Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Ten years of probabilistic estimates of biocrystal solvent content: new insights *via* nonparametric kernel density estimate

Christian X. Weichenberger[a] and Bernhard Rupp[b,c]*

[a]Center for Biomedicine, European Academy of Bozen/Bolzano (EURAC), Viale Druso 1, I-39100 Bozen/Bolzano, Italy, [b]Department of Forensic Crystallography, k.-k. Hofkristallamt, 991 Audrey Place, Vista, CA 92084, USA, and [c]Department of Genetic Epidemiology, Innsbruck Medical University, Schöpfstrasse 41, A-6020 Innsbruck, Austria

Correspondence e-mail: br@hofkristallamt.org

Received 6 January 2014
Accepted 11 March 2014

The probabilistic estimate of the solvent content (Matthews probability) was first introduced in 2003. Given that the Matthews probability is based on prior information, revisiting the empirical foundation of this widely used solvent-content estimate is appropriate. The parameter set for the original Matthews probability distribution function employed in *MATTPROB* has been updated after ten years of rapid PDB growth. A new nonparametric kernel density estimator has been implemented to calculate the Matthews probabilities directly from empirical solvent-content data, thus avoiding the need to revise the multiple parameters of the original binned empirical fit function. The influence and dependency of other possible parameters determining the solvent content of protein crystals have been examined. Detailed analysis showed that resolution is the primary and dominating model parameter correlated with solvent content. Modifications of protein specific density for low molecular weight have no practical effect, and there is no correlation with oligomerization state. A weak, and in practice irrelevant, dependency on symmetry and molecular weight is present, but cannot be satisfactorily explained by simple linear or categorical models. The Bayesian argument that the observed resolution represents only a lower limit for the true diffraction potential of the crystal is maintained. The new kernel density estimator is implemented as the primary option in the *MATTPROB* web application at http://www.ruppweb.org/mattprob/.

1. Introduction

The first step in the process of structure determination is almost always the estimation of the molecular unit-cell content. Such an analysis can be performed as soon as the unit-cell dimensions and possible lattice types have been determined from indexed diffraction data, and solvent-content analysis often informs the choice of internal symmetry and point-group symmetry. Not only can the numbers of possible molecular entities in the asymmetric unit cell be estimated, but improbable values for these numbers can indicate problems such as the presence of twinning, pseudo-symmetry or incorrect point (space) group assignment (Zwart *et al.*, 2008), and even the possibility that a different species than intended has been crystallized. The accurate estimate of the solvent content is also an important parameter in density-modification techniques which are used to break phase-angle ambiguity (Wang, 1985) in single-wavelength anomalous diffraction phasing (Dauter *et al.*, 2002; Mueller-Dieckmann *et al.*, 2007) for phase improvement (Abrahams & Leslie, 1996) and for phase extension (Sheldrick, 2010).

Table 1
Historic and current descriptive statistics of V_M and V_S distributions for protein crystals.

All V_S calculations are based on the same partial specific volume of 0.741 g cm^{-3}. Values extracted from original figures were used to compute the historic V_M and V_S values (Matthews, 1968, 1976), which are printed in italics. The listed mean V_M is followed by the standard deviation of the V_M distribution, while the range of the 99% confidence interval (CI) of the mean V_M is given in square brackets. With increasing availability of experimental data the precision of the mean also increases (smaller CI), but the actual distributions became wider (increasing standard deviation). N/A, not available.

Year	N	Mode V_M (Å^3 Da^{-1})	Mean V_M (Å^3 Da^{-1})	Median V_M (Å^3 Da^{-1})	Mean V_S (%)	Reference
1968	120	*2.1*	*2.325 ± 0.38 [±0.091]*	*2.3*	*45.8*	Matthews (1968)
1976	224	*2.1*	*2.432 ± 0.50 [±0.086]*	*2.3*	*47.7*	Matthews (1976)
2003	10471	2.34	2.691 ± 0.74 [±0.015]	2.52	51.9	Kantardjieff & Rupp (2003)
2008	9081	N/A	2.68 ± 0.78 [N/A]	2.48	N/A	Chruszcz *et al.* (2008)
2013	60218	2.32	2.665 ± 0.71 [±0.007]	2.49	51.4	This work

1.1. The Matthews coefficient

Based on the analysis of 116 different crystal forms of globular proteins, Matthews observed in 1968 that the fraction of the crystal volume occupied by solvent ranged from 27 to 78% (Matthews, 1968, 1976), with the most common value being about 43% (Fig. 1, Table 1). Matthews defined V_M, known as the Matthews coefficient, as the crystal (asymmetric unit) volume V_A per unit of protein molecular weight, M,

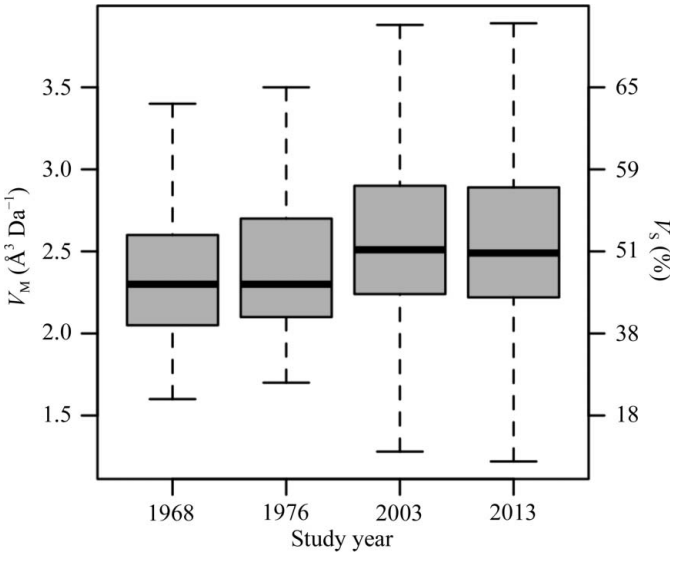

Figure 1
Box plot of historic and current values of V_M for protein crystals. On the x axis we plot the year of the study (1968, Matthews, 1968; 1976, Matthews, 1976; 2003, Kantardjieff & Rupp, 2003) and the current study (2013). The gray boxes display values that fall in between the first and third quartile, the black bar represents the median and the whiskers extend to data points no more than 1.5 times the inner quartile range. A plausible explanation for the small but significant increase in mean solvent content in the post-1970s analyses may be the now widespread availability of PCR-based molecular-biology techniques which enable heterologous overexpression of crystallizable variants of more intricate and rare proteins compared with the earlier, more stable and abundant proteins which almost exclusively had to be isolated from natural sources. Some of the few structures exceeding the closed-sphere packing limit of 26% solvent content have been analysed (Trillo-Muyo *et al.*, 2013). The structures of dehydrated monoclinic lysozyme (Nagendra *et al.*, 1998) may serve as examples of extremely compact structures with a solvent content as low as 9%.

$$V_M = \frac{V_A}{M} \tag{1}$$

and showed that V_M bears a straightforward relationship to the fractional volume of solvent in the crystal. Matthews further remarked in 1968 that a relationship between solvent content and resolution of the diffraction data is plausible and could exist.

Definition of solvent fraction. A protein crystal contains protein and solvent, so the asymmetric unit with volume V_A consists of the volume occupied by the protein, V_P, and by the solvent, V_L, such that $V_A = V_P + V_L$, or equivalently

$$1 = \frac{V_P}{V_A} + \frac{V_L}{V_A}. \tag{2}$$

Matthews assigned the fractions of the crystal volume occupied by the protein and the solvent as the dimensionless quantities $V_{prot} = V_P/V_A$ and $V_{solv} = V_L/V_A$, respectively, such that (2) becomes $1 = V_{prot} + V_{solv}$, or

$$V_{solv} = 1 - V_{prot},$$

or in subsequent notation,

$$V_S = 1 - V_{prot}. \tag{3}$$

Derivation. Given a value for the protein specific density ρ (or its reciprocal, the partial specific volume $\overline{v}$) of the protein, the volume occupied by a protein molecule, V_P, can be calculated from the molecular weight M_P as $V_P = M_P\overline{v}$. The weight of the protein molecule M_P in grams can be readily obtained from M using Avogadro's number N_A (6.022×10^{23} mol^{-1}), $M_P = M/N_A$.

With the widely accepted experimental value of ρ = 1.350 g cm^{-3} or the corresponding $\overline{v}$ = 0.741 cm^3 g^{-1} and the conversion 1 cm^3 = 10^{24} Å^3, we obtain V_P in Å^3:

$$V_P = M_P\overline{v} = \frac{M_P}{\rho} = \frac{M\overline{v}}{N_A} = \frac{M \times 0.741 \times 10^{24}}{6.022 \times 10^{23}}$$
$$= M\frac{(\mathrm{g})}{(\mathrm{mol})} \times 1.230 \times \frac{(\mathrm{Å}^3)(\mathrm{mol})}{(\mathrm{g})}. \tag{4}$$

We finally obtain the actual dimensionless solvent fraction V_S as per (3) expressed in terms of V_M (Å^3 Da^{-1} or Å^3 mol g^{-1}) from the definition of $V_{prot} = V_P/V_A$ as

$$V_S = 1 - V_{prot} = 1 - \frac{V_P}{V_A} = 1 - \frac{M\overline{v}}{N_A V_A} = 1 - \frac{1.230}{V_M}. \quad (5)$$

1.2. Probabilistic estimates of solvent content

About a decade ago, a conditional probabilistic estimate for possible unit-cell contents as a function of resolution, termed the Matthews probability (MP), was developed (Kantardjieff & Rupp, 2003) and the *MATTPROB* web applet, the corresponding probability distribution function and its parameters for cumulative resolution bins have been provided (http://www.ruppweb.org/mattprob/).

The original MP estimator, which has been cited in the literature over 300 times, has been implemented in some form in the major crystallographic structure-determination packages [*MATTHEWS_COEF* of *CCP*4 (Winn, 2003), *Phaser* (McCoy *et al.*, 2007) and *PHENIX* (Adams *et al.*, 2010)]. Since the original publication of the MP estimator in 2003, the available database of published structures in the PDB has increased almost fivefold, and a detailed analysis of the correlation of solvent content with crystal symmetry, space group and oligomeric state has been published (Chruszcz *et al.*, 2008). Given that the MP is based on prior information, revisiting the empirical foundation of this widely used solvent-content estimate seems to be appropriate. The accuracy of the MP estimates is particularly important in the case of large numbers of molecules in the asymmetric unit, where the solution landscape for the most probable number becomes increasingly degenerate. While the distinction between a

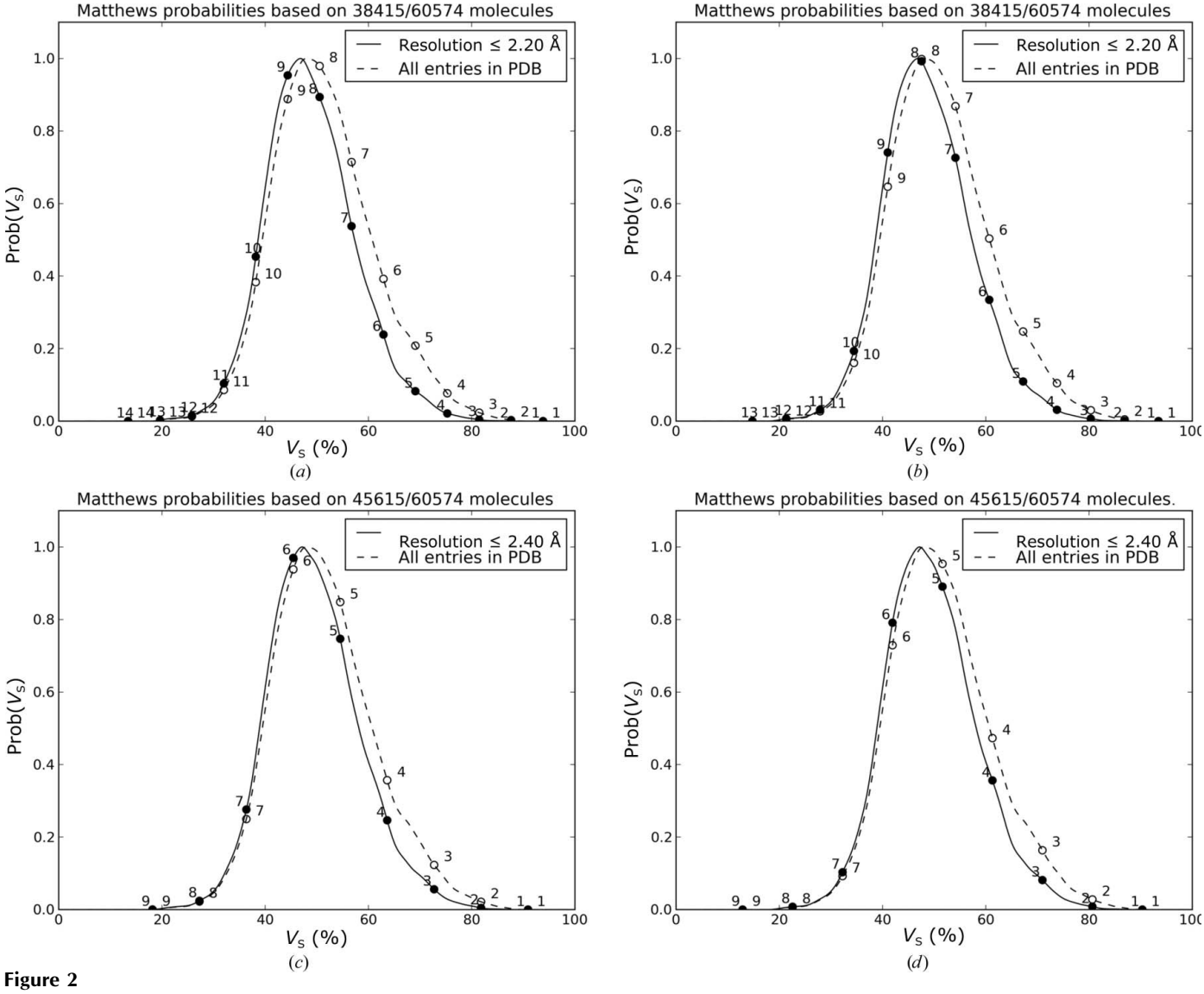

Figure 2
Correct molecular weight determines the outcome of the solvent-content predictions. Illustrated are the examples of PDB entries 3orx (top row) and 1xja (bottom row). (*a*, *c*) An incorrect, too low molecular weight estimated from generic mean residue weights (values of 9 and 6, respectively) overestimates the unit-cell content, while (*b*, *d*) the correct modular weight predicts the actually determined values (values of 5 and 8, respectively). The MP predictions were calculated and plotted with the kernel density estimator implemented in *MATTPROB* on http://www.ruppweb.org and described in §3.3.

single molecule or a dimer in the asymmetric unit is almost always clear, the discrimination between a pentamer or hexamer is much more subtle and depends on accurate prior information (including the correct molecular weight; §2.1.2). A unique feature of the 2003 MP calculator is its implicit assumption of a weak Bayesian prior that the observed resolution represents an empirical lower limit for the true diffraction potential of a crystal: the selected crystal has been demonstrated to diffract *at least* to the reported particular resolution under given experimental circumstances, but in principle could have diffracted better. The assumption of this Bayesian prior is also compatible with the fact that the distribution of the reported resolution cutoffs is distinctly skewed towards cutoff values higher than the mean $I/\sigma(I)$ mode of 2.0 (*cf.* Supplementary Fig. S1[1]). No agreement on objective criteria exists for the nontrivial selection of a resolution cutoff for model refinement, but it seems that present resolution cutoffs under-report the actual diffraction potential of the crystals (Diederichs & Karplus, 2013; Luo *et al.*, 2014).

The availability of a larger 2013 training data set facilitates the examination of secondary effects such as the dependence of the MP on molecular weight or symmetry. These analyses have been attempted before, but have not revealed significant correlations in previous smaller training data sets.

2. Parameters affecting solvent-content predictions

2.1. Fundamental dependencies

Equation (5) demonstrates that on a fundamental level the calculated solvent fraction is a function of (i) the (asymmetric) unit-cell volume, (ii) the molecular weight of the molecular species occupying that (asymmetric) unit cell and (iii) the specific density (or its reciprocal the partial specific volume) of the protein species.

2.1.1. Calculation of solvent content and reported solvent content. In our analysis V_M is calculated straightforwardly according to (1) from the asymmetric unit-cell volume V_A (using parsed unit-cell parameters and general position multiplicity of the reported space group) and the molecular weight M of the asymmetric unit contents (computed from the SEQRES records). The solvent content V_S then follows directly from (5). V_M was not calculated for 338 PDB protein structure entries as a result of nonstandard space-group settings, and 297 PDB entries with calculated $V_M >$ 10.0 Å^3 Da^{-1} ($V_S > 88\%$) or $V_M < 1.23$ Å^3 Da^{-1} (negative V_S) were removed as suspected outliers. For a comparison of deposited solvent content with our computed values, we could not include 1805 PDB entries owing to a lack of reported V_M or V_S values in the PDB files, but of the remaining 69 895 entries about three quarters report percentage V_S values within a single percentage unit of our computed values. Only 1.9% of entries differed by more than 10 units in percentage V_S. With the exception of obvious outliers, the solvent content reported in the PDB entries seems to be reasonably accurate.

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: DZ5231).

Table 2
Test cases with low molecular weight and high copy number.

In none of these cases was any difference observed whether the molecular-weight dependency of ρ according to (4) was accounted for or not. N, observed oligomerization state; P, predicted most probable state.

PDB entry	M (Da)	N	P	Reference
n/a	5000	12	12	n/a
1qoh	7204	20	21	Ling *et al.* (2000)
4otc	6795	9	9	Taylor *et al.* (1998)
1h64	8472	28	28	Thore *et al.* (2003)

2.1.2. Protein molecular weight. The accuracy of the molecular weight can and does have a significant effect on the V_S estimate, particularly in cases of large numbers of molecules in the asymmetric unit cell. Two examples that have been brought to our attention illustrate the importance of entering the correct actual molecular weight of the protein in the MP calculation. In case of PDB entry 3orx (Sadowsky *et al.*, 2011) using the actual molecular weight of 36 051 Da calculated from the sequence predicts the presence of an octamer, while an inaccurate estimate of 34 349 Da [given 316 residues and a (species-dependent) mean residue molecular weight of 108.7 Da optionally available in the 2003 MP estimator] predicts nine molecules in the asymmetric unit (Fig. 2). Similarly, PDB entry 1xja (Weldon *et al.*, 2007) is correctly predicted as a pentamer with the actual molecular weight, while the generic molecular weight calculation from the number of residues estimates six molecules in the asymmetric unit cell. The correct results obtained by both the classical MP calculator and the new kernel estimator (§3.3) are shown in Fig. 2. Given the importance of the accurate molecular weight for the most accurate solvent-content probabilities, the choice to compute M from residue number is no longer available in our *MATTPROB* application, and a link to the calculation of the actual molecular weight from the sequence is provided.

2.1.3. Protein specific density. Questions have been raised whether protein specific density is a function of the molecular weight (Fischer *et al.*, 2004), although deviations from the experimentally determined average value for the protein specific density ρ of $\sim$1.350 g cm^{-3} (or a $\overline{v}$ of $\sim$0.741 cm^3 g^{-1}) have only been found for proteins below 20 kDa molecular weight. Otherwise, as discussed in Quillin & Matthews (2000), there seems to be little indication of deviation from the experimental average value, because the only significant differences in reported theoretical values seem to result from systematic errors in Voronoi volume calculations.

To test whether a correction of ρ is necessary in practice and in extreme cases, we implemented a correction to V_P (4) following the empirical fit by Fischer *et al.* (2004) for the protein specific density ρ as a function of molecular weight M,

$$\rho(M) = \rho_\infty + \Delta\rho_0 \exp(-M/K), \qquad (6)$$

with the experimentally determined value of 1.350 g cm^{-3} and the remaining parameters those established by Fischer *et al.* (2004). As the maximum difference of densities is in the few percent range, it is already obvious that in the case of clear MP predictions when only a few molecules are present in the

asymmetric unit no changes for the most probable number of molecules are expected. We therefore examined the situation for various examples of homo-oligomers with a high number of units in the asymmetric unit cell and low molecular weight (i) without ρ correction, (ii) with the ρ correction of the monomer applied to all possible oligomers and (iii) with the ρ correction applied to the corresponding oligomer molecular weight. Case (iii) seems to represent the most realistic scenario, as no dependency of the solvent content on the oligomerization state of a protein in the asymmetric unit has been discovered (Chruszcz *et al.*, 2008; *cf.* the discussion and findings in §2.2.2).

For a sample of homo-oligomer entries in the PDB with the highest number of molecules in the asymmetric unit (N) and a low molecular weight M, as well as for a putative 12-mer of 5 kDa, no changes in the predicted number of monomers (P) in the asymmetric unit as a function of $\rho(M)$ resulted, regardless of whether the most probable values were actually observed or not. The results are summarized in Table 2. A correction to the solvent-content predictions as a function of $\rho(M)$ does not improve the MP model, and the assertion of Quillin & Matthews (2000) that the empirical average specific protein density value of 1.350 g cm^{-3} suffices in practice for solvent-content estimates still holds.

2.1.4. General remarks about high homo-oligomeric states. The fact that MP predictions become degenerate for high oligomerization states [meaning that multiple high numbers of the (same) molecule in the asymmetric unit are almost equally probable] is less of a concern than it may appear from a purely statistical point of view. Additional prior information often exists about probable oligomeric states from biological evidence. As a consequence, the MP calculator allows the entering of information about known or presumed obligate oligomerization states (known dimer, trimer *etc.*) and searches only for multiples of those known, presumed obligate, homo-oligomers.

In addition, the simple, but often neglected, methods of native Patterson and self-rotation Patterson function analysis, as reviewed, for example, in Rupp (2009), will frequently reveal the presence of noncrystallographic symmetry indicating actual oligomerization states. Finally, many high oligomerization states that are in principle possible are rarely found in practice; for example, a predicted 15-mer or 17-mer is probably a dimer of octamers, a tetramer of tetramers or some reasonable 16-mer assembly indicated by other, for example biological, evidence. In some cases the revealed local symmetry may be compatible with and indicative of higher space-group symmetry. Should, for example, a presumed proper NCS axis coincide with (that is, become) a crystallographic axis, there will be not sufficient space available in the asymmetric unit given the selected oligomeric state, providing further incentive to carefully examine the (space) point-group assignment or the presumed oligomerization state.

It is also evident that there are no 'wrong' predictions provided by a probabilistic model calculator. As long as the solvent content of an unknown protein crystal is close to the mode of the empirically observed distribution, the prediction for the most probable value will be the actual one, even for high oligomerization states. It is in the case of unusual solvent content that the actual value will be less probable than a predicted value based on the empirical probability distribution function.

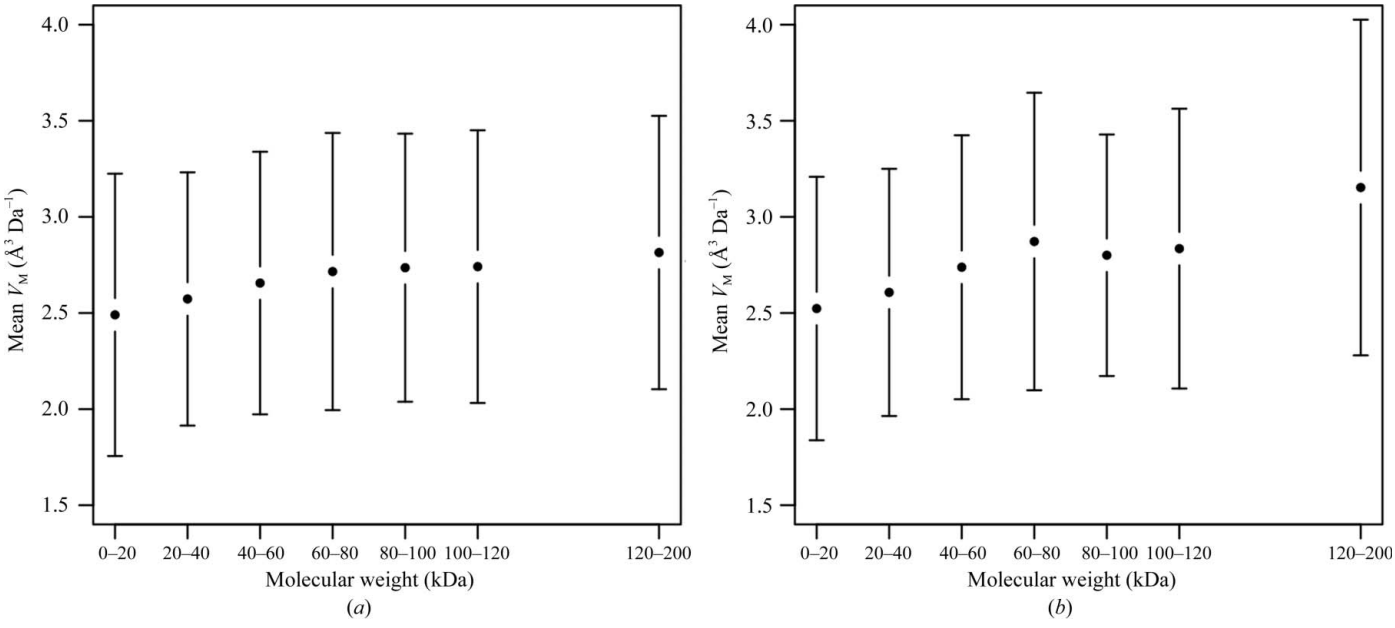

Figure 3
Dependence of mean V_M on molecular weight. (*a*) Total molecular weight; (*b*) restricted to the 50 190 PDB entries that contain monomers or homo-oligomers only. Dots correspond to the mean V_M of the corresponding bin shown on the x axis, and bars represent standard deviations of V_M for the respective bin. From both graphs, the justification for a linear model is not obvious, and past 60–80 kDa the initial increase of V_M with molecular weight seems to stagnate. Over the entire molecular-weight range, the increase based on a linear model would be only about 8–12%.

2.2. Empirically discovered dependencies

In addition to the fundamental mathematical dependencies outlined in §2.1, the solvent-content data extracted from the PDB entries have been repeatedly analysed for empirical parameters that may affect the actual solvent content.

2.2.1. Symmetry. Matthews himself noted in 1968 that

> there does not appear to be any correlation between the degree of symmetry of the crystals and the amount of solvent contained in them.

In the analysis of a larger data set in 2008 (Chruszcz *et al.*, 2008), a correlation was found between V_M and the 'degree of symmetry' [a linear measure that in essence is the L value of Wukovitz & Yeates (1995) and the additional assumption that the hexagonal system with the same L value of 2 as the tetragonal system has a higher symmetry]. As Chruszcz *et al.* (2008) note, there is no satisfactory explanation for this weak correlation, and the justification for a linear regression against the categorical L measure is not obvious. A plot (without categorical regression) showing the same weak trend in the present data set has been deposited as Supplementary Fig. S2. Ultra-tight structures with solvent contents below the closed-sphere packing limit (26%) have been reported in low-symmetry as well as in high-symmetry space groups (Trillo-Muyo *et al.*, 2013), essentially reflecting the expected empirical space-group frequency distribution (Wukovitz & Yeates, 1995; Kantardjieff & Rupp, 2003; Chruszcz *et al.*, 2008). We see no justification to include a linear regression model of the categorical relationship between symmetry and solvent content in the MP predictions.

2.2.2. Protein molecular weight. From the few data available to him in 1968, Matthews concluded that

> there appears to be a tendency for molecules of higher molecular weight to form crystals containing a relatively higher fractional volume of solvent.

Such a trend has been confirmed by Kantardjieff & Rupp (2003) but was not found to be significant based on the 2003 data set, and no trend was reported by Chruszcz *et al.* (2008).

The crucial question here is whether any linear regression of resolution *versus* molecular weight (binned or not) is in fact physically sensible. Given that the molecular weight of single protein molecule chains rarely exceeds approximately 50–100 kDa, it is reasonable to assume that all molecular weights at the extreme high end of the distribution present larger oligomeric assemblies of smaller subunits (an extreme is exemplified by the 60-fold symmetry in some virus capsids). Fig. 3 shows that when the total molecular weight is plotted against V_M, a weak dependency can in fact be observed for binned molecular weights up to 60–80 kDa, but the curve then flattens and no statistically significant dependency for high molecular weights can be inferred. Any linear regression of the overall M would therefore vastly overestimate the molecular-weight correction for very high total molecular-weight oligomers. A similar but weaker trend is observed if only the molecular weight of the monomer of (homo)oligomer chains is examined. The principal component analysis discussed in §3.1 also indicates that molecular weight is only a weak and not a primary determinant for solvent content.

2.2.3. Oligomerization state. We also examined the dependency of solvent content on the oligomerization state[2] of (homo-oligomeric) proteins and discovered no dependency (Fig. 4), confirming the findings of Chruszcz *et al.* (2008). The absence of any dependency means that a large oligomer with high M containing multiple small molecular-weight subunits does *not* tend to have a higher solvent content, which again makes a linear regression of solvent content *versus* molecular weight questionable for higher molecular weights. We therefore do not consider a linear regression model of the relationship between molecular weight and resolution physically meaningful.

2.2.4. Polymorphism. Analysis of polymorphism (*i.e.* the same molecular moiety crystallizing in different crystal forms) is not possible based on solvent content alone, although certain polymorphs can be distinguished. Different space groups clearly identify different polymorphs, but within the same space group V_S alone does not suffice to clearly distinguish polymorphs. It is possible that within the same space group polymorphs with different packing exist that have similar V_S and even similar unit-cell parameters. We have identified in our data set 1683 protein chains that have been crystallized as homo-oligomers (including monomers) in more than one space group. Bovine ribonuclease A has been deposited in seven different space groups, the most that we have observed. Space groups $P2_1$ and $P3_221$ have the highest

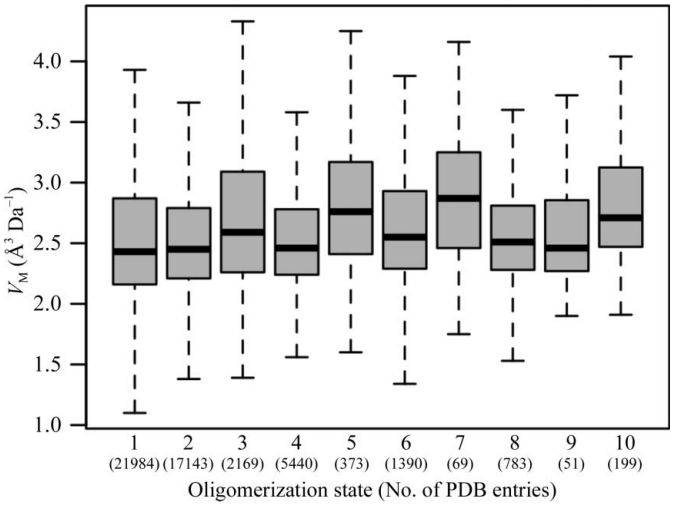

Figure 4
Relationship between protein oligomerization status and V_M. Box plot for the 50 190 PDB entries that were identified as homo-oligomers, including monomers ranging up to decamers. Higher oligomerization states have been omitted owing to limited number counts. Shown on the x axis is the oligomerization state, and the corresponding number of PDB entries found in our data set is shown in parentheses. There is no apparent relationship between these two variables. Also notice that some multimers such as heptamers have low number counts.

[2] A surprising finding during the analysis of the oligomeric state was that only half of the PDB entries where an homo-oligomeric state was determined from the SEQRES records also had a NCS matrix record, *i.e.* it seems that (at least according to the PDB entries) they were refined without the use of NCS restraints.

number of representatives and exhibit clearly different distributions of V_S, with means of 42.53 (±2.54) (99% confidence interval ±1.3) and 56.46 (±1.19) (99% confidence interval ±0.7), respectively. Conversely, we find the solvent-content distributions of human cathepsin K in space groups $P2_12_12_1$ and $P4_32_12$ highly overlapping. An extreme example of intra-space-group variability is given by bovine trypsin crystallized in space group $P2_12_12_1$, where (in the absence of any detailed packing analysis) we observe a multimodal distribution of V_S with modes ranging from approximately 43 to 59% solvent content (Supplementary Fig. S3).

2.2.5. Resolution. The most significant empirical correlation remains the clear tendency established by Kantardjieff & Rupp (2003) for the solvent content to decrease with higher resolution, meaning that crystals with tighter molecular packing tend to diffract better. As shown in Fig. 5, when the current 2013 data are binned similarly as in Kantardjieff & Rupp (2003), the same significant trend is observed in a simple empirical, linear fit of V_M against the resolution d. Other plausible, reciprocal resolution-dependent models that we tested ($1/d$, $1/d^2$) show a worse correlation. Irrespective of whether a linear relationship between V_M and resolution can be rationally explained, so far it remains a parsimonious and sufficient model for MP calculations.

A fundamental question that arises is whether the binned distributions of V_M are the best way to predict the MPs, or whether a more direct way through (i) a nonparametric fit and (ii) directly through V_S can improve the predictions and simplify future updates by eliminating the need for binning and adjusting multiple fit parameters of the modified logistic extreme function as implemented by Kantardjieff & Rupp (2003). The situation is examined in the following section.

3. Analysis of the 2013 data and implementation of a nonparametric probability estimator

The original calculation of the MP was based on a purely empirical but nonetheless highly successful logistic regression function of V_M *versus* resolution. To maintain compatibility with the various previous implementations of the MP calculator, we have kept this function but have updated the downloadable parameter set in the *MATTPROB* calculator (http://www.ruppweb.org/mattprob). Given the fivefold increase in the number of available PDB entries, it is worthwhile examining (i) which changes or improvements over the analysis of the 2003 data exist and (ii) which other empirical correlations are practically significant and could be included in a probabilistic calculation of the solvent content. The large amount of data may also (iii) allow a more direct, empirical estimate for the solvent content as a function of resolution than the parameterized fit to binned resolution ranges as implemented in the original *MATTPROB* calculator.

3.1. Principal component analysis

To examine whether molecular weight, resolution or both should be used in the MP predictor, we performed principal component analysis on the data set of 50 190 entries containing homo-oligomeric proteins using the observed variables molecular weight per chain, resolution and V_M. A Cattell scree plot (Cattell, 1966) was used to determine the number of significant factors for linear fitting. Here, the eigenvalues of the principal components are plotted in decreasing order, and all components with eigenvalues less than 1 (which corresponds to the information contributed by an average single variable) are then dropped as not meaningful. In our case, the scree plot suggests that a single variable is already sufficient for a linear model: the first principal component explains about 51% of the total variance of the data set. Resolution and V_M contribute more to the first principal component than does molecular weight, and the sample correlation coefficient between resolution and V_M (0.40) is higher than that between molecular weight and V_M (0.15). These numbers minimally increase when investigating all 60 218 PDB entries and total molecular weight (54% of the total variance is explained by principal component 1, and the correlation coefficients between resolution and V_M and between M and V_M are 0.43 and 0.16, respectively). Therefore, in the kernel density estimator (§3.3) we continue to use resolution as the sole parameter for the dependency of V_M in a linear model.

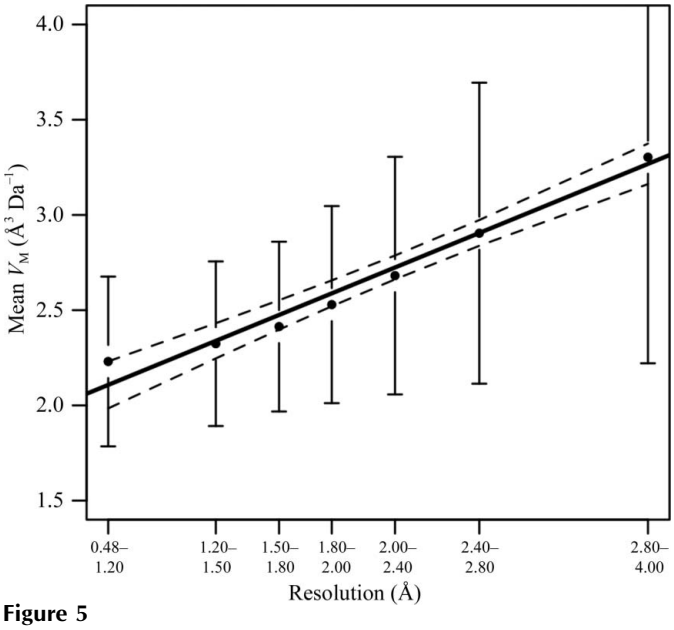

Figure 5
Linear regression on mean V_M values *versus* resolution. The mean values of V_M (shown as filled circles) were computed for each of the resolution bins indicated on the x axis for the subset of 50 190 homo-oligomers. Weighted linear least-squares regression analysis was applied to the mean V_M values with weights corresponding to the standard deviation (shown as vertical error bars) of the V_M distribution for the respective interval. The solid line represents the obtained linear regression (R^2 = 0.9796, p value = 2.031×10^{-5}), whereas the dashed lines indicate the upper and lower confidence interval at the 95% level. Extrapolation of the regression line intercepts the y axis at V_M = 1.69 Å^3 Da^{-1} (not shown in the graph), which is equivalent to a solvent content of 27%, corresponding to the empty space in close-packed spheres (approximately 26%). Over the examined resolution range, the mean V_M increases by approximately 57%.

3.2. Update of the parameterized 2003 MP calculator

To maintain compatibility with the original Matthews probability calculator *MATTPROB* (http://www.ruppweb.org/mattprob) and its implementation in other crystallographic programs, we have updated the binned parameter set used in the original parameterized extreme function fit (Kantardjieff & Rupp, 2003). The functionality remains unchanged and is described in the original publication and on the web site, from which the updated parameter files can be downloaded. Fig. 6 shows the relative changes between the 2003 and 2013 data for proteins (1.9 Å resolution bin), protein–DNA complexes and nucleic acids (all resolutions).

3.3. Nonparametric kernel density estimator of Matthews probabilities

We have repeated the raw data extraction and analysis described in Kantardjieff & Rupp (2003) using the same Fortran software with only minor adaptions to reflect the updated PDB format (Henrick *et al.*, 2008). We used the PDB advanced query interface to retrieve 77 481 crystal structures deposited in the PDB as of 6 February 2013. Obvious outliers with exceptionally low or high V_M values were removed according to §2.1.2 and the protocol presented in Kantardjieff & Rupp (2003). A nonredundant data set was constructed from this selection where only the highest resolution entry was chosen within a group of entries sharing the same space group and a maximum of 1% difference in molecular weight and unit-cell volume. Applying all of these steps resulted in a total of 60 218 protein structures, 998 nucleic acid structures and 2414 structures of protein–nucleic acid complexes. Analysis was carried out with the *R* statistical computation software (v.3.0.2; http://www.r-project.org).

The analysis carried out in Kantardjieff & Rupp (2003) was founded on fitting a modified logistic extreme function on binned V_M data, which was needed to model the tail of the V_M distribution. This tail is much less pronounced in the distribution of V_S values, which is reflected by a tenfold lower

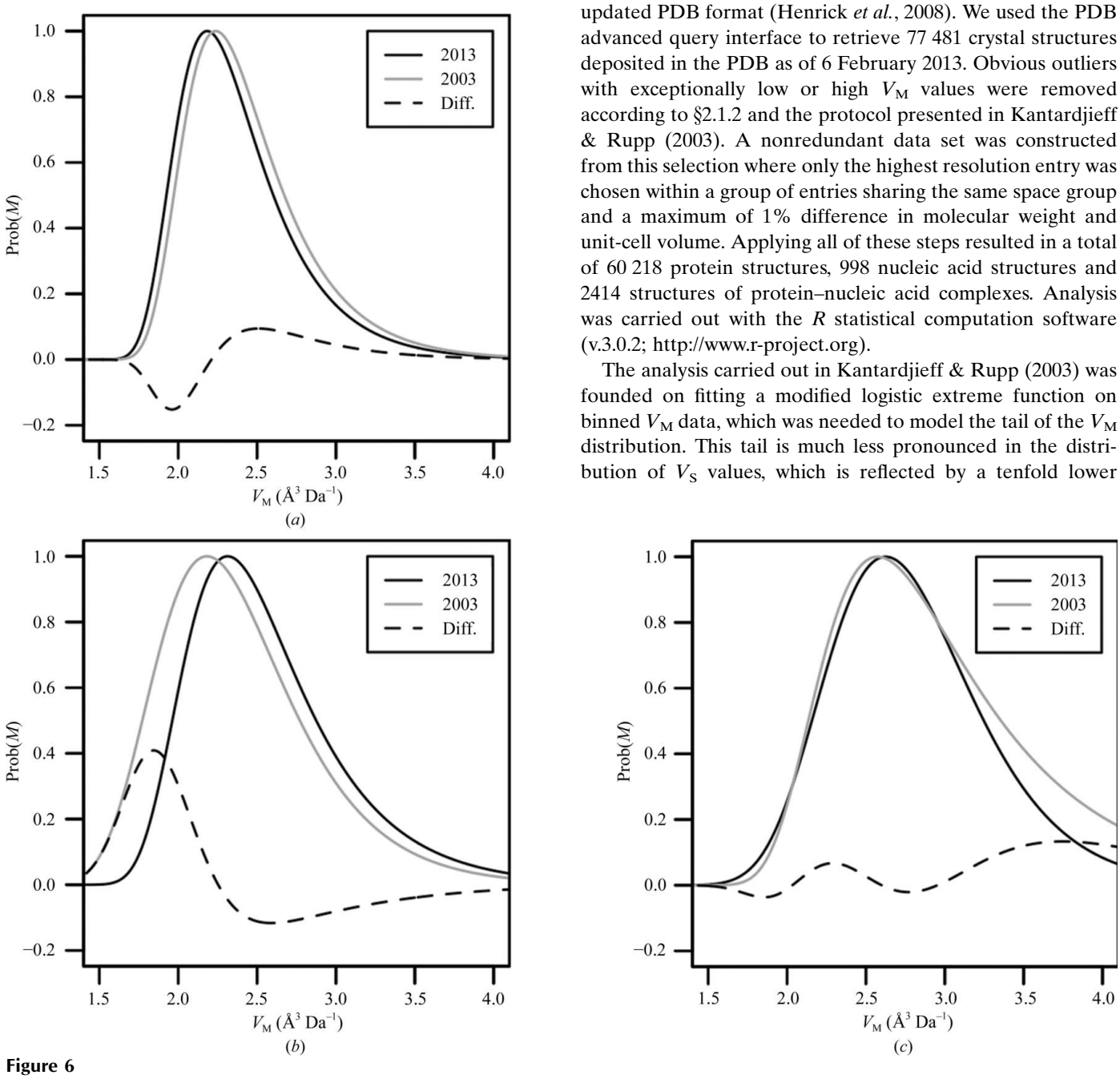

Figure 6
Relative changes in the 2013 update of the parameterized 2003 *MATTPROB* calculator. We compare the fitted, resolution-dependent original (gray line) and updated (black line) curves used in the 2003 *MATTPROB* calculator. The difference between the originally published and the updated curve is shown as a dashed line. In (*a*) the fitted curve is based on protein crystal structures with resolutions better than 1.9 Å. We observe a slight shift towards lower values of the Matthews coefficient, indicating that the number of higher packing protein structures increased in the 2013 database. The converse phenomenon is detected for nucleic acids, where we see a trend towards higher V_M in the set of 998 nucleic acid crystal structures (*b*). This trend is also visible less severely for the 2414 nucleic acid–protein complexes (*c*). The probability density function has been normalized to have a maximum of 1.

sample skewness of V_S (0.26 *versus* 2.57; see Fig. 7). The high skewness of the V_M distribution motivated us to favor V_S over V_M when reconstructing the probability density function

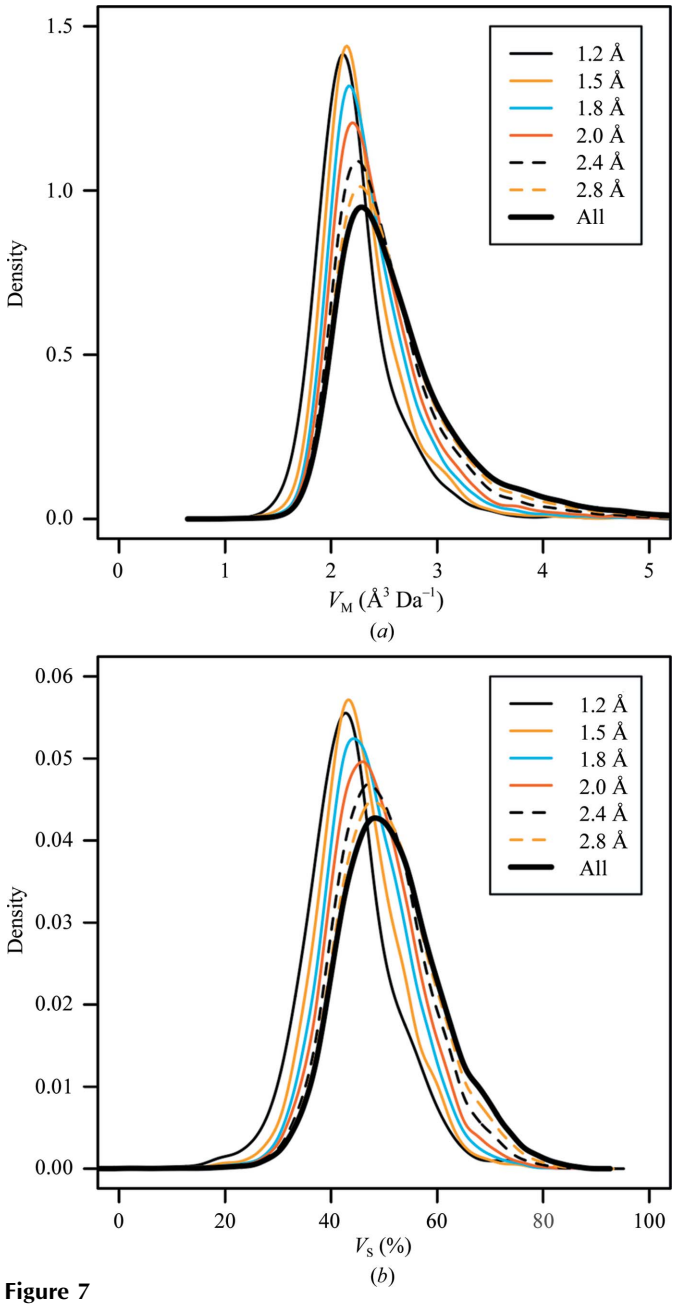

Figure 7
Distribution of V_M (*a*) and V_S (*b*) for 60 218 protein crystal forms for various resolution limits using binned kernel density estimates. Both distributions have in common that they are much broader when taking into account all resolutions of protein crystals, and peaks shift towards lower values of V_M (V_S) when limiting the observations to higher resolutions. For example, $V_M \geq 3$ Å^3 Da^{-1} ($V_S \geq 59\%$) is observed in 17.7% of protein structures determined at 2.8 Å resolution or better, whereas this value is seldom seen (4.2%) in crystal structures resolved at 1.2 Å resolution or better. The highest resolution threshold of 1.2 Å still contains 1538 protein crystals; that is, 2.6% of the overall number of protein structures investigated. Notice the broad tails of the V_M distributions (*a*) are less present in the V_S distributions (*b*). The latter appear to be distributed much more symmetrically, which is also expressed by the sample skewness of the distribution over all protein crystal forms of 2.57 Å^3 Da^{-1} (V_M) and 0.26 (V_S), respectively.

P(resolution, V_S) for pairs of resolution and solvent content observed in the PDB by using a two-dimensional kernel estimator (Fig. 8), which allows the computation of nonparametric, conditional MPs for a given resolution r, $P_r(V_S) = P(V_S|\text{resolution} \leq r)$, as illustrated in Fig. 7 for selected resolutions.

The Bayesian argument that the observed resolution represents a minimum resolution (that is, the crystal could actually diffract better but certainly not worse than the observed value) is maintained. Compared with the previously described parametric MP, the new kernel density estimator is independent of any binning and can therefore be probed with any resolution currently available in the PDB. From the definition of $P_r(V_S)$ it is evident that the calculator relies on fewer data with increasing experimental resolution, which should be kept in mind when applying this tool. The larger range of possible input resolutions generates better discrimination, which becomes apparent when reinvestigating the default MP example presented in Kantardjieff & Rupp (2003): a protein with 23 500 Da molecular weight located in a unit cell with dimensions a = 71.18, b = 79.38, c = 93.81 and space group $P2_12_12_1$. At 1.6 Å resolution, where the parametric MP utilizes data up to 1.5 Å resolution, a trimer is favored over a dimer compared with a resolution-independent approach. With the nonparametric MP we find the same result for 1.5 Å resolution, but the predicted oligomeric state is reversed for data resolved up to only 1.65 Å resolution, with only a small difference in probabilities. We have noticed that the 2013

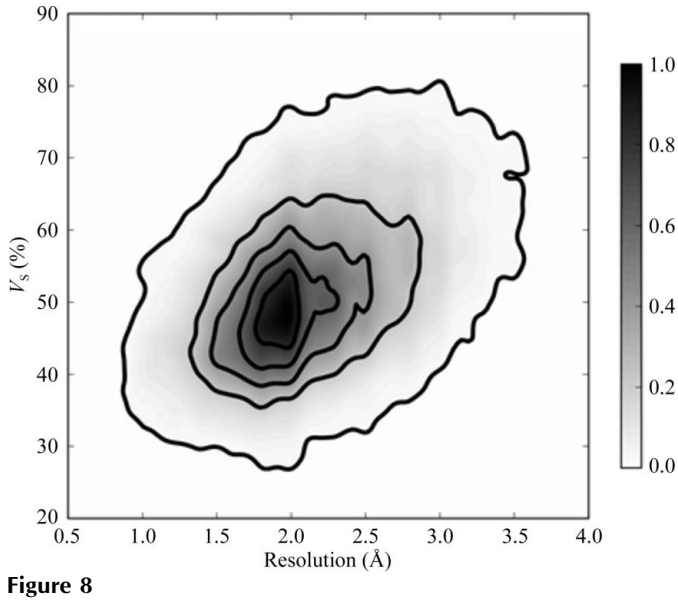

Figure 8
Two-dimensional density function of V_S for 60 218 protein crystal forms using binned two-dimensional kernel estimates. The scale of V_S from 20 to 90% on the *y* axis corresponds to values of V_M between 1.54 and 12.3 Å^3 Da^{-1}. The plot has been normalized to have a maximum value of 1. Isocontour lines are drawn as solid bold lines in increments of 0.2. There is a clear trend towards lower values of V_S for higher resolutions, which is in agreement with the findings from the previous publication (Kantardjieff & Rupp, 2003). Most density is centered about approximately 1.9 Å resolution and V_S = 50% and typical values for V_S range between 30 and 80%. This figure was generated with *matplotlib* (Hunter, 2007).

update of the parametric MP tends to have distributions shifted towards lower values of V_M, so that compared with the 2003 data set a higher oligomeric state is favored down to lower resolutions. In the 2003 example above a trimer is predicted for resolutions of 1.8 Å or better, but using the 2013 data set this prediction is maintained up to 2.0 Å resolution. This might be a result of the original parameter-fitting process necessary for the binned V_M data, since the nonparametric kernel density estimator predicts a trimer only for resolutions up to 1.6 Å. The independence of the kernel density estimator of any raw data binning, combined with the much larger learning data set compared with 2003, allows a more accurate estimate of the most probable number of molecules for any given resolution.

4. Conclusions

The kernel density estimator-based Matthews probability calculator provides an updated, parameter-free tool for estimating solvent content at any given resolution. It can be readily updated by simple filtering of the raw data without any need for binning and empirical parameter fitting. It validates the previous predictions based on the 2003 *MATTPROB* web application, which has also been updated with the 2013 data. No other correlations (dependency of ρ on molecular weight, symmetry, molecular weight or oligomerization state) were significant enough to be implemented in the prediction of solvent content as a function of resolution. The *MATTPROB* program is available online at http://www.ruppweb.org/mattprob, and the Python implementation and raw data as well as the filtering programs are available from the authors on request.

BR acknowledges support from the European Union under a FP7 Marie Curie People Action, grant PIIF-GA-2011-300025 (SAXCESS). The authors thank Dr Katherine Kantardjieff (California State University San Marcos, College of Science and Mathematics) for critical reading and comments.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Cattell, R. B. (1966). *Multivariate Behav. Res.* **1**, 245–276.
Chruszcz, M., Potrzebowski, W., Zimmerman, M. D., Grabowski, M., Zheng, H., Lasota, P. & Minor, W. (2008). *Protein Sci.* **17**, 623–632.
Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* D**58**, 494–506.
Diederichs, K. & Karplus, P. A. (2013). *Acta Cryst.* D**69**, 1215–1222.
Fischer, H., Polikarpov, I. & Craievich, A. F. (2004). *Protein Sci.* **13**, 2825–2828.
Henrick, K. *et al.* (2008). *Nucleic Acids Res.* **36**, D426–D433.
Hunter, J. D. (2007). *Comput. Sci. Eng.* **9**, 90–95.
Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.
Ling, H., Pannu, N. S., Boodhoo, A., Armstrong, G. D., Clark, C. G., Brunton, J. L. & Read, R. J. (2000). *Structure*, **8**, 253–264.
Luo, Z., Rajashankar, K. & Dauter, Z. (2014). *Acta Cryst.* D**70**, 253–260.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Matthews, B. W. (1976). *Annu. Rev. Phys. Chem.* **27**, 493.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mueller-Dieckmann, C., Panjikar, S., Schmidt, A., Mueller, S., Kuper, J., Geerlof, A., Wilmanns, M., Singh, R. K., Tucker, P. A. & Weiss, M. S. (2007). *Acta Cryst.* D**63**, 366–380.
Nagendra, H. G., Sukumar, N. & Vijayan, M. (1998). *Proteins*, **32**, 229–240.
Quillin, M. L. & Matthews, B. W. (2000). *Acta Cryst.* D**56**, 791–794.
Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st ed. New York: Garland Science.
Sadowsky, J. D., Burlingame, M. A., Wolan, D. W., McClendon, C. L., Jacobson, M. P. & Wells, J. A. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 6056–6061.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Taylor, A. B., Czerwinski, R. M., Johnson, W. H. Jr, Whitman, C. P. & Hackert, M. L. (1998). *Biochemistry*, **37**, 14692–14700.
Thore, S., Mayer, C., Sauter, C., Weeks, S. & Suck, D. (2003). *J. Biol. Chem.* **278**, 1239–1247.
Trillo-Muyo, S., Jasilionis, A., Domagalski, M. J., Chruszcz, M., Minor, W., Kuisiene, N., Arolas, J. L., Solà, M. & Gomis-Rüth, F. X. (2013). *Acta Cryst.* D**69**, 464–470.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Weldon, J. E., Rodgers, M. E., Larkin, C. & Schleif, R. F. (2007). *Proteins*, **66**, 646–654.
Winn, M. D. (2003). *J. Synchrotron Rad.* **10**, 23–25.
Wukovitz, S. W. & Yeates, T. O. (1995). *Nature Struct. Biol.* **2**, 1062–1067.
Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 99–107.