Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Use of massively multiple merged data for low-resolution S-SAD phasing and refinement of flavivirus NS1

David L. Akey,[a] W. Clay Brown,[a] Jamie R. Konwerski,[a] Craig M. Ogata[b] and Janet L. Smith[a]*

[a]Life Sciences Institute, University of Michigan, 210 Washtenaw Avenue, Ann Arbor, MI 48109-2216, USA, and [b]GM/CA@APS, Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439, USA

Correspondence e-mail: janetsmith@umich.edu

Received 8 June 2014
Accepted 30 July 2014

PDB reference: *West Nile virus* NS1, 4tpl

An emergent challenge in macromolecular crystallography is the identification of the substructure from native anomalous scatterers in crystals that diffract to low to moderate resolution. Increasing the multiplicity of data sets has been shown to make previously intractable phasing problems solvable and to increase the useful resolution in model refinement. For the *West Nile virus* nonstructural protein 1 (NS1), a protein of novel fold, the utility of exceptionally high multiplicity data is demonstrated both in solving the crystal structure from the anomalous scattering of the native S atoms and in extending the useful limits of resolution during refinement. A high-multiplicity data set from 18 crystals had sufficient anomalous signal to identify sulfur sites using data to 5.2 Å resolution. Phases calculated to 4.5 Å resolution and extended to 3.0 Å resolution were of sufficient quality for automated building of three-quarters of the final structure. Crystallographic refinement to 2.9 Å resolution proceeded smoothly, justifying the increase in resolution that was made possible by combining multiple data sets. The identification and exclusion of data from outlier crystals is shown to result in more robust substructure determination.

1. Introduction

The potential for using the anomalous scattering of native elements (sulfur and phosphorus) for determination of the substructure and subsequent phasing has long been recognized (Hendrickson & Teeter, 1981). As the anomalous signal of native atoms is weak when compared with traditional heavy-atom incorporation, accurate determination of the anomalous differences typically requires some combination of low-energy X-rays, high-resolution data and high-multiplicity data (Lehmann *et al.*, 1993; Ramagopal *et al.*, 2003; Wagner *et al.*, 2006; Doutch *et al.*, 2012). Recent experiments have demonstrated the utility of high-multiplicity data to enhance the weak anomalous signal for solving crystal structures by single-wavelength anomalous diffraction (SAD) with data of moderate (2.3–2.8 Å) resolution (Liu *et al.*, 2012, 2013). Simultaneously, new metrics have redefined the limits of useful data for crystallographic refinement using high-resolution data ($d_{\rm min}$ beyond 2.0 Å; Diederichs & Karplus, 2013; Karplus & Diederichs, 2012). Both of these advances suggest beneficial effects from increasing the number of independent observations of each Bragg reflection in routine problems in macromolecular crystallography.

Accurate measurement of very small signals is critical at the limit of diffraction, where intensities are weak, and in the measurement of Bijvoet differences in cases of weak

anomalous scattering (for example, native sulfur and phosphorus). Residual 'R' values ($R_{\rm merge}$, $R_{\rm ano}$) and 'signal to noise' [average $I/\sigma(I)$, average $\Delta F^{\pm}/\sigma\Delta F^{\pm}$] have traditionally been used as proxies for accuracy. However, R values increase with either increasing multiplicity or decreasing intensity, and the correlation with data accuracy remains unclear. Correction factors for multiplicity in intensity data have been addressed with the introduction of $R_{\rm p.i.m.}$ (Weiss, 2001) and $R_{\rm meas}$ (Diederichs & Karplus, 1997), but neither of these metrics assesses the accuracy of low-intensity data. The correlation coefficient between randomly selected half data sets, $CC_{1/2}$, and the analytically related estimate of the correlation between the full data set and the 'true' intensities, CC* {$= [2CC_{1/2}/(1 + CC_{1/2})]^{1/2}$}, were introduced as more reliable estimates of the overall accuracy of a data set (Karplus & Diederichs, 2012). In contrast to R values, the correlation coefficients and the average $I/\sigma(I)$ benefit from increased multiplicity.

Multi-crystal data sets were the norm in the pre-cryogenic era of macromolecular crystallography (Blundell & Johnson, 1976), but are uncommon today. However, the effects of radiation damage from increasingly brilliant synchrotron sources, especially as applied to microcrystals (Smith *et al.*, 2012; Cherezov *et al.*, 2009), generally require the collection of data from many samples. Recent advances in detector technology have speeded data collection and made the collection of ultrahigh-multiplicity data feasible for many projects. While merging data from many crystals should significantly increase the accuracy of intensities and Bijvoet differences, a strategy for determining which data to include in a final meta data set needs to account for differences in intensity measurements owing to both radiation damage (Garman, 2010) and sample non-isomorphism. Outlier non-isomorphic crystals can be rapidly identified by comparison of unit-cell parameters (Foadi *et al.*, 2013) or by rigorous comparison of intensities or anomalous differences (Giordano *et al.*, 2012; Liu *et al.*, 2012).

Here, we present the case of the *West Nile virus* NS1 protein (Akey *et al.*, 2014), where massively multiple merged data from many weakly diffracting crystals were used to determine the sulfur substructure, to phase at low resolution and to extend the limit of the data used during refinement. NS1 is an essential protein of the *Flavivirus* genus that is required both for genome replication and immune-system evasion (Muller & Young, 2013). Flaviviruses cause many diseases for which neither antiviral drugs nor vaccines exist, including dengue fever and West Nile fever. NS1 is highly glycosylated, membrane-associated and replete with disulfide bonds, and thus the production of stable native NS1 was challenging and slowed progress in understanding the molecular basis of its multiple and enigmatic functions. As the production of selenomethionyl protein in eukaryotic cells can be complicated and since NS1 contains an abundance of native anomalous scatterers, with six disulfides and five methionines per 352-residue monomer (370 with the expression tag), S-SAD was our first choice for phasing (Liu *et al.*, 2012, 2013). Data were collected from 28 crystals of recombinant NS1 purified from insect cells. Initially, the combined data did not have a useful anomalous signal. We used a culling strategy to identify crystals that did not appear to contribute to the anomalous signal and eliminated the data from ten crystals. The resultant data set was successfully used for determination of the sulfur substructure at 5.2 Å resolution, for the calculation of initial phases to 4.5 Å resolution, for phase extension to 3.0 Å resolution and for crystallographic refinement to 2.9 Å resolution.

2. Methods and results

2.1. NS1 expression and crystallization

Production and crystallization of NS1 has been described by Akey *et al.* (2014). Both dimeric and hexameric forms of the protein were detected by gel filtration. NS1 ($\sim$7 mg ml^{-1} in 50 m*M* Tris pH 8.5, 50 m*M* ammonium sulfate, 10% glycerol) was crystallized by vapor diffusion against a reservoir solution consisting of 20–25% polyethylene glycol (PEG) 3000 or PEG 3350, 5% glycerol, 150–300 m*M* sodium citrate pH 5.5. Matthews coefficient probability calculations with two, three or four copies of NS1 in the asymmetric unit of crystals of space group *P*321 yielded estimated solvent contents of 74, 61 and 48%, respectively (Matthews, 1968; Kantardjieff & Rupp, 2003). Crystals of hexagonal plate morphology (50–150 µm across the threefold axis, 20–50 µm along the threefold axis) were harvested directly from the growth solution without additional cryoprotection and flash-cooled in liquid nitrogen. Approximately one in every eight crystals yielded suitable diffraction, and the degree of heterogeneity of the samples was unknown.

2.2. Data collection and processing

Data were collected on GM/CA@APS beamline 23ID-D using a MAR Mosaic 300 CCD detector at 225 mm from the sample with a 105 mm helium box mounted on the front. Data were collected at 7.1 keV as this energy was suggested to be a suitable compromise between the strength of the sulfur anomalous signal and the absorption of X-rays by the sample (Liu *et al.*, 2012). Assuming that each of the six disulfides increases the scattering power of the two cysteines by $2^{1/2}$, the NS1 crystals have an expected anomalous signal of 1.49% at 7.1 keV (Hendrickson *et al.*, 1985). The 75 × 35 µm (horizontal × vertical, full-width half-maximum) beam was reduced to 50 × 35 µm by slits, approximating the dimensions of the crystals. To maximize the detection of the anomalous signal, data were collected using an inverse-beam protocol (Hendrickson *et al.*, 1985; Smith & Hendrickson, 2001) with 0.5° rotation per image and 5° of data (ten images) per wedge for a total of 2 × 90° of data from each crystal. The morphology of the crystals allowed us to determine the crystal orientation visually, and data were collected alternately starting along the threefold (face), the crystal edge or at a 45° offset. Thus, for any subset of reflections data were collected for both Friedel mates early in dose time from a significant number of samples. In total, 12 data sets were collected starting from a face orientation, 11 from an edge orientation and five from a 45° offset.

Data were processed using *XDS* and were scaled and merged using *XSCALE* (Kabsch, 2010*a*,*b*). Although *XDS* required separate processing of the forward 90° and inverse 90° of data for each crystal, a common crystal orientation matrix was applied to both 90° data sweeps. Friedel pairs were treated separately in scaling (STRICT_ABSORPTION_CORRECTION=TRUE and FRIEDEL'S_LAW=FALSE). Subsequent analysis confirmed the importance of these settings (based on failed attempts to solve the S substructure after scaling with STRICT_ABSORPTION_CORRECTION =FALSE). The 28 crystals appeared to be isomorphous based on the similarity of their unit-cell parameters; however, initial attempts to merge data from all 28 crystals were unsuccessful. Although data from individual crystals had positive anomalous correlation coefficients (AnomCC, calculated between random half sets of anomalous differences) at low resolution (50.0–8.57 Å), the merged meta data set did not. Suspecting that some incompatible data sets 'poisoned' the final data and obliterated the anomalous signal, we identified outlier crystals by pairwise comparison of data sets (28 crystals × 2 wedges each). Data-set pairs were scaled together and the AnomCC and $R_{\rm merge}$ values at low resolution (to $d_{\rm min}$ = 8.57 Å) were evaluated for compatibility. Crystals were excluded from the final data set if either the forward or inverse data consistently had negative AnomCC values or $R_{\rm merge}$ values greater than 7.5% when scaled with individual data sets from other crystals. We settled on the data from 18 crystals, eight of which were collected starting with the beam perpendicular to the crystal face, eight with it perpendicular to the edge and two at a 45° offset.

The data (forward and inverse combined, d = 50.0–2.9 Å) for 18 individual crystals had $\langle I/\sigma(I)\rangle$ values of 10.7–3.0, an $R_{\rm merge}$ of 12.2–45.4% and a $CC_{1/2}$ of 99.6–93.8%. When scaled into a single set (*XSCALE*), the combined data set had an $\langle I/\sigma(I)\rangle$ of 19.6, an $R_{\rm merge}$ of 31.3% and a $CC_{1/2}$ of 100% (Table 1, Supplementary Table S1[1]). The resultant data set had an anomalous multiplicity of 100 to $d_{\rm min}$ = 3.0 Å. Importantly, whereas the AnomCC and anomalous signal (average $\Delta F^{\pm}/\sigma\Delta F^{\pm}$) were weak for individual crystals, the AnomCC was 60% for data to $d_{\rm min}$ = 8.57 Å and was positive to ∼5.4 Å, with an anomalous signal greater than 1.0 to the same limit (Figs. 1*a* and 1*b*). While data from individual crystals were weak when assessed by traditional metrics, the combined data were of much improved quality according to the $\langle I/\sigma(I)\rangle$ and $CC_{1/2}$ metrics, although not according to $R_{\rm merge}$ (Figs. 1*c*–1*e*).

2.3. Determination of anomalous scatterer substructure

SHELXD was used for substructure determination (Sheldrick, 2010). 10 000 trials were requested for each *SHELXD* run using a range of resolution cutoffs, 4.5–5.5 Å, in 0.1 Å increments. Searches were conducted in both *P*321 and the enantiomorphic space groups $P3_121$ and $P3_221$. Encouragingly, significant trial solutions were found for only one space group, *P*321, and not for either of the enantiomorphic alternatives. A scatter plot of *SHELXD* correlation coefficients between observed anomalous difference E values and those calculated from each of the 10 000 trials for all data and for the weakest E values ($CC_{\rm all}$ *versus* $CC_{\rm weak}$) showed a continuum of trial solutions extending from the main grouping of presumably random trials (Fig. 1*f*) and not the typical cluster of strong trials well separated from the main grouping (Liu *et al.*, 2012). We evaluated the results of several *SHELXD* runs with different parameters (space group, $d_{\rm min}$, number of NS1 monomers per asymmetric unit) by comparing the maximum values of $CC_{\rm all}$ and $CC_{\rm weak}$ and the density of trials that were outside the main group. In *SHELXD* tests with the 18-crystal data set, the best trial solutions had roughly equivalent maximum values for both $CC_{\rm all}$ and $CC_{\rm weak}$. In contrast, the density of trials outside the main grouping (the largest values of $CC_{\rm all}$ and $CC_{\rm weak}$) was highly dependent on the resolution cutoff, with the best results using a 5.2 Å cutoff roughly correlating with the 5.4 Å limit of the anomalous signal estimated by *XSCALE*. The trial solution used for further phasing correctly identified all 12 disulfides and seven out of ten methionine sulfurs (Figs. 2*a* and 2*b*). The methionine sulfurs not found by *SHELXD* have higher B factors than the other

Table 1
Data processing and refinement.

Values in parentheses are for the outermost shell.

	18 crystals combined		'Best' crystal
	I^+, I^- separate	I^+, I^- combined	I^+, I^- separate
Data			
Space group	*P*321		
Wavelength (Å)	1.7462		
$d_{\rm min}$ (Å)	2.90 (2.96–2.90)		
Unit-cell parameters (Å)	$a = b = 167.80$, $c = 93.82$		$a = b = 167.64$, $c = 93.86$
Observations	6267610 (198290)	6266198 (198299)	362149 (13317)
Unique reflections	65517 (3942)	34031 (2023)	65112 (3592)
Average multiplicity	95.7 (50.3)	184.1 (98.0)	5.6 (3.7)
$\langle I/\sigma(I)\rangle$	19.6 (0.69)	27.31 (0.96)	10.7 (0.34)
$R_{\rm merge}$ (%)	31.3 (900.0)	31.3 (903.8)	12.4 (348.0)
Completeness (%)	100.0 (99.9)	100.0 (100.0)	99.5 (92.5)
$CC_{1/2}$† (%)	100.0 (17.1)	100.0 (33.2)	99.7 (5.4)
CC*‡ (%)	100.0 (54.0)	100.0 (70.6)	99.9 (32.0)
Wilson B (Å^2)	97.3	91.4	90.8
Refinement			
No. of reflections		33928	
$R_{\rm work}/R_{\rm free}$		0.172/0.204	
$CC_{\rm work}/CC_{\rm free}$ (%)		88.7/90.8	
R.m.s.d., bonds (Å)		0.006	
R.m.s.d., angles (°)		1.012	
No. of atoms			
Protein		5493	
Solvent		61	
Sugar/detergent		229	
B factors (Å^2)			
Protein		99.4	
Solvent		84.0	
Sugar/detergent		147.0	
Ramachandran			
Favored (%)		94.46	
Allowed (%)		5.10	
Outliers (%)		0.44	
PDB code		4tpl	

† $CC_{1/2}$ is the correlation of one-half of the observations with the other half. ‡ $CC^* = [2CC_{1/2}/(1 + CC_{1/2})]^{1/2}$.

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: DZ5338).

sulfurs in the refined model and yielded weak peaks in an anomalous difference map calculated with density-modified experimental phases (Fig. 2*b*).

2.4. Phase improvement and initial model building

Initial phasing from the sulfur sites was tested with data to 5.0, 4.5, 4.0 and 3.5 Å resolution, followed by phase refinement

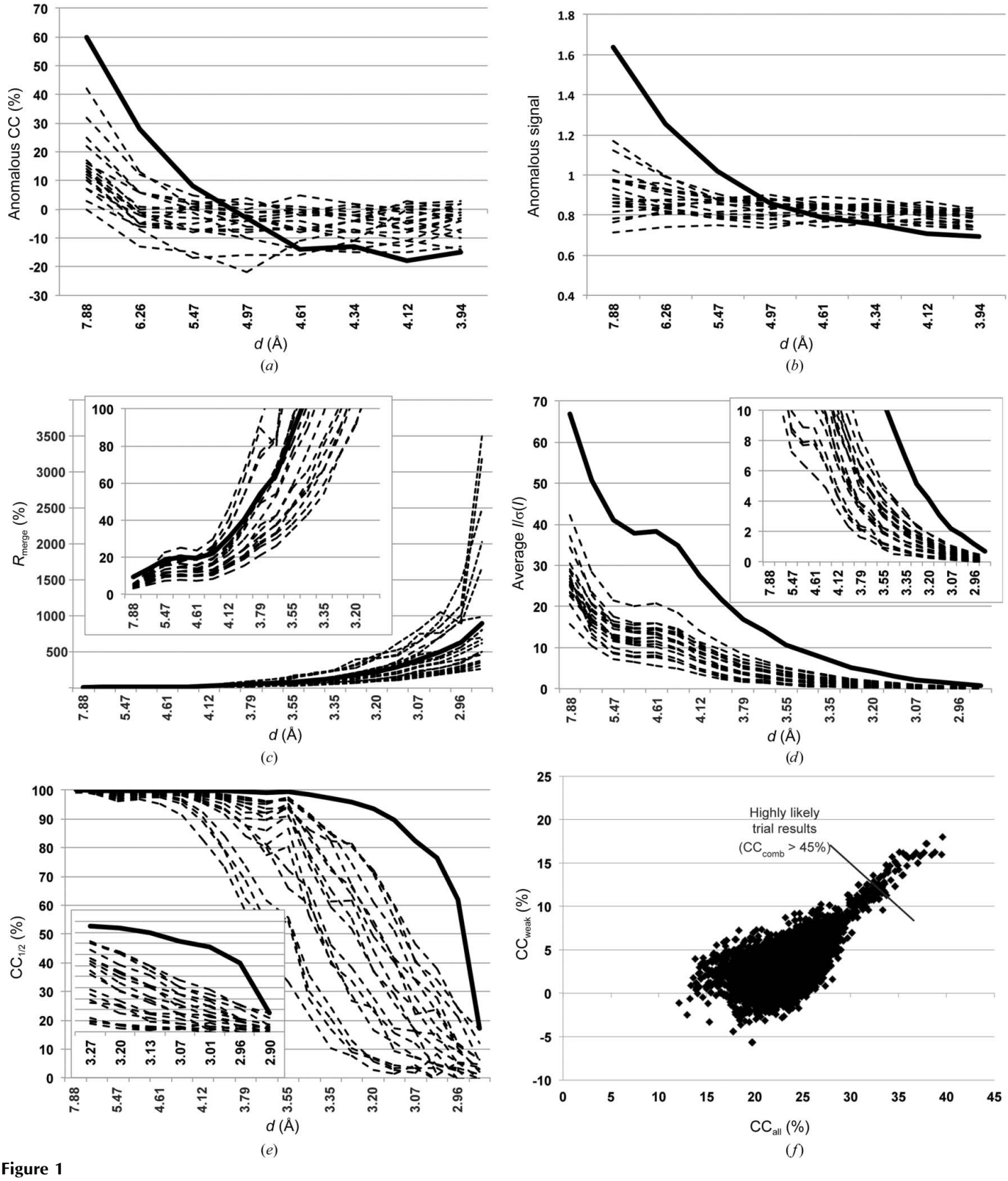

Figure 1
Anomalous signal strength (*a*, *b*) and data quality (*d*–*f*) for the 18-crystal data set used to solve the NS1 crystal structure (solid line, combined data with 100-fold multiplicity; dashed lines, individual crystal data sets; insets, high-resolution detail). (*c*) Results of a *SHELXD* substructure search with combined 18-crystal data (d_{min} = 5.2 Å, 10 000 trials).

and extension to 3.0 Å resolution. A 4.5 Å resolution limit for initial phasing was selected based on the interpretability of electron-density maps. Refinement of phases (to 4.5 Å resolution, 50 cycles) by *SHELXE* (Sheldrick, 2010) indicated a

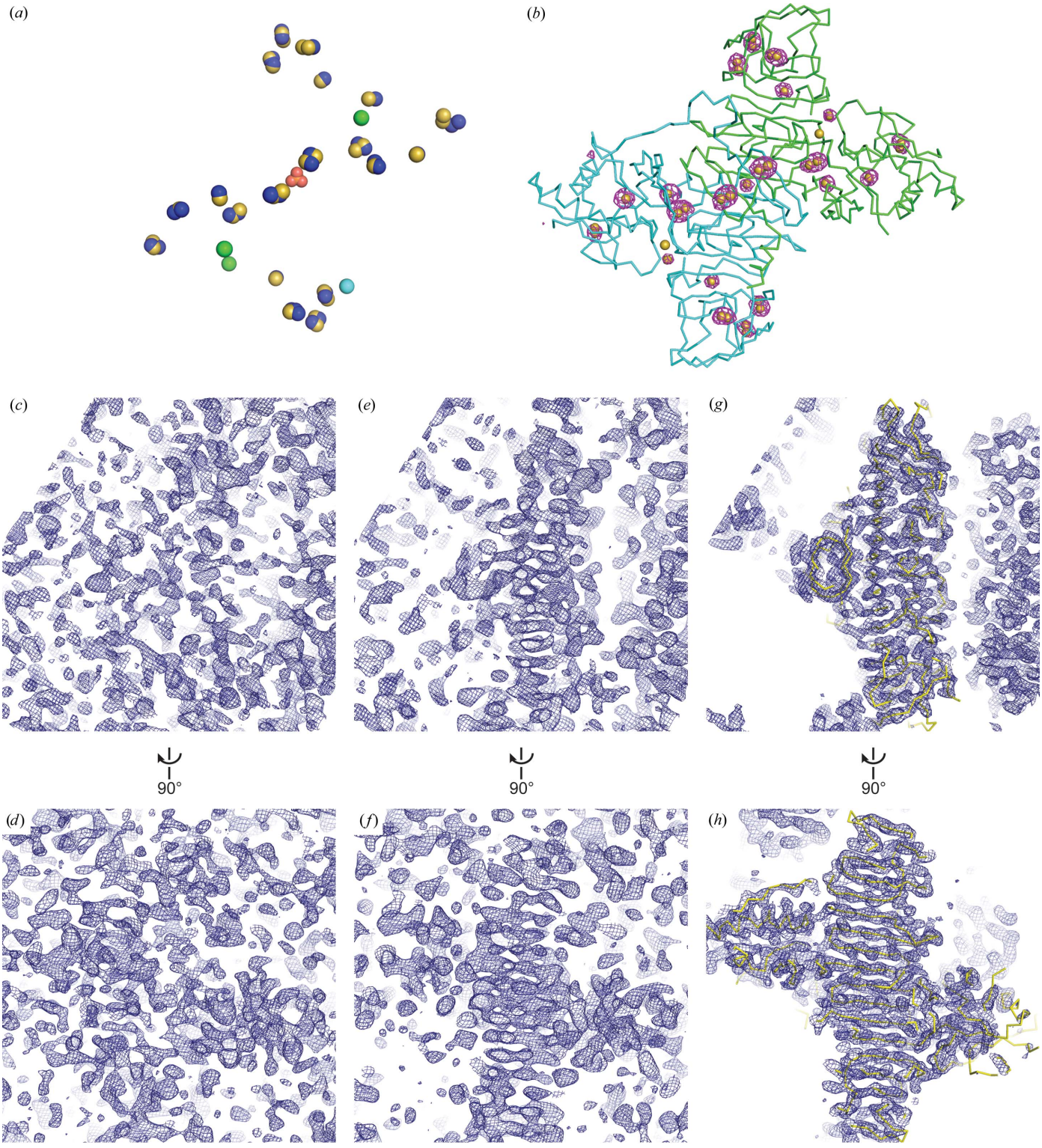

Figure 2
Sulfur sites and phasing. Sulfur atomic positions (*a*) identified by *SHELXD* (true sites, blue; false sites, cyan) and in the final refined model (yellow, found by *SHELXD*; green, missed; orange/red, sulfate on the molecular twofold) and (*b*) anomalous difference electron density (*DM* phases, 5.0σ) on a C^{α} trace of the NS1 dimer viewed along the noncrystallographic molecular twofold. (*c*–*h*) Orthogonal views of electron density at progressive stages of phase refinement. (*c*–*f*) *SHELXE* phases at 4.5 Å resolution (1.5σ contour) before (*c*, *d*) and after (*e*, *f*) 50 rounds of phase refinement. (*g*, *h*) *DM* phases at 3.0 Å resolution (1.5σ contour) after 200 rounds of phase refinement and extension with the NS1 C^{α} trace (yellow).

clear preference in hand both by visual inspection and from map statistics (*SHELXE* connectivity and contrast metrics).

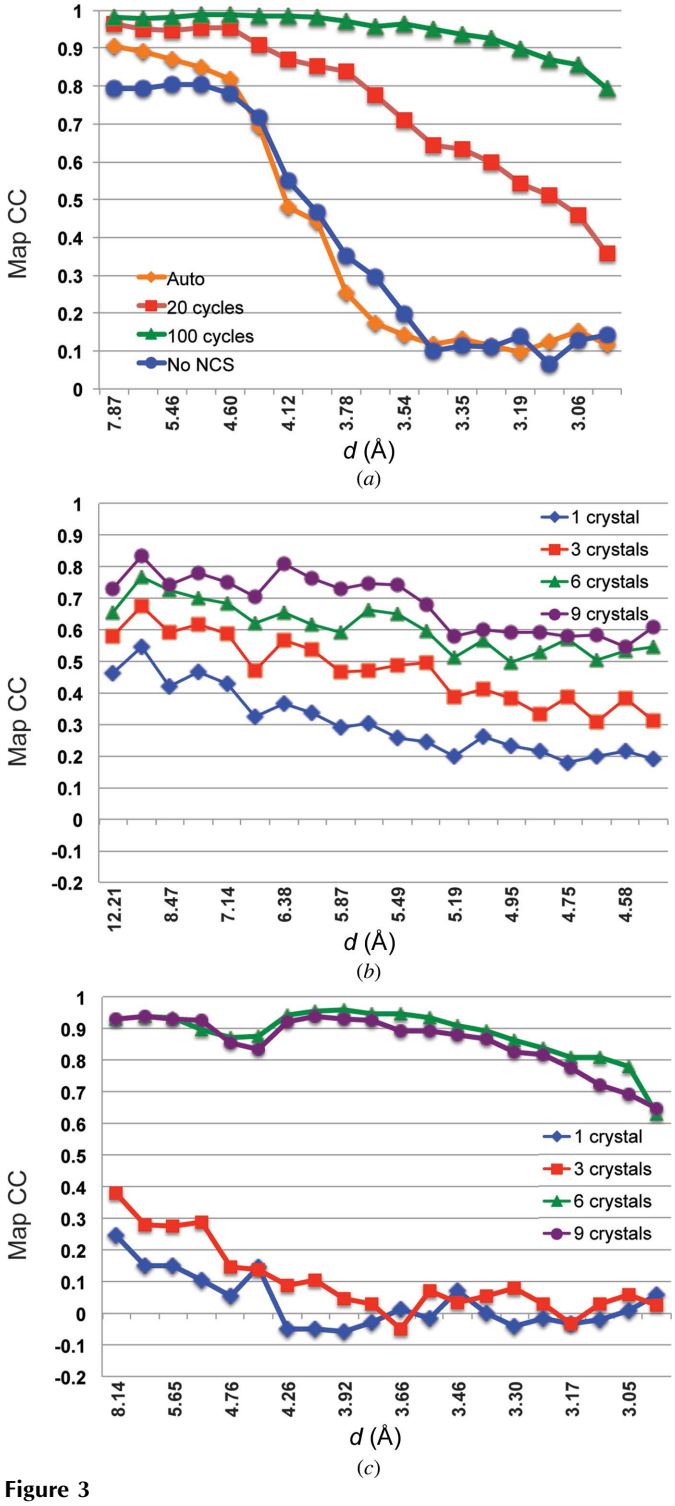

Figure 3
Phase-refinement and phase-extension experiments. (*a*) Comparison of *DM* phase-refinement and phase-extension protocols by correlation with the map calculated with a 200-cycle protocol employing 75% solvent flattening and twofold averaging. (*b*, *c*) Efficacy of phasing as a function of data multiplicity. Maps were calculated after (*b*) *SHELXE* (4.5 Å resolution, 50 cycles of phase refinement) followed by (*c*) *DM* (phase refinement and extension to 3.0 Å resolution in 200 cycles) by comparison with maps from the same stage using the combined 18-crystal data set.

A twofold local symmetry axis was readily apparent in the anomalous scatterer sites of the top *SHELXD* trial solution (Fig. 2*a*), consistent with two monomers per asymmetric unit (74% solvent). The twofold noncrystallographic symmetry (NCS) operator was calculated from the sulfur sites (*LSQKAB*; Winn *et al.*, 2011) and used for phase refinement and extension (200 cycles, from 4.5 to 3.0 Å resolution) in *DM* (Cowtan, 1994). Phase refinement dramatically improved the maps (Figs. 2*c*–2*h*) and the final 3.0 Å resolution map was of outstanding quality, sufficient for autobuilding ∼75% of the amino-acid residues with *Buccaneer* (Cowtan, 2006).

We tested the requirement of both cycle count (*DM* NCYCLE parameter) and twofold NCS averaging for the final map quality. The use of fewer *DM* cycles (20 or 100) or setting NCYCLE to 'auto' reduced the quality of the maps as judged by the combined figure of merit (FOM) and by the correlation coefficient (map CC; Fig. 3*a*). While the 20-cycle and auto NCYCLE maps were of noticeably poorer quality by visual inspection, the 100-cycle map appeared to be of equivalent quality to the 200-cycle map. Phase extension without NCS (solvent flattening only) resulted in a high estimated FOM, but the map was poorly correlated with the 200-cycle map and was of lesser quality by visual inspection. Thus, even with 74% solvent content, twofold NCS averaging was critical for successful phase extension.

With a known sulfur substructure, the phase calculation and extension process was successful with data of much lower multiplicity than the 100-fold used to determine the substructure. Single-, three-, six- and nine-crystal data sets were tested to calculate and refine phases to a 4.5 Å resolution limit in *SHELXE*. Maps at 4.5 Å resolution from the nine-crystal data were highly correlated to the 18-crystal maps, with diminishing correlation as the data multiplicity was reduced (Fig. 3*b*). Map interpretability is nearly impossible to evaluate at 4.5 Å resolution, so the phases from *SHELXE* were further refined and extended to 3.0 Å resolution by *DM*. The resulting six- and nine-crystal maps closely matched the 18-crystal map, whereas extension failed from the starting phases from single- and three-crystal data sets (Fig. 3*c*). The quality of the starting phases was more important than data multiplicity during extension. When initial 4.5 Å resolution phases were calculated and refined by *SHELXE* using the 18-crystal data set, phase extension using *DM* proceeded smoothly even with a single-crystal data set (data not shown).

2.5. Use of high-multiplicity data for extension of resolution during refinement

Model refinement was carried out with *PHENIX* (Adams *et al.*, 2010) and model building was performed with *Coot* (Emsley & Cowtan, 2004). The high-multiplicity data allowed us to significantly extend the useful resolution for crystallographic refinement beyond what would have been possible with data of ordinary multiplicity. Using traditional metrics, the data from the single best crystal would be considered to be useful to (optimistically) 3.2 Å resolution (Figs. 1*c*–1*e*). We used combined data to 2.9 Å resolution during model building

and refinement. The use of extended data was justified by visual inspection of maps calculated from density-modified experimental phases (Figs. 2*g*, 2*h* and 4*a*) or from model phases (Fig. 4*b*) as well as by refinement metrics (Figs. 4*c*–4*f*, Table 1). Refinement statistics (both traditional *R* values and correlation coefficients) indicate good agreement between the model and observed amplitudes for the full range of data. Additionally, the theoretical agreement of the merged data with the 'true' intensities, as estimated by CC* (Karplus & Diederichs, 2012; Figs. 4*c* and 4*d*, dashed red line), supports the use of the merged data to the 2.9 Å resolution limit. We compared F_{calc} from the final NS1 model with F_{obs} for each individual crystal data set. Unsurprisingly, the refined model agrees considerably better with the combined data than with any individual crystal data set (Figs. 4*c*–4*f*), even when the model was subjected to simulated annealing followed by conventional refinement with single-crystal data sets (data not shown). The magnitude of the differences between F_{calc} for the refined model and F_{obs} from individual crystals suggests that refinement against individual data sets would not be

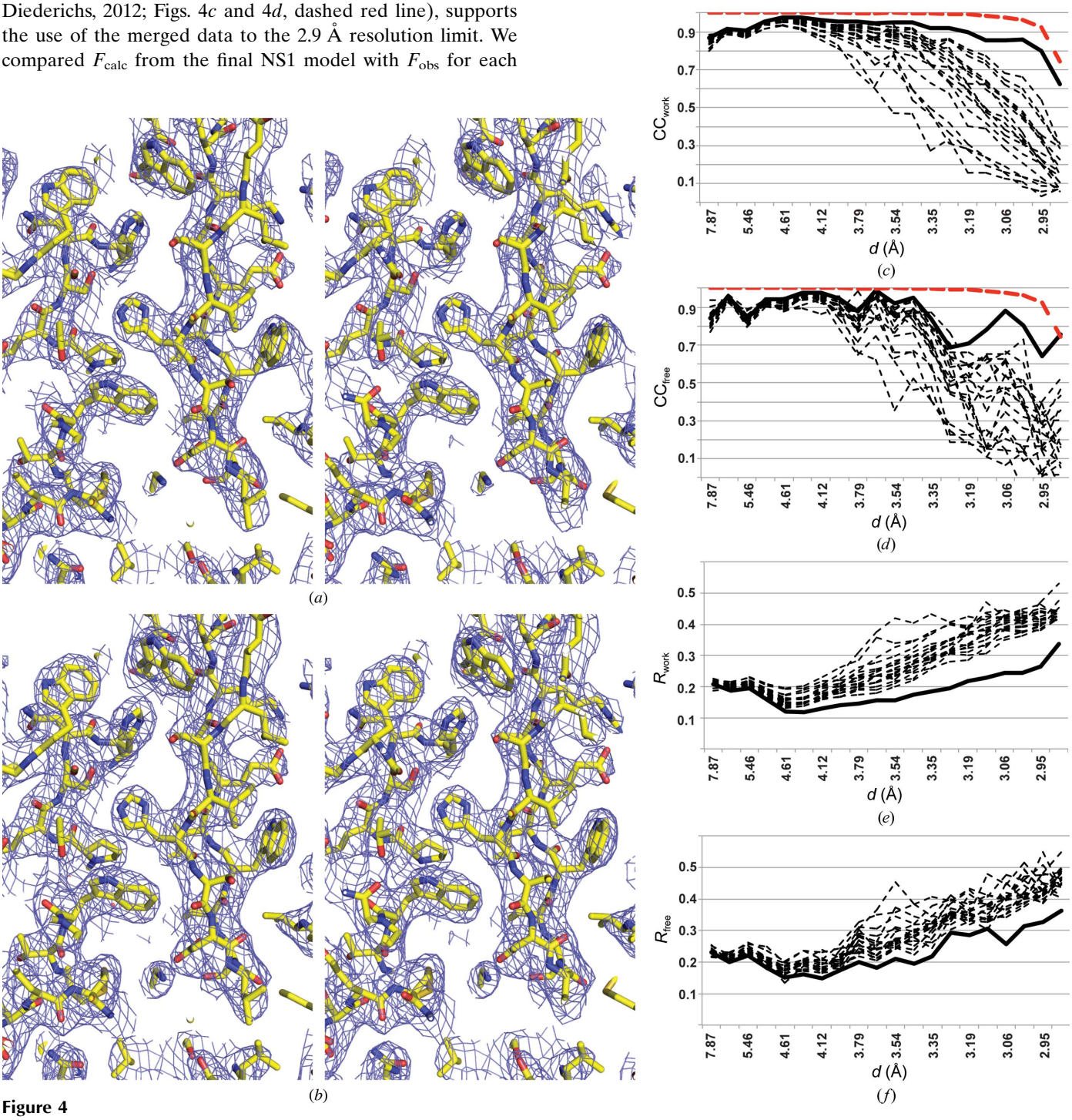

Figure 4
Crystallographic refinement using multi-crystal data. Stereodiagrams of electron density for the NS1 core from (*a*) model-free *DM*-modified phases (3.0 Å resolution, 1.5σ contour) and (*b*) model phases (*PHENIX*, $2mF_o - DF_c$, 2.9 Å resolution, 1.5σ contour). (*c*–*f*) Refined model fit to the 18-crystal data (bold lines) and individual crystal data (dashed lines). CC* is shown as a dashed red line.

productive to the 2.9 Å resolution limit. The NS1 structure is complete with the exception of one internal 20-residue disordered loop. Five of the six glycosylated asparagines are well ordered and one to five sugar residues were built at these sites. Additionally, electron density consistent with the head groups of two detergent molecules (Triton, from the purification buffers) was clear in the experimental 3.0 Å resolution maps from *DM*. Fragments of additional detergent molecules were built on a hydrophobic surface of NS1. A sulfate ion on the twofold NCS axis was identified in the anomalous difference electron density (Figs. 2*a* and 2*b*).

2.6. Utility of data 'culling' prior to merging : are more data always better?

In this *post mortem* analysis, we set out to answer two related questions. Can we use lower-multiplicity data to solve the sulfur substructure? How do we choose the optimal subset of data for determining the sulfur substructure? The initial step was to re-examine the compatibility of the data from 28 crystals. We reprocessed the data from the ten crystals not included in the 18-crystal data set used to solve the structure. Three had simple indexing inconsistencies that were readily corrected. Others appeared to be cases of varying degrees of

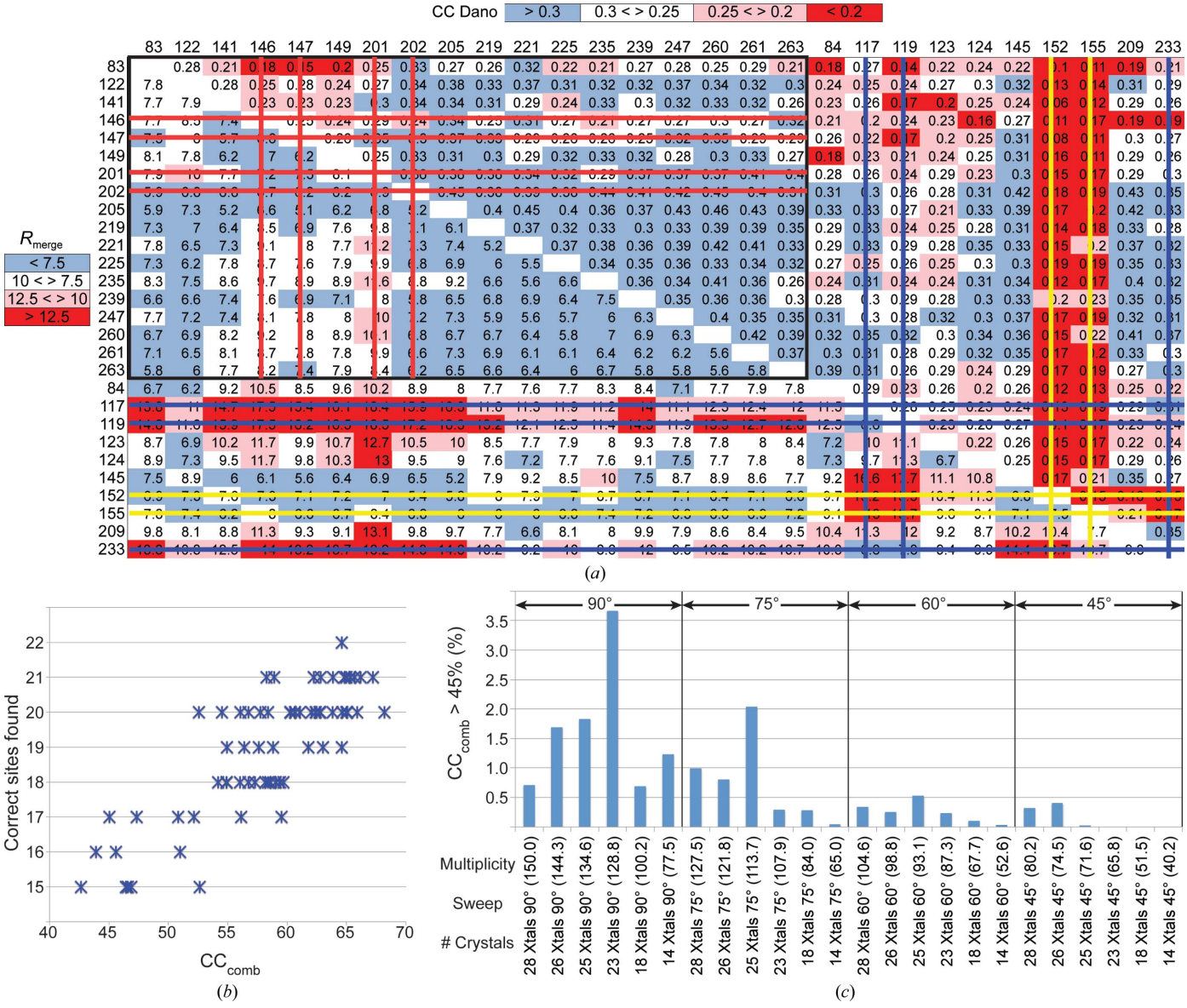

Figure 5
Data-culling experiments. (*a*) Pairwise comparison of low-resolution data (d_{min} = 7.88 Å) from 28 crystals. The 18 crystals used to solve the NS1 structure are grouped at the upper left (black box); crystal numbers label columns and rows. (*b*) Number of true sulfur sites (23 possible) as a function of CC_{comb} for the 'most likely' trial solution in each of 64 *SHELXD* tests. (*c*) Percentage of 'highly likely' trial solutions from *SHELXD* tests with various data combinations. 28 Xtals, all crystals; 26 Xtals, excluding crystals with low CC Dano [yellow strikethrough in (*a*)]; 25 Xtals, excluding crystals with high R_{merge} (blue strikethrough); 23 Xtals, excluding crystals with low CC Dano or high R_{merge} (yellow and blue strikethrough); 18 Xtals, original 18 crystals (black box); 14 Xtals, original 18 crystals excluding four crystals as described in the text (red strikethrough). The half-angular swaths of data (rotation per crystal) and multiplicity (50.0–3.5 Å, in parentheses) are indicated for each data combination. All data sets also included the true Friedel images for each of the angular widths indicated (2 × 90° *etc.*).

non-isomorphism. The compatibility of the data from the 28 crystals was evaluated with the modified criteria. As the goal was to use anomalous differences to identify native anomalous atom substructure, we again focused on the correlation of anomalous signal and on isomorphism, limiting the analysis to low-resolution data. Whereas previously individual observations from different crystals were scaled together and merged and the *XDS*/*XSCALE* AnomCC was used as a metric, here the anomalous differences for each sample (measured using an inverse-beam protocol) were preserved by scaling and merging data from individual crystals prior to pairwise comparison. The correlation of anomalous differences between data sets from crystal pairs, based on merged data from each crystal (CC Dano, calculated by *SFTOOLS*; Winn *et al.*, 2011), were all positive and ranged from 0.06 to 0.47. To assess isomorphism, unmerged data from crystal pairs were merged and scaled (*XSCALE*) and the $R_{\rm merge}$ value ($d_{\rm min}$ = 7.88 Å) for the pairs ranged from 5.1 to 18.5%. The results (Fig. 5*a*) identified several outliers as assessed by either CC Dano or $R_{\rm merge}$. The original 18 crystals used to solve the NS1 structure (black box, upper left) were reasonably compatible by both the $R_{\rm merge}$ and CC Dano metrics. Of the additional ten crystals, two (152 and 155) had significantly lower pairwise CC Dano values and three (117, 119 and 233) had significant non-isomorphism as judged by $R_{\rm merge}$ values with other data sets.

Various combinations of the 28 data sets were scaled, merged and tested in *SHELXD* (50.0–5.2 Å resolution data). Each *SHELXD* test employed 10 000 trials to search for 22 S sites, of which 12 were specified as disulfides. Substructure determination was more robust when the data were scaled to 3.5 Å resolution [approximately the limit where $I/\sigma(I) > 1$]; thus, we limited the scaling in *XSCALE* to 3.5 Å resolution for the purposes of locating S sites and used the separately scaled 2.9 Å resolution data for phase extension and density modification. As the results from the *SHELXD* tests did not have a clear delineation between random and 'correct' trial solutions (Fig. 1*f*), we used the sum of $CC_{\rm all}$ and $CC_{\rm weak}$ (here called $CC_{\rm comb}$) as a single metric to compare trial solutions. The top trial solution from any given *SHELXD* test was evaluated by the number of correctly identified S sites (within 3 Å of a S atom); each *SHELXD* trial solution yielded 31 sites (1.4 × 22 requested). Among the 64 *SHELXD* tests, $CC_{\rm comb}$ for the best trial solution of the 10 000 was highly correlated with the number of correct S sites (Fig. 5*b*). We then used the best trial solution (31 sites) from each *SHELXD* test to calculate phases to 4.5 Å resolution and to refine and extend phases to 3.0 Å resolution using the correct twofold NCS operator. All of the resulting 3.0 Å resolution maps were interpretable. Based on these results, we used a $CC_{\rm comb}$ cutoff of 0.45 to identify 'highly likely' trial solutions.

Use of a data set from all 28 crystals in *SHELXD* resulted in an equivalent number of 'highly likely' trial solutions as with the original 18-crystal data set (Fig. 5*c*; 28 crystals *versus* 18 crystals). When we excluded the crystals that appeared to be outliers based on CC Dano (152 and 155; Fig. 5*b*; 26 crystals) or on $R_{\rm merge}$ (117, 119 and 233; Fig. 5*c*; 25 crystals), the fraction of 'highly likely' trials increased significantly. (Crystals 152 and 155 are the only samples for which only 2 × 45° of data were collected, and $R_{\rm merge}$ may be artificially suppressed for comparisons involving these crystals.) Excluding the data from all five of these suspect crystals (Fig. 5*b*; 23 crystals) had an additive effect, with over fivefold more 'highly likely' trials than for the data from all 28 crystals. We also conducted a series of trials in which data from individual crystals were excluded from the original 18-crystal data set. Of these, data from any of four crystals (146, 147, 201 and 202) could be excluded with either neutral or beneficial results (data not shown). Exclusion of all four of these data sets (Fig. 5*c*; 14 crystals) yielded a data set with half the multiplicity of the 28-crystal data set but with slightly more 'highly likely' solutions.

In addition to non-isomorphism, radiation damage can degrade the measured anomalous signal. For the crystal combinations above, we investigated this effect in *SHELXD* tests using the first 2 × 75°, 2 × 60° and 2 × 45° of data, which preserved Friedel pairs from the inverse-beam geometry of data collection. There was a clear trend that excluding the later images had a detrimental effect on the number of 'highly likely' trial solutions from *SHELXD* (Fig. 5*c*). In data combinations that yielded roughly equal overall multiplicity, this trend was still evident. For combinations with a multiplicity of ∼100 (18 crystals 90°, 23 crystals 75°, 28 crystals 60°, 26 crystals 60°, 25 crystals 60°) or a multiplicity of ∼80 (14 crystals 90°, 18 crystals 75°, 23 crystals 60°, 28 crystals 45°, 26 crystals 45°), the full 2 × 90° of data per crystal from fewer crystals resulted in substantially more 'highly likely' solutions than did reduced angular swaths of data from each of a greater number of crystals. We note that the multiplicity of Bijvoet pairs per crystal (up to sixfold in the case of 2 × 90° of data) would be substantially reduced by reducing the angular range. This suggests that in the case of NS1 the multiplicity of anomalous pairs within data from each crystal was more important to the integrity of multi-crystal anomalous pair data than was minimizing radiation damage.

3. Discussion

To accurately measure the sulfur anomalous signal from NS1 crystals, several approaches were used during data collection and subsequently during assembly of the meta data set. For data collection, we attempted to maximize the accuracy of anomalous measurements through the use of an inverse-beam protocol, the inclusion of helium in the diffracted beam path and the use of alternating starting positions to collect different unique Friedel pairs early in dose-decay time. Ultrahigh-multiplicity data led to a successful solution of the substructure of sulfur anomalous scatterers with data limited to a $d_{\rm min}$ of 5.2 Å, roughly corresponding to the limit of positive values for AnomCC and an anomalous signal greater than 1.0. An atomic model cannot be built into a map at this resolution, particularly as the protein fold was unknown. Phase calculation to 4.5 Å resolution and extension to 3.0 Å resolution resulted in an easily interpreted map of exceptional quality

(Figs. 2*g* and 2*h*) when we employed a protocol that included solvent flattening, twofold NCS averaging and at least 100 cycles during phase extension (Fig. 3*a*).

In *post mortem* experiments, we sought to determine whether the 100-fold multiplicity was essential to successful substructure determination. Multiplicities of 50-fold to 150-fold were examined in several dozen tests using *SHELXD* with 10 000 trial solutions per test. The CC_{comb} statistic (the sum of the *SHELXD* CC_{all} and CC_{weak}) was strongly correlated with the number of correct sulfur sites in any trial solution (Fig. 5*b*). This implies that the maximum CC_{comb} value from a set of *SHELXD* tests should indicate the optimal selection of resolution, crystal subset or any other variable examined. Among the 10 000 trials in each *SHELXD* test, the 'most likely' trial solution yielded two-thirds or more of the correct sulfur sites (at least 15 of 22 sites; Fig. 5*b*). However, these 'most likely' trial solutions were polluted with up to 16 false sites, as trial solutions typically yielded 40% more sites than requested. In most of the trial solutions that were analyzed in detail, the sites with highest occupancy were valid. However, occupancy diminished gradually for the remaining sites and there was no clear demarcation to distinguish the valid sites from the false sites. Between three and 15 (with an average of seven) false sites had higher occupancies than the lowest-occupancy true site. In the general case, contamination owing to false sites complicates the determination of NCS operators from sulfur atomic positions. Nevertheless, for all trial solutions in which CC_{comb} was greater than 45% the phase-extension and refinement protocol from the 31 sites (valid and false) produced an interpretable 3.0 Å resolution map if the correct twofold NCS operator was applied. As optimal phase extension required proper identification of the twofold NCS operator, when searching for the NCS operator it may be useful to pare the lower occupancy sites, which are more likely be incorrect. We did not investigate the possibility that solvent flattening alone may have improved the phases sufficiently to yield a map that revealed the NCS operator (or the false sites).

The NS1 structure determination benefited from an ability to collect complete anomalous pair data from almost every sample, as the crystals in space group *P*321 were sufficiently robust for the measurement of $2 \times 90^\circ$ of true Friedel data collected in 5° wedges with inverse geometry. We anticipated that *SHELXD* results of similar quality could be obtained by excluding the most radiation-damaged images for each crystal while preserving true Friedel pairs for the included data. However, the number of 'highly likely' trial solutions from *SHELXD* dropped precipitously as the data range was reduced to $2 \times 75^\circ$ and below (Fig. 5*c*), corresponding to a drop in anomalous pair multiplicity for individual crystals from six to four or fewer. This somewhat surprising result indicates that the multiplicity of anomalous pairs within the data from each crystal may have been important for the estimation of the true anomalous signal for weakly diffracting NS1 crystals. We note that the *XDS* data processing and scaling took no account of the experimental design to record true Friedel pairs in inverse geometry and that other scaling protocols may diminish the need for high anomalous multiplicity in data from each crystal.

The multiplicity required to solve the sulfur substructure was considerably greater than that needed for initial phasing and for phase extension. Data with approximately 30-fold multiplicity were sufficient for initial phase calculation to 4.5 Å resolution (*SHELXE*) and extension to 3.0 Å resolution using the known sulfur substructure (Fig. 3*c*). However, phases to 4.5 Å resolution calculated from data of 15-fold multiplicity failed in the phase-extension protocol and did not produce an interpretable 3.0 Å resolution map. In contrast, if initial phases to 4.5 Å resolution were calculated from data of at least 30-fold multiplicity, phase extension to 3.0 Å resolution was successful using data from only one crystal (12-fold multiplicity). Thus, the path of NS1 structure determination had diminishing demands on data multiplicity: 50-fold to solve the sulfur substructure at 5.2 Å resolution, 30-fold to calculate initial phases at 4.5 Å resoution and 12-fold to extend the phases to 3.0 Å resolution.

When compiling extremely high multiplicity data it is beneficial to determine which samples to include in the final data set by identifying isomorphous, or conversely 'outlier', crystals. In the case of NS1, the exclusion of non-isomorphous outlier crystals increased the apparent success rate of sulfur substructure determination. We used pairwise comparisons of data sets, a strategy that is justified when data are complete for each sample. Unit-cell parameter differences are a rapid way to detect non-isomorphism (Foadi *et al.*, 2013), but equivalence of unit-cell parameters does not guarantee isomorphism. When complete data are not available from single samples, it may be necessary to first compile a reference data set with which all partial data sets will be compared (Liu *et al.*, 2012). Multiple metrics can be considered when determining which data to include. For solving the sulfur substructure at low resolution, we chose low-resolution anomalous correlation coefficients and R_{merge} values. Conclusive identification of the 'best' data set would likely require hierarchical cluster analysis (Giordano *et al.*, 2012), which can be implemented in data-processing pipelines.

In addition to the utility for phasing from a weak anomalous signal, ultrahigh multiplicity was critical for extending the resolution limit of weak data. Whereas traditional metrics would indicate that the data from the single best crystal were useful to ~3.2 Å resolution [average $I/\sigma(I) > 2$], the combined data were useful to 2.9 Å resolution, resulting in a one-third increase in the number of unique reflections for refinement. The utility of the combined data is apparent both in the quality of the electron-density maps (Figs. 4*a* and 4*b*) and in the agreement of calculated and observed amplitudes (Fig. 4*c*–4*f*). The use of additional data from 3.2 to 2.9 Å resolution is particularly important for NS1, where the protein fold was unknown and novel. Karplus and Diederichs showed that for problems with data limits beyond 2.0 Å resolution the correlation coefficients of random half data sets ($CC_{1/2}$ and CC*) are better indicators of the useful data limit than are the traditional metrics R_{merge} and average $I/\sigma(I)$ (Diederichs & Karplus, 2013; Karplus & Diederichs, 2012). The NS1 case

shows that this is also true for problems of moderate diffraction limit.

Taken together, the recent demonstrations by others of the utility of ultrahigh-multiplicity data recorded at nonheroic X-ray energies for sulfur SAD phasing (Liu *et al.*, 2012, 2013) and of new correlation-coefficient metrics to evaluate data quality (Karplus & Diederichs, 2012) were instrumental in solving and refining the crystal structure of the *West Nile virus* NS1 protein. The moderate quality of NS1 crystals indicates that these advances in crystallographic methodology have wide applicability, particularly in combination with high-brilliance X-ray sources and high-speed detectors.

This work was supported by a grant from the National Institutes of Health (P01AI055672) to JLS and the Martha L. Ludwig Professorship of Protein Structure and Function to JLS. The beamlines of GM/CA@APS were supported by the National Institute of General Medical Sciences ('GM', Y1-GM-1104) and the National Cancer Institute ('CA', Y1-CO-1020).

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Akey, D. L., Brown, W. C., Dutta, S., Konwerski, J., Jose, J., Jurkiw, T. J., DelProposto, J., Ogata, C. M., Skiniotis, G., Kuhn, R. J. & Smith, J. L. (2014). *Science*, **343**, 881–885.
Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.
Cherezov, V., Hanson, M. A., Griffith, M. T., Hilgart, M. C., Sanishvili, R., Nagarajan, V., Stepanov, S., Fischetti, R. F., Kuhn, P. & Stevens, R. C. (2009). *J. R. Soc. Interface*, **6**, S587–S597.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.
Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.
Diederichs, K. & Karplus, P. A. (2013). *Acta Cryst.* D**69**, 1215–1222.
Doutch, J., Hough, M. A., Hasnain, S. S. & Strange, R. W. (2012). *J. Synchrotron Rad.* **19**, 19–29.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Cryst.* D**69**, 1617–1632.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Cryst.* D**68**, 649–658.
Hendrickson, W. A., Smith, J. L. & Sherrif, S. (1985). *Methods Enzymol.* **115**, 41–55.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 133–144.
Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Lehmann, M. S., Müller, H.-H. & Stuhrmann, H. B. (1993). *Acta Cryst.* D**49**, 308–310.
Liu, Q., Dahmane, T., Zhang, Z., Assur, Z., Brasch, J., Shapiro, L., Mancia, F. & Hendrickson, W. A. (2012). *Science*, **336**, 1033–1037.
Liu, Q., Liu, Q. & Hendrickson, W. A. (2013). *Acta Cryst.* D**69**, 1314–1332.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Muller, D. A. & Young, P. R. (2013). *Antiviral Res.* **98**, 192–208.
Ramagopal, U. A., Dauter, M. & Dauter, Z. (2003). *Acta Cryst.* D**59**, 1020–1027.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Smith, J. L., Fischetti, R. F. & Yamamoto, M. (2012). *Curr. Opin. Struct. Biol.* **22**, 602–612.
Smith, J. L. & Hendrickson, W. A. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 299–303. Dordecht: Kluwer Academic Publishers.
Wagner, A., Pieren, M., Schulze-Briese, C., Ballmer-Hofer, K. & Prota, A. E. (2006). *Acta Cryst.* D**62**, 1430–1434.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.