delta\bar{R}_n^{(4)}$, making this step size also unsuitable. This leads to 0.4–0.8 Å as the range for the grid-step search.

### 3.2. Acceptable combinations of parameters

Next, for each model $n$, we analysed the parameter values that lead to the lowest $R_n^{(4)}$ value across all combinations of ($r_{solv}$, $r_{shrink}$, $d_{grid}$),

$$\bar{R}_n^{(4)} = \min_{r_{solv}, r_{shrink}, d_{grid}} \{R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid})\} = \min_{d_{grid}} \{\bar{R}_n^{(4)}(d_{grid})\}. \quad (6)$$

It is possible that, for a given structure, several combinations of ($r_{solv}$, $r_{shrink}$, $d_{grid}$) result in $R_n^{(4)}$ values that are close to the global minimum of (6). To address such small fluctuations in the $R_n^{(4)}$ values, we introduce a parameter $\varepsilon_R$ considering all values $R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid}) \leq \bar{R}_n^{(4)} + \varepsilon_R$ to be as good as $\bar{R}_n^{(4)}$, where the value of $\varepsilon_R$ varies in the range 0.001–0.002.

The parameter values ($r_{solv}$, $r_{shrink}$, $d_{grid}$) corresponding to (6) are expected to vary from one structure to another, and we are looking for the combinations that are persistent over all structures. As a formal quantitative measure, for each set of parameters ($r_{solv}$, $r_{shrink}$, $d_{grid}$) and for each structure $n$, we calculate a non-negative value

$$\Delta R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid}; \varepsilon_R) = \max\{R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid}) - \bar{R}_n^{(4)} - \varepsilon_R; 0\}. \quad (7)$$

To be able to combine the distribution of ($r_{solv}$, $r_{shrink}$, $d_{grid}$) for each structure into one cumulative distribution over all structures, we convert $\Delta R_n^{(4)}$ in (7) into

$$P_n(r_{solv}, r_{shrink}, d_{grid}) = \exp[-C\,\Delta R_n^{(4)}(r_{solv}, r_{shrink}, d_{grid}; \varepsilon_R)] \quad (8)$$

with constants $C > 0$ and $\varepsilon_R$. The product of (8) over all structures,

$$P(r_{solv}, r_{shrink}, d_{grid}) = \prod_n P_n(r_{solv}, r_{shrink}, d_{grid}), \quad (9)$$

reflects both the contrast of the lowest $R_n^{(4)}$ for an individual structure and the persistence of the parameter values over all structures. $P(r_{solv}, r_{shrink}, d_{grid})$ varies from 0 to 1; the higher the value of (9), the more preferable the combination of

parameters. Calculations with data sets of different sizes suggested the choice of $C$ in the range between 0.01 and 1.0 and $\varepsilon_R$ as stated above in order to obtain a decent contrast while keeping points with neighbouring $R_n^{(4)}$ values. In general, we observed that the variation in the constants $C$ and $\varepsilon_R$ around the values given above obviously modifies the contrast of the distribution (9) while not influencing the location of its peaks.

Fig. 2 shows the results with $\varepsilon_R = 0.002$ and $C = 0.5$. As expected, the distribution shows that mask shrinking does not impact the results when $r_{\rm shrink} < d_{\rm grid}$ (rectangular bottom-left regions). Also, the distribution shows a clear nearly linear correlation between $r_{\rm solv}$ and $r_{\rm shrink}$, giving preferable values roughly at the line $r_{\rm solv} - r_{\rm shrink} \simeq 0.2$ Å for each grid step. Finally, it shows that the optimal radii ($r_{\rm solv}$, $r_{\rm shrink}$) cluster in the range (1.1 ± 0.1 Å, 0.9 ± 0.1 Å).

As expected, increasing the grid step makes $R_n^{(4)}$ worse. In agreement with the first test (Fig. 1), most frequently the lowest $R_n^{(4)}$ occurred for a grid step size of 0.3–0.4 Å ($P$ values up to 0.99). Consequently, a step of 0.4 Å may be considered as a candidate for the most accurate calculations since a smaller step of 0.3 Å does not significantly improve the $R$ factors while leading to an increased computational time. Using a grid with step sizes of 0.5–0.6 Å makes it possible for $R_n^{(4)}$ to be close to $\bar{R}_n^{(4)}$, indicated by high values of the function $P(r_{\rm solv}, r_{\rm shrink}, d_{\rm grid})$ in the range 0.91–0.93. Increasing the step further reduces the maximum $P$-function value to 0.83. Since a larger grid step is preferable to reduce the computing time,

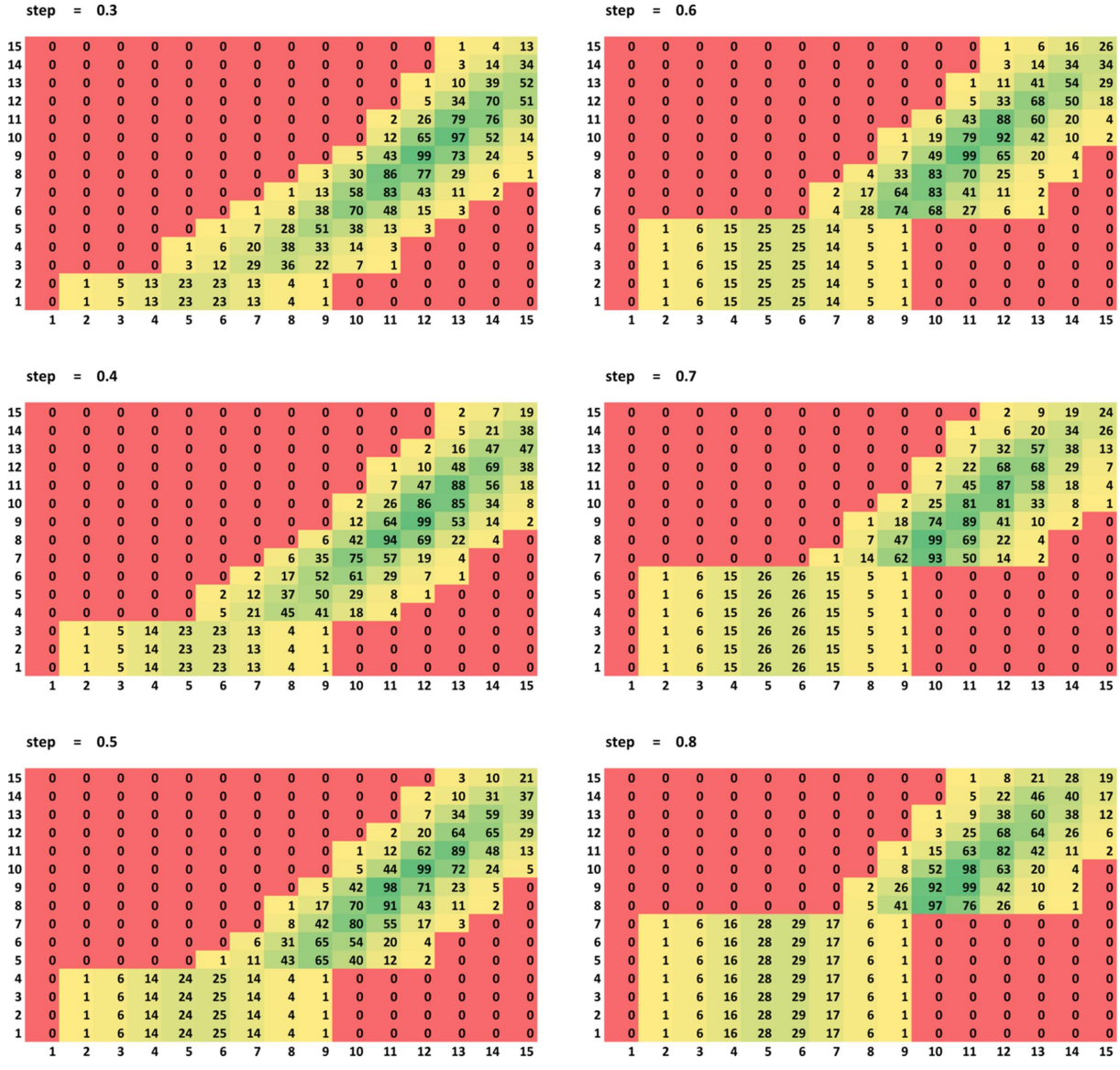



**Figure 2**
The distribution $P(r_{\rm solv}, r_{\rm shrink}, d_{\rm grid}) \times 10^2$ of the variation in $R_n^{(4)}$ with respect to the best value $\bar{R}_n^{(4)}$ for all combinations of the mask parameters. The values on the axes are $r_{\rm solv} \times 10$ (horizontal) and $r_{\rm shrink} \times 10$ (vertical). High $P$ values correspond to the ($r_{\rm solv}$, $r_{\rm shrink}$, $d_{\rm grid}$) combinations giving $R_n^{(4)}$ close to the minimum best value of $\bar{R}_n^{(4)}$. The colour scheme indicates the $P$-function value ranges: red ($P < 0.01$), yellow ($0.01 \le P < 0.2$), light green ($0.2 \le P < 0.9$) and dark green ($P \ge 0.9$).

$d_{\rm grid}$ = 0.6 Å is a good potential compromise candidate for the universal value.

The values of the radii ($r_{\rm solv}$, $r_{\rm shrink}$) leading to the minimum $\bar{R}_n^{(4)}$ value in (6) also varied only slightly over the trial grid step sizes. This provided a relatively small number of tentative combinations ($r_{\rm solv}$, $r_{\rm shrink}$, $d_{\rm grid}$) to identify the optimal ones which we denote ($r^0_{\rm solv}$, $r^0_{\rm shrink}$, $d^0_{\rm grid}$).

### 3.3. Optimal set of parameters

The analysis described in Sections 3.1–3.2 results in a range of ($r_{\rm solv}$, $r_{\rm shrink}$, $d_{\rm grid}$) parameters minimizing $R^{(4)}$ on average for all test models. These values, however, do not necessarily lead to the lowest $R^{(4)}$ for a particular model.

Next, we can ask which of these combinations, if any, lead to a value of $R_n^{(4)}(r_{\rm solv}, r_{\rm shrink}, d_{\rm grid})$ that is larger than, and by how much, the lowest value of $\bar{R}_n^{(4)}$ in (6) for what fraction of structures. To answer this question, we calculate the fraction $p(\Delta R)$ of the structures with the difference

$$R_n^{(4)}\left(r_{\rm solv}, r_{\rm shrink}, d_{\rm grid}\right) - \bar{R}_n^{(4)} > \Delta R \qquad (10)$$

for different $\Delta R$ values. The expected solution corresponds to the minimum of the $p(\Delta R)$ function. For the sake of

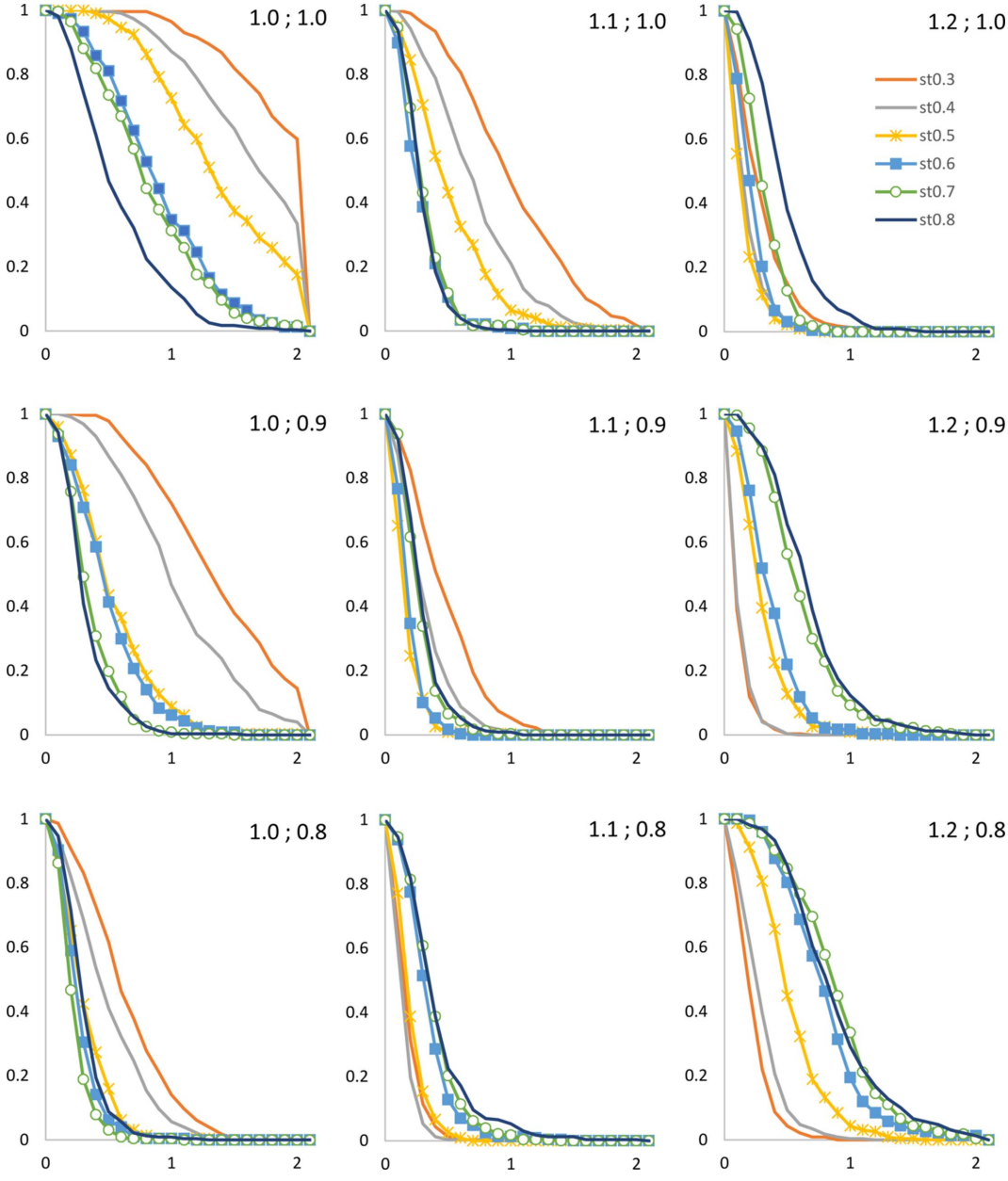



**Figure 3**
The fraction of the models that satisfy $R_n^{(4)}(r_{\rm solv}, r_{\rm shrink}, d_{\rm grid}) - \bar{R}_n^{(4)} > \Delta R$ shown as a function of $\Delta R$. The colour scheme for the grid step is the same for all plots and is given in the top right plot. The values of ($r_{\rm solv}$; $r_{\rm shrink}$) are indicated in each plot.

completeness, we calculated $p(\Delta R)$ for all triplet values of the parameters considered above.

The combination ($r_{solv}$ = 1.0 Å, $r_{shrink}$ = 1.0 Å) gives poor results for all grid step sizes and the combination ($r_{solv}$ = 1.2 Å, $r_{shrink}$ = 0.8 Å) gives results acceptable only for very small grid steps, $d_{grid}$ = 0.3–0.4 Å (Fig. 3). The best results are observed for the sets of $r_{solv}$, $r_{shrink}$ with $r_{solv} - r_{shrink}$ = 0.2 Å, with a slight preference for the combination ($r_{solv}$ = 1.1 Å, $r_{shrink}$ = 0.9 Å, $d_{grid}$ = 0.6 Å). Here, $R_n^{(4)}$ increased by less than 0.5% for all structures of the search set except one, for which this value was below 0.6%. The same plots (Fig. 3) indicate that if more accurate calculations are required, then the combination ($r_{solv}$ = 1.1 Å, $r_{shrink}$ = 0.8 Å, $d_{grid}$ = 0.4 Å) is optimal. However, using this finer grid step would lead to a nearly fourfold increase in the number of grid points and, consequently, in the number of computational operations required. Conversely, for a very large structure, if a coarser grid is acceptable, a possible combination would be ($r_{solv}$ = 1.0 Å, $r_{shrink}$ = 0.8 Å, $d_{grid}$ = 0.7 Å), resulting in only a slight increase in overall $R$ factors.

### 3.4. New versus old mask calculation parameters

Finally, for each model from the complete data set, we analysed how much the $R^{(4)}$ and the $R$ factor calculated using all reflection data change if the new mask calculation parameter values ($r^0_{solv}$ = 1.1 Å, $r^0_{shrink}$ = 0.9 Å, $d^0_{grid}$ = 0.6 Å) are used instead of the values of ($r_{solv}$ = 1.1 Å, $r_{shrink}$ = 1.0 Å, $d_{grid}$ = $d_{min}$/4) used previously.

Fig. 4 shows that $R_n^{(4)}$ varies little and typically remains within ±0.3% for most structures, with the exception of a few cases where it varies within ±0.5%. We consider these variations negligible. As expected, $R_n$ changes even less than $R_{n_o}^{(4)}$, since the bulk-solvent contribution vanishes beyond 3–4 Å resolution.

The only notable outlier is PDB entry 3b6a (Willems *et al.*, 2008), for which $\Delta R_n^{(4)}$ = 0.92% ($\Delta R_n$ = 0.64%). This structure was solved at a resolution of $d_{min}$ = 3.0 Å, which means that the original algorithm used a mask with a grid step size of 0.75 Å, coarser than the proposed 0.6 Å. This seemingly counterintuitive result can be rationalized as follows. The bulk-solvent mask typically consists of a large region and several (often many) smaller isolated regions (Afonine *et al.*, 2024). These small regions are typically cavities inside the protein or computational artefacts. The number and size of such regions vary based upon the choice of mask parameters ($r_{solv}$, $r_{shrink}$, $d_{grid}$). With a step size $d^0_{grid}$ = 0.6 Å, the mask for 3b6a contains about 20 isolated small regions incapable of containing even a single disordered water molecule. Excluding these regions from the bulk-solvent mask reduces $\Delta R_n^{(4)}$ and $\Delta R_n$ to 0.36% and 0.26%, respectively, suggesting that these regions are computational artefacts.

## 4. Conclusions

The choice of mask parameters for the flat bulk-solvent model, *i.e.* solvent and mask shrinkage radii and the grid sampling step size, affects both the accuracy of the fit between model and data at medium to low resolution and the speed of the calculations. Since accounting for the bulk solvent typically occurs in crystallographic calculations that involve atomic model and reflection data, from simple operations like $R$-factor or map calculation to complex procedures like model refinement and building, the computational efficiency of this step is critical. The parameters governing the speed and accuracy of the flat bulk-solvent model are the solvent radius $r_{solv}$, the mask shrinkage radius $r_{shrink}$ and the grid step $d_{grid}$ for the mask sampling. When this model was introduced (Jiang & Brünger, 1994) the choice of values for these parameters, of $r_{solv}$ = 1.0 Å, $r_{shrink}$ = 1.1 Å and $d_{grid}$ = ($d_{min}$/4) Å, was based on only two study cases at medium resolution (around 2 Å). A decade later, this choice was revisited for low-resolution cases by Rees *et al.* (2005), resulting in the suggestion that much

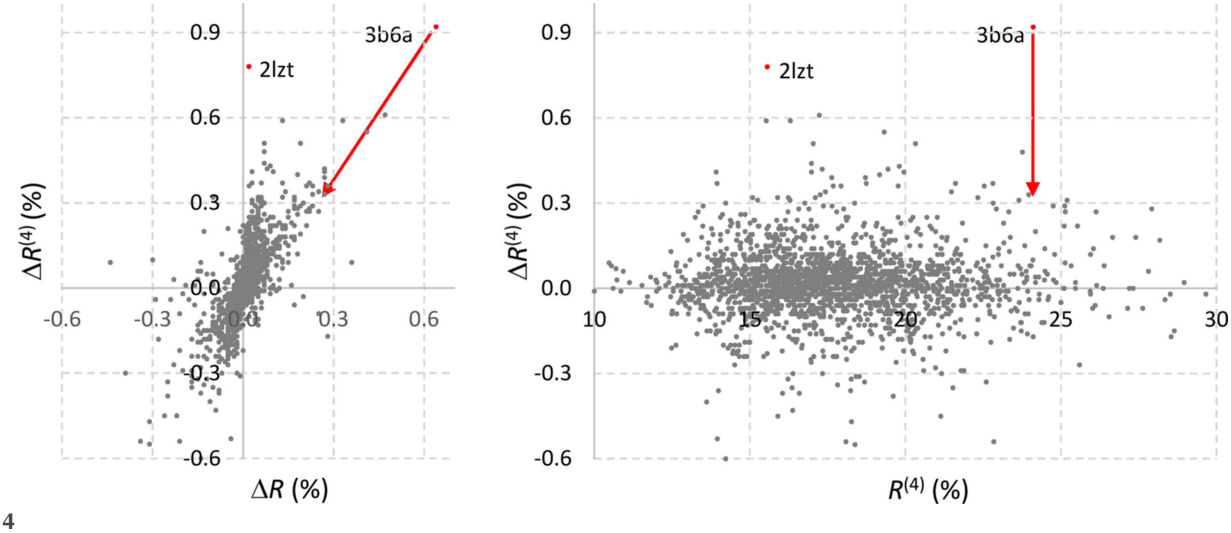



**Figure 4**
The variation in $\Delta R^{(4)}$ calculated for the test set of 2077 models when moving from the conventional set of mask parameters to the selected set $r^0_{solv}$, $r^0_{shrink}$, $d^0_{grid}$ of optimal values. Each point corresponds to an individual model. The plots show the distribution of the variation in $\Delta R^{(4)}$ (vertical) versus (left) the variation $\Delta R$ in the overall $R$ factor calculated for the whole set of reflections and (right) the $R^{(4)}$ value itself. For the model of 3b6a (indicated by red arrows) the values (0.64; 0.92) become (0.26; 0.36) after the removal of isolated small-volume regions appearing in the new mask.

finer steps are needed at these resolutions. While these finer steps, calculated as a fraction of $d_{\rm min}$, address the problem of accuracy for low-resolution models, they in turn create the problem of substantially increasing the computational time for higher-resolution cases. A universal resolution-independent choice of mask calculation parameters is therefore highly desirable, and we here show that this is possible. A systematic study across hundreds of models from the PDB performed in this work reveals an optimal choice of these parameters to be $r^0_{\rm solv}$ = 1.1 Å, $r^0_{\rm shrink}$ = 0.9 Å and $d^0_{\rm grid}$ = 0.6 Å. Validation of this choice with a much larger test set of models shows that these values are broadly applicable. In the last stages of refinement, a finer grid with a step $d_{\rm grid}$ = 0.4 Å and radii $r_{\rm solv}$ = 1.1 Å, $r_{\rm shrink}$ = 0.8 Å may possibly be used to improve the results further. The parameters described here are implemented in *CCTBX* and are used in the *Phenix* suite (Liebschner *et al.*, 2019), where applicable, starting from Version 1.20rc4-4425.

## APPENDIX *A*

### Discrete Fourier transform and Fourier coefficients

For a periodic function $f(x)$ of a single variable with period $a$ = 1, the integral Fourier transform results in an infinite number of Fourier coefficients $\mathbf{F}(h)$, where $h$ is an integer. The infinite Fourier series defined by these coefficients converges to the function $f(x)$. The sharper the function, the faster the convergence is (we do not specify the formal, mathematically strict, conditions of convergence when studying smooth functions like density distributions). When such a function is sampled on a regular grid with $N$ points per interval, using the obvious equation for integer $m$

$$\exp\left(i2\pi n\frac{h+mN}{N}\right)=\exp\left(i2\pi n\frac{h}{N}\right) \tag{11}$$

gives the convergent Fourier series on the grid nodes $x_n = n/N$ which can be expressed as

$$\begin{aligned} f(x_n) &= \sum_{h=-\infty}^{\infty}\mathbf{F}(h)\exp(i2\pi h x_n)\\ &= \sum_{h=-\infty}^{\infty}\mathbf{F}(h)\exp\left(i2\pi h\frac{n}{N}\right)\\ &= \sum_{h=-\infty}^{\infty}\mathbf{F}(h)\exp\left(i2\pi n\frac{h}{N}\right)\\ &= \sum_{h=H}^{H+N-1}\left\{\sum_{m=-\infty}^{\infty}[\mathbf{F}(h+mN)]\right\}\exp\left(i2\pi n\frac{h}{N}\right). \end{aligned} \tag{12}$$

Here $H$ is any integer number, and convergence of the original Fourier series proves convergence of each internal series in the right-hand expression of (12). In other words, given $N$ real numbers on a regular grid, the discrete Fourier transform results in a set of values

$$\mathbf{F}_{\rm grid}(h)=\sum_{m=-\infty}^{\infty}[\mathbf{F}(h+mN)], \tag{13}$$

which possess Hermitian symmetry. There are only $N/2$ independent values since, according to their definition in (13), $\mathbf{F}_{\rm grid}(h)$ are periodic and so $\mathbf{F}_{\rm grid}(h) = \mathbf{F}_{\rm grid}(h + N)$ for every $h$. Usually, $\mathbf{F}_{\rm grid}(h)$ taken with the consecutive indices are used as approximations to $\mathbf{F}(h)$. In some applications, the range $0 \le h < N$ can be chosen. However, since the Fourier coefficients $\mathbf{F}(h)$ for convergent series generally decrease with $|h|$, it is more practical to choose $-N/2 < h \le N/2$, which results in a smaller $|\mathbf{F}_{\rm grid}(h) - \mathbf{F}(h)|$ difference.


## Funding information

Pavel V. Afonine and Paul D. Adams thank the NIH (grant Nos. R01GM071939, P01GM063210 and R24GM141254) and the *Phenix* Industrial Consortium for support of the *Phenix* project. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231. Alexandre Urzhumtsev thanks the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01 and Instruct-ERIC.


## References


Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst.* D**69**, 625–634.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

Afonine, P. V., Sobolev, O. V., Adams, P. D. & Urzhumtsev, A. (2024). *Protein Science*. In the press.

Afonine, P. V. & Urzhumtsev, A. (2004). *Acta Cryst.* A**60**, 19–32.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 2210–2221.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y., Voigt, M., Westbrook, J., Young, J. Y., Zardecki, C. & Zhuravleva, M. (2021). *Nucleic Acids Res.* **49**, D437–D451.

Fenn, T. D., Schnieders, M. J. & Brunger, A. T. (2010). *Acta Cryst.* D**66**, 1024–1031.

Jiang, J. S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *J. Mol. Biol.* **2**, 38–I, N12.

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.

Lunin, V. Y., Urzhumtsev, A. & Bockmayr, A. (2002). *Acta Cryst.* A**58**, 283–291.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–225.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Navaza, J. (2002). *Acta Cryst.* A**58**, 568–573.
Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.
Rees, B., Jenner, L. & Yusupov, M. (2005). *Acta Cryst.* D**61**, 1299–1301.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.
Sayre, D. (1951). *Acta Cryst.* **4**, 362–367.
Sheldrick, G. M. (2015). *Acta Cryst.* C**71**, 3–8.
Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. (2015). *Acta Cryst.* D**71**, 1023–1038.
Willems, A. R., Tahlan, K., Taguchi, T., Zhang, K., Lee, Z. Z., Ichinose, K., Junop, M. S. & Nodwell, J. R. (2008). *J. Mol. Biol.* **376**, 1377–1387.