Macromolecular X-ray structure determination using weak, single-wavelength anomalous data

Gábor Bunkóczi1, Airlie J McCoy1, Nathaniel Echols2, Ralf W Grosse-Kunstleve2, Paul D Adams2, James M Holton2,3, Randy J Read1 & Thomas C Terwilliger4

We describe a likelihood-based method for determining the substructure of anomalously scattering atoms in macromolecular crystals that allows successful structure determination by single-wavelength anomalous diffraction (SAD) X-ray analysis with weak anomalous signal. With the use of partial models and electron density maps in searches for anomalously scattering atoms, testing of alternative values of parameters and parallelized automated model-building, this method has the potential to extend the applicability of the SAD method in challenging cases.

SAD phasing is the dominant X-ray crystallographic method for determination of macromolecular structures by experimental phasing, accounting for 73% of such structures deposited in the Protein Data Bank (PDB; http://www.pdb.org/)1 in 2013. In the SAD method, the X-ray diffraction from anomalously scattering atoms in a molecule provides X-ray phase information for the entire crystal structure2,3. The anomalous differences between X-ray amplitudes for ‘Bijvoet pairs’ of reflections related by inversion are used first to find the positions of the anomalously scattering atoms, known as the substructure, that are consistent with these differences4,5. The substructure is then used along with the X-ray data (including Bijvoet pairs) to estimate phases for the entire structure and to calculate an electron density map2,6–8. The phases can then be improved in a third step by an iterative process of phase improvement, model-building and refinement9, which can often yield an accurate electron density and a relatively complete model.

The SAD phasing method can be challenging if the anomalous signal-to-noise ratio is low10,11. The magnitudes of the anomalous differences between Bijvoet pairs of reflections depend on the types and numbers of anomalously scattering atoms in the structure and the wavelength of data collection, and the accuracies of the anomalous differences depend on the details of data collection, particularly the number of times each X-ray reflection is measured. In experiments using small numbers of selenium atoms in SAD phasing of large structures, in sulfur SAD phasing and in experiments using X-ray free-electron lasers with SAD phasing, exceptional efforts may be necessary to obtain sufficient signal-to-noise ratios12,13.

A step that can be particularly difficult when the signal-to-noise ratio is low is identifying the positions of the anomalously scattering atoms in a structure. The substructure is often determined with ‘dual-space’ algorithms that use the anomalous differences and alternate between real-space peak-picking and reciprocal-space direct-methods phase improvement4,5. Possible substructures generated by dual-space algorithms are scored according to agreement between the structure factors calculated from the substructure model and the measured anomalous differences.

Here we introduce the use of a likelihood function to find the substructure7. The SAD likelihood function describes the probability of measuring the observed data given a model of the substructure and can be used to rank possible substructures. Likelihood functions have been used for some time for estimation of crystallographic phases using the anomalous data and the substructure7,8 and for finding missing sites in a nearly complete substructure6,14,15. After finding candidate partial substructures from the anomalous difference Patterson function, we use likelihood-based maps to complete the substructure and the SAD likelihood function to evaluate potential solutions.

We compared the dual-space completion and correlation-based scoring method of finding the anomalously scattering substructure with the log-likelihood gradient map (LLG) completion and SAD likelihood-based approach (Fig. 1). We took data sets with known structures from the PDB1. The 162 data sets (Online Methods and Supplementary Data 1) included anomalous signal from selenium, iodine, mercury, iron and zinc, contained from 1 to 74 anomalously scattering atoms in the asymmetric unit, and had high-resolution limits from 1 Å to 3.3 Å. Most of these SAD data sets were taken from multiwavelength experiments and include not just the peak wavelength with maximal anomalous signal but also weaker remote and inflection data that were not previously used alone to determine structures. We evaluated the relative strengths of the anomalous signal in the various data sets by calculating model-phased anomalous difference Fourier maps and noting the density in these maps at positions of atoms in the known substructure. An anomalous difference Fourier map based on accurate anomalous differences should have high peaks at these atomic positions. For this comparison, we implemented each approach with the same software (HySS16 in the Phenix software suite17)
We determined the fraction of sites identified correctly for each data set using dual-space completion and correlation-based scoring (Fig. 1a). For data sets with anomalous signal below 7.5, none of the substructures could be determined (with at least 50% of the sites found); for those above 7.5, 71% could be determined. We carried out the same analysis using LLG completion and likelihood-based scoring of solutions (Fig. 1b). With the likelihood-based approach, nearly all (96%) of the substructures for data sets with signal over 7.5, along with some of the substructures (4%) for data sets with signal below 7.5, could be determined. This difference is substantial because it means that 37 more of the 162 substructures could be determined by LLG completion and likelihood-based scoring than by traditional dual-space completion and structure factor correlation scoring.

We carried out a second test of the likelihood-based methods for substructure determination using sulfur-SAD data sets that had been collected to test multicrystal approaches for cases with a weak anomalous signal. In that work, sulfur-SAD anomalous data were collected from seven crystals of the membrane protein CysZ (PDB entry 3TX3) at a wavelength of 1.7432 Å to a resolution of 2.3 Å and were merged to form a composite data set. The anomalous substructure could be determined with the seven-crystal data set and with at least some combinations of data sets assembled from three or more crystals. We created a set of 28 merged data sets using from 1 to 7 crystals and tested likelihood-based substructure determination using each original or merged data set (Supplementary Data 2). To check whether the likelihood-based methods were comparable to implementations of dual-space methods in other software packages, we carried out Shelxc/d dual-space substructure determination with the same data sets.

We examined the number of sites correctly identified as a function of the anomalous signal in these data sets (Fig. 2a). In our tests, dual-space substructure determination succeeded in at least some cases for merged data sets with an anomalous signal of about 8.4 or greater, whereas likelihood-based determination succeeded in cases with an anomalous signal as low as 7.4. We plotted the same data as a function of the number of crystals used in each merged data set (Fig. 2b). The likelihood-based approach correctly identified the sulfur substructure in four of the eight two-crystal data sets examined. Beginning with a two-crystal merged data set and sulfur substructure, the sequence of the protein, the number of sulfur atoms and the wavelength of X-ray data collection, our automated structure determination algorithm yielded a high-quality electron density map (Fig. 2c). The resulting model produced a free R value of 0.26, where a value of 0 is perfect and the deposited structure (3TX3) has a value of 0.24. The model had 435 of 453 residues assigned to sequence (the polypeptide backbone for 18 residues was not automatically associated with a particular segment of the amino acid sequence).

**Figure 1** Comparison of substructure completion algorithms. (a) Dual-space substructure completion. Fraction of sites correctly identified using the dual-space algorithm implemented in HySS within the Phenix software system, plotted as a function of the anomalous signal in each SAD data set. Anomalous signal is the mean peak height of a normalized anomalous difference Fourier map, phased using the deposited model or a refined model based on the deposited model (for data sets where the deposited model did not correspond to the anomalous data set), at the coordinates of the atoms in the anomalous substructure (see main text). Substructure searches were carried out with default parameters and include trials at varying resolutions. (b) Likelihood-based substructure determination as in a, except that the scoring and substructure completion was carried out using the SAD likelihood function.

**Figure 2** Application of substructure completion algorithms to CysZ data sets merged from varying numbers of crystals. (a) Dual-space substructure determination with Shelxc/d.18 and brute-force likelihood-based completion are shown. A cutoff of 3.5 Å was used throughout. We also tested a cutoff of 2.8 Å for Shelc/d and obtained similar results (at least 8 of 21 sites were found for six data sets using a cutoff of 2.8 Å and for seven data sets using a cutoff of 3.5 Å). Correct sites found (out of a possible 21) are shown as a function of the anomalous signal in the merged data sets (mean peak height at positions of atoms in the known substructure in model-phased anomalous difference Fourier map). (b) Substructure determination in a, but showing sites found as a function of the number of crystals included in merging. The values for numbers of crystals are slightly offset so that multiple values can be seen. (c) Model and density-modified electron density map obtained by default application of Phenix structure determination algorithms beginning with the sites marked with the arrow in a (merged data from crystals 2 and 6 of Liu et al.18).
Comparisons with algorithms implemented by other groups are difficult to carry out without bias. The developers of an algorithm normally use their software more adeptly than they can that of others; and comparison software is normally static, whereas the software being developed may be optimized using the test data. In this CysZ comparison, we attempted to reduce the expertise effect by using very extensive searches with Shelxc/d. The Phenix software does, however, have the advantage of having been developed, and choices of strategies and default parameter values made, in the presence of this data and of all the other data used in this work. We tested whether this use of the data in development affected the results of this test by reanalyzing all the data sets (Fig. 2a) with Phenix code developed before any use with this data. This analysis yielded numbers of correct sites very similar to those for the fully developed Phenix version (maximum difference of two sites, mean difference of less than 0.1 sites; Online Methods and Supplementary Data 2). Overall we found that likelihood-based methods can be exceptionally powerful for substructure solution in a challenging case such as this CysZ membrane-protein structure.

Once the substructure is identified, it is used along with the original data to estimate crystallographic phases.6–8 This is typically followed by iterative phase improvement, model-building and refinement.9 Our approach for phase improvement once the substructure is determined has four key features. These are the use of statistical density modification19 for integrating information from density modification with phase information from the anomalous differences; optimization of parameters during the structure determination process; iteration14,15 of the process of identifying the positions of anomalously scattering atoms, calculating phases and density modification; and parallel automated model-building.

We applied likelihood-based substructure determination and our enhanced phase improvement approaches to 159 SAD data sets (Supplementary Data 3). We evaluated our methods by using the same software (Phenix17) without and with the use of the new approaches presented in this work. We examined the correlation between final electron density maps produced by previously available algorithms in Phenix and the model map (calculated from final deposited structures) for these data sets (Fig. 3a). The map correlation, a standard metric of the quality of the structure determination process20, is plotted as a function of the anomalous signal in the data. The higher the map correlation, the more closely the map corresponds to the final structure. We define a structure to be ‘solved’ here if the map correlation is 0.50 or greater, though lower correlations indicate some degree of correctness of the electron density map.

Without including the improved algorithms described here, 50% of the data sets with anomalous signal in the range of 8–30 could be solved (map correlations with the model-phased map of 0.50 or greater; Fig. 3a). Applying the current algorithms allows solution of 79% of these data sets (Fig. 3b). To place the capabilities of the Phenix algorithms in the context of other available software, and noting that a comparison was made recently between the Crank2 software and earlier Phenix algorithms using many of the same data sets, we carried out a comparison of our current Phenix with the Crank2 software.11 We compared map quality obtained using the current algorithms as implemented in Phenix on the basis of 73 of the most challenging data sets (from Fig. 3a, high-resolution limits ranging from 1.3 to 3.0 Å) with those obtained with the recently improved algorithms in the Crank2 software (Fig. 3c). To focus on the structure determination algorithms, we started each analysis with the substructures determined by Phenix. We found that 6 out of 73 of these difficult SAD data sets could not be solved by Crank2 but could be analyzed with Phenix to obtain maps highly correlated with those calculated from the deposited structure (Fig. 3c). We also carried out a comparison of Phenix and Crank2, performing with each the entire process of finding the substructure through phase improvement and model-building. Eight of these data sets could be determined by Phenix but not by Crank2 using sites determined by Shelxc/d (Supplementary Fig. 1), and 41 could be determined by Phenix but not Crank2 using sites determined by Crunch2 (Supplementary Fig. 2).

We conclude that likelihood-based determination of the anomalous substructure, combined with improvements in methodology for phase improvement, can be a powerful approach for structure determination using SAD phasing. Additional improvements in substructure determination may potentially be obtained by optimizing the likelihood scoring function and possibly also by combining the most powerful aspects of likelihood-based methods, such as scoring of partial substructures and identification
of additional sites based on a substructure, with the rapidity and extensive exploration of possible substructures possible with direct-methods approaches.

All the Phenix tools and code described here are available from the Phenix website at http://www.phenix-online.org/.

METHODS
Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank Q. Liu, W. Hendrickson and colleagues for generously providing the CysZ sulfur SAD data sets. Support received from the US National Institutes of Health (grant P01GM063210 to P.D.A., T.C.T. and R.J.R. and grant GM073210 to J.M.H.) and the Wellcome Trust (Principal Research Fellowship to R.J.R., grant 082961/Z/07/Z) is gratefully acknowledged. This work was partially supported by the US Department of Energy under contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS
R.J.R. and T.C.T. led the work; R.J.R. and A.J.M. developed the likelihood-based scoring and LLG maps in Phaser; R.W.G.-K. and P.D.A. developed the HySS framework; J.M.H. developed the synthetic data tests; N.E., T.C.T. and G.B. integrated the likelihood-based methods into HySS; T.C.T. carried out the tests and optimized the Phenix AutoSol and AutoBuild algorithms.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Berman, H.M. et al. Nucleic Acids Res. 28, 235–242 (2000).
2. Hendrickson, W.A. & Teeter, M.M. Nature 290, 107–113 (1981).
3. Wang, B.-C. Methods Enzymol. 115, 90–112 (1985).
4. Weeks, C.M., DeTitta, G.T., Miller, R. & Hauptman, H.A. Acta Crystallogr. D Biol. Crystallogr. 49, 179–181 (1993).
5. Schneider, T.R. & Sheldrick, G.M. Acta Crystallogr. D Biol. Crystallogr. 58, 1772–1779 (2002).
6. de La Fortelle, E. & Bricogne, G. Methods Enzymol. 276, 472–494 (1997).
7. McCoy, A.J., Storoni, L.C. & Read, R.J. Acta Crystallogr. D Biol. Crystallogr. 60, 1220–1228 (2004).
8. Pannu, N.S. & Read, R.J. Acta Crystallogr. D Biol. Crystallogr. 60, 22–27 (2004).
9. Langer, G., Cohen, S.X., Lamzin, V.S. & Perrakis, A. Nat. Protoc. 3, 1171–1179 (2008).
10. Liu, Q. et al. Science 336, 1033–1037 (2012).
11. Skubák, P. & Pannu, N.S. Nat. Commun. 4, 2777 (2013).
12. Weiss, M.S., Sicker, T. & Hilgenfeld, R. Structure 9, 771–777 (2001).
13. Barends, T.R. et al. Nature 505, 244–247 (2014).
14. McCoy, A.J. & Read, R.J. Acta Crystallogr. D Biol. Crystallogr. 66, 458–469 (2010).
15. Read, R.J. & McCoy, A.J. Acta Crystallogr. D Biol. Crystallogr. 67, 338–344 (2011).
16. Grosse-Kunstleve, R.W. & Adams, P.D. Acta Crystallogr. D Biol. Crystallogr. 59, 1966–1973 (2003).
17. Adams, P.D. et al. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).
18. Sheldrick, G.M. Acta Crystallogr. D Biol. Crystallogr. 66, 479–485 (2010).
19. Terwilliger, T.C. Acta Crystallogr. D Biol. Crystallogr. 56, 965–972 (2000).
20. Terwilliger, T.C. et al. Acta Crystallogr. D Biol. Crystallogr. 65, 582–601 (2009).
ONLINE METHODS

Likelihood-based substructure determination. The substructure determination procedures implemented in the Phenix hybrid substructure search (HySS)16 were modified to allow scoring based on a SAD likelihood function17 and substructure completion using log-likelihood gradient maps as described. The HySS infrastructure was further extended to allow automatic searches using data of varying resolution and parallel evaluation of substructures in which rapid dual-space algorithms are automatically alternated with slower likelihood-based algorithms, and in which the search is terminated if equivalent solutions are repeatedly found16. A typical command for mixed dual-space/LLG substructure determination is

\[
\text{phenix.hyss data.sca 21 Se wavelength=0.9792}
\]

A brute-force likelihood-based substructure completion procedure was developed that uses m (typically 100) of the top-scoring two-site trial solutions to the anomalous difference Patterson function as seeds. A log-likelihood gradient map is calculated on the basis of a trial solution, and the n (typically 30) highest peaks in the map are added two at a time to the trial solution. All the resulting four-site trial solutions are used in a step of automatic substructure and likelihood scoring, and the top p (typically five) resulting trial solutions are used in additional cycles (typically three) of completion and scoring. The top-scoring solution overall is then returned. A typical Phenix command for brute-force substructure determination is

\[
\text{phenix.hyss strategy=brute_force merge_23.sca 21 S wavelength=1.7432 \ resolution=3.5 rescore=phaser-complete nproc=6}
\]

Structure determination algorithms. The automated structure-determination procedures in the Phenix tools AutoSol20 and AutoBuild21 were extended and used in this work.

Statistical density modification is carried out as described19 and is used both in the absence and presence of a partial model of the structure. This density modification approach has the advantage, as do the approaches used in refinement in Buster,22 that it is possible to specify the regions in the crystal that contain disordered solvent, those that contain modeled structure, those that contain unmodeled structure and the distribution of as-yet-unmodeled density in each region23.

Optimization of parameters is carried out during the structure determination process. Some parameters are optimized within individual steps (many parameters are optimized in Phaser SAD phasing), and others are optimized using a scoring procedure based on the analysis of features in the resulting electron density maps. One parameter tested is the value of the smoothing radius used in identification of the solvent boundary in density modification2, scored on the basis of the agreement factor (R value) obtained from density modification20. A second parameter tested automatically is the sharpening and anisotropy-correction of the data used in the substructure search process, with decision-making based on the electron density maps obtained20. Additionally, if the figure of merit of phasing is low (typically less than 0.35), then the number of cycles of density modification is reduced to four, with one overall cycle of mask identification19.

Map-based iteration of anomalous substructure determination14,15 is carried out if the figure of merit of phasing is low (typically if less than 0.35). Model-based iteration is also carried out in this case if the model that is built is very incomplete (for example, if the R-value is higher than 0.35). The likelihood-based procedure for completion of a partial model of the substructure can take into account information on the remainder of the structure. This algorithm can use partial structure information from either density-modified electron density maps or partial models built into these electron density maps to find the locations of anomalously scattering atoms not identified in the initial stages of structure solution. These improved models for the substructure can then be used to obtain improved phases, density-modified maps and models.

Parallel automated model-building is carried out in cases where standard model-building yields a very incomplete model. It extends the use of model averaging in iterative model-building24 by carrying out an iterative model-building procedure multiple times, followed by map-averaging to improve the resulting electron density maps and choice of working models based on their agreement with the data (R values).

The use of these extensions is controlled by individual keywords or by the “thoroughness” parameter. When this parameter is set to “medium,” all the new algorithms described here except for brute-force substructure completion and parallel autobuilding are used. This is the value of “thoroughness” used in the comparisons shown in Figure 3.

A typical command used in this work for automated structure determination with phenix.autosol was

\[
\text{phenix.autosol unit_cell=65.648 70.734 93.922 90 90 90 data=w3.sca \ atom_type=se lambda=0.97936 seq_file=1vlm.fa throughness=medium}
\]

where the unit cell is provided in this case because the data file does not contain this information. This is followed by a phenix.autobuild command such as

\[
\text{phenix.autobuild data=AutoSol_run_1_/overall_best_refine_data.mtz \ seq_file=./1vlm.fa ha_file=AutoSol_run_1_/overall_best_ha_pdb.pdb \ map_file=AutoSol_run_1_/overall_best_denmod_map_coeffs.mtz \ model=AutoSol_run_1_/overall_best.pdb extreme_dm=False \ rebuild_in_place=False}
\]

Parallel model-building was carried out using the Phenix tool phenix.parallel_autobuild. This procedure consists of n (typically 8–16) parallel runs of the automated model-building, density-modification and refinement algorithm implemented in phenix.autobuild. Each run uses a different random seed, generating variation in the linkages between peptide fragments when models are built and yielding slightly or even substantially different final models. When all runs are completed, the resulting density-modified electron density maps are averaged, the model with the lowest R value is chosen, and the averaged map and chosen model are used as starting points for the next cycle of parallel model-building. This entire process is repeated (typically three times total) to yield a final model and density-modified
electron density map. This procedure is carried out with a command such as

```
phenix.parallel_autobuild_run_command=qsub nproc=48
```

data=data/AutoSol_run_2_overall_best_refine_data.mtz seq_file=~/1vlm.fa

```
ha_file=AutoSol_run_2_overall_best_ha.pdb
```

map_file=data/AutoSol_run_2_overall_best_denmod_map_coeffs.mtz

```
model=AutoSol_run_2_overall_best.pdb extreme_dm=True
```

rebuild_in_place=False

---

**Data from the Protein Data Bank.** All the data used in this work except the CysZ data sets and synthetic data were downloaded from the Protein Data Bank (PDB)\(^1\). SAD data sets, along with the anomalously scattering atoms, the wavelengths of data collection and the deposited models, are automatically extracted using the Phenix tool phenix.sad_data_from_pdb. The data sets used are available in [Supplementary Data 1](#) and [Supplementary Data 3](#). The model-phased anomalous difference maps and \(2\Delta F_c\) maps\(^25\) were calculated using the models deposited in the PDB, except that any combinations of SAD data and model that had an \(R\) value greater than 0.30 were rerefined with Phenix\(^17\) before use. This included data from PDB entries 2B78, 2PRR, 3P96, 2HBA, 2A6B, and 2AVN.

Data from the following PDB entries were used in this work (for additional details of data sets used and results for each data set, see [Supplementary Data 1](#) and [Supplementary Data 3](#)):

- 1VJN, 1VJR, 1VJZ, 1VK4, 1VKM\(^26\), 1VLM, 1VQR\(^27\), 1Z82, 1ZYG, 2A3N, 2A6B, 2AML, 2AVN, 2B8M, 2ETD, 2ETJ, 2ETS\(^28\), 2ETV, 2EVR\(^29\), 2F4F, 2FDN\(^30\), 2FEA\(^31\), 2FFI, 2FGQ\(^29\), 2FG9, 2FNA\(^32\), 2FQF, 2FUR, 2FZ7, 2GC9, 2LN3\(^33\), 2NUI, 2NWV\(^33\), 2Q08, 2Q1Q, 2Q2X, 2Q2Z, 2Q3L, 2Q62, 2Q7T, 2Q8Q, 2Q8P, 2Q5C, 2OD5, 2OD6, 2OHD, 2OKC, 2OKF\(^33\), 2O0J, 2OPK, 2Q5D, 2QTM, 2QZG, 2QZI, 2P16, 2P4O, 2P7I, 2P97, 2PG3, 2PG4, 2PGC, 2PIM, 2PN1, 2PPV, 2PR7, 2PRV, 2PVX, 2PV4, 2PV5, 3K9G\(^34\), 3KM3\(^34\), 3QQC\(^35\), 2AZP, 2HZG, 2QDN\(^36\), 2W1Y\(^37\), 4J8S\(^38\), 2I52, 2ZY6\(^39\) and 3GB5\(^40\).

**Reanalysis of CysZ merged data sets using Phenix code developed before any use of the CysZ data sets.** In order to examine whether the availability of the CysZ data sets during development of Phenix brute-force substructure determination caused a bias in our comparison of alternative methods, we created an unbiased Phenix version by combining the release version 1.9-1692 of Phenix with working updates developed before our first examination or use of any CysZ data sets. (Normally there are working versions of Phenix built every night that we could use for this purpose, but during this period the installer software was being updated, and no nightly builds were available.) These updates are available along with instructions for combining them with 1.9-1692 of Phenix at http://www.phenix-online.org/phenix_data/terwilliger/\(^\text{Supplementary Data 2}\). We used this version of Phenix to analyze each data set in [Figure 2](#) using the commands

```
phenix.hysmmerge_16.sca 21 S wavelength=1.7432 resolution=3.5
```

```
rescore=phaser-complete strategy=brute_force n_llg_add_at_once=2
```

```
max_multiple=1 n_top_patt=20 nproc=64
```

where the name of the data file was changed for each data set but the other commands were fixed. Correct sites were assessed by the distance between sites in each solution and the corresponding symmetry-equivalent sites in the sulfur atoms in PDB entry 3TX3, with sites within 3 Å considered as matching. As discussed in the main text, this analysis yielded a number of correct sites for each data set differing by at most two sites from the number found with the fully developed Phenix version used in [Figure 2](#) (Supplementary Data 2). For example, for the data set merge_16.sca corresponding to the data point marked with an arrow in [Figure 2a](#), the solution obtained contained 25 sites, of which 17 were within 3 Å of a sulfur site in PDB entry 3TX3 and which had an r.m.s. difference from corresponding sulfur sites in 3TX3 of 0.50 Å.

**Comparison of Phenix and Crank2 structure determination with substructure determination carried out by Crunch2 or Shelx/c/d.** We performed comparisons of Phenix and Crank2, carrying out with each the entire process of finding the substructure through phase improvement and model-building. The Phenix structure determinations and overall procedures are the same as those shown in [Figure 2c](#). The Crank2 structure determinations began either with sites obtained by Crunch2\(^41\) or with sites determined by Shelx/c/d\(^5,18\), in each case using default parameters in the CCP4i interface\(^12\). Eight of these data sets could be determined by Phenix but not by Crank2 using sites determined by Shelx/c/d\(^5,18\) ([Supplementary Fig. 1](#) and [Supplementary Data 2](#)), and 41 could be determined by Phenix but not Crank2 using sites determined by Crunch2\(^41\) ([Supplementary Fig. 2](#) and [Supplementary Data 2](#)).

We note that there are many powerful software algorithms and suites for automatic or semiautomatic determination of macromolecular structures (for example, refs. 9,11,18,42–45) and that we could have chosen any of these for comparisons. We chose the Crank2 software\(^11\) because it had been recently compared with Phenix and because we used many of the same PDB entries in this work as were used in that comparison (though we have used remote and edge data and eight sulfur SAD data sets not used in that previous work). As most of these data sets were available for both algorithms tested, this choice reduced the bias that can be introduced by using the same data sets in testing and development. We reanalyzed all the data with Crank2 in the CCP4 suite\(^42\) as the edge and remote data sets had not been analyzed previously and as the map correlation information for individual peak data sets was not available from the previous work\(^11\).

The parameters used for Phenix structure determination are as described above

```
phenix.autosol unit_cell=65.648 70.734 93.922 90 90 90 data=w3.sca
atom_type=se lambda=0.97936 seq_file=1vlm.fa thoroughness=medium
```

Parameters used for Crank2 substructure determination were default parameters in the CCP4i interface\(^42\) except for the wavelength and scattering factors which were taken from the Phenix analysis. For the data set above for example, the Phenix analysis estimated that the scattering factors were \(f^= -8.0\) and \(f^"= 4.5\) based on the atom type of selenium and wavelength of 0.97936. Data sets for which no result was obtained using Crank2 (owing to software crashes) are excluded from the analysis (these seven data sets are listed in [Supplementary Data 3](#)).
Parameters for the Shelxc/d substructure determination for this data set were

```
TITL CRANKFa.ins SAD in P21212  
CELL 0.980000 65.65 70.73 93.92 90.00 90.00 90.00  
LATT -1  
SYMM -X, -Y, Z  
SYMM 1/2-X, 1/2+Y, -Z  
SYMM 1/2+X, 1/2-Y, -Z  
SFAC SE  
UNIT 192  
SHEL 999 3.3  
PATS  
FIND 12  
MIND -1.5 -0.1  
NTRY 500  
SEED 1  
HKLF 3  
END
```

Comparison of methods for substructure determination using model SAD data. We compared the overall LLG completion and likelihood-based scoring approach with other widely used methods for finding the anomalously scattering substructure that process the data differently, obtain Patterson-based seeds differently, and use different implementations of completion and scoring. We used a set of synthetic data sets that have been challenging tests of the ability of crystallographic software to determine macromolecular structures using data sets with very low anomalous signal (http://bl831.als.lbl.gov/~jamesh/challenge/anom/). These data sets were created with varying simulated levels of substitution of sulfur with selenium at methionine residues and therefore varying levels of anomalous signal. The simulated data sets contain 12 selenium sites. The high-resolution limit of the data used in all tests (3.5 Å) was chosen to be the resolution at which the anomalous signal (the mean model-phased anomalous difference Fourier peak height at positions of the substructure) was maximal. The tests were carried out within the Phenix17 and CCP4 (ref. 42) software packages.

The numbers of sites identified correctly by each of several approaches are shown (Supplementary Fig. 3 and Supplementary Data 4) as a function of the anomalous signal in the data sets. The dual-space approach implemented in HySS and the difference Fourier approach in SOLVE46 correctly generally identified most of the sites when the anomalous signal was about 12 or greater (though the SOLVE approach was less consistent and solved the substructure in one data set with a signal of 10 but did not solve it in a data set with a signal of 13). The dual-space methods in Crunch2 (ref. 41) and Shelxc/d5,18 correctly identified most of the substructure in data sets with anomalous signal of 10.5 or greater and 9.4 or greater, respectively. The LLG completion and likelihood-scoring approach described here identified most of the substructure in data sets with anomalous signal of 8.7 or greater. As the substructure determination methods tested here have some flexibility in how extensive a search is carried out, we also tested Shelxc/d substructure identification with a thorough search (100,000 tries compared to a typical 1,000 tries) and our brute-force LLG completion approach, in which pairs of sites identified from LLG maps were tested together rather than adding a single site at a time. The Shelxd search correctly determined most of the substructure for data sets with anomalous signal of about 8.7 or greater, and the brute-force LLG approach was successful for those with signal of 8.1 or greater (Supplementary Fig. 3).

As discussed in the main text, it is difficult to compare algorithms with those developed by others without bias. The Phenix brute-force combinatorial approach was developed specifically to solve this set of data sets, whereas the Shelxc/d software was static, so it is possible that Shelxc/d could be used or modified in a way that would allow it to solve a greater fraction of these data sets. We tried to partially compensate for this by allowing very extensive sampling with Shelxc/d, involving even more computation than that used for the brute-force approach. For the data set with the lowest anomalous signal (8.9) that could be solved by Shelxc/d, 106 min were required for 100,000 tries using Shelxc/d on a four-processor machine, and 49 min were required for the same data set and machine for calculations using the brute-force likelihood-based approach. Taken together, our analyses (Fig. 1 and Supplementary Fig. 1) indicate that LLG completion and likelihood scoring can be at least as effective for finding the anomalous substructure in these data sets as the most powerful existing methods.

Parameters used for the Phenix brute-force and Shelxc/d analysis of the data with anomalous signal of 8.7 and used for the timing comparison. The Phenix command used for this analysis was

```
phenix.hyss frac0.83_2.3.mtz n_top_llg=30 \ comparison_emma_model=perfect_ha.pdb \ 12 se resolution=3.5 rescore=phaser-complete \ strategy=brute_force wavelength=0.9792 nproc=4 max_multiple=1
```

where the comparison_emma_model allowed monitoring the number of correct sites during the analysis. For verification that this comparison model had no effect on the outcome, a run was carried out without this keyword. This run also yielded 12 correct sites.

The Shelxc/d parameters (obtained with a default use of the Crank2 CCP4i interface) were

```
TITL frac0_83_1_shelxc_fa.ins SAD in P65  
CELL 0.980000 52.65 52.65 217.04 90.00 90.00 120.00  
LATT -1  
SYMM -Y, X, X, 2/3+Z  
SYMM -X+Y, -X, 1/3+Z  
SYMM -X, -Y, 1/2+Z  
SYMM Y, -X+Y, 1/6+Z  
SYMM X-Y, X, 5/6+Z  
SFAC SE  
UNIT 288  
SHEL 42.036 3.5  
PATS  
FIND 12  
MIND -3.5  
NTRY 100000  
SEED 1  
HKLF 3  
END
```

doi:10.1038/nmeth.3212
Comparison of sensitivity of likelihood-based scoring with correlation scoring. We used the CysZ sulfur-SAD data sets in a test comparing the sensitivity of likelihood-based scoring with that of correlation scoring. Trial solutions for the CysZ anomalous substructure were constructed by seeding Phenix dual-space substructure determination with 1–21 correct sites and generating substructures with 29 sites. After this process, 2,068 trial solutions were obtained containing 0–18 correct sites (within 3 Å of a corresponding sulfur position in the deposited model). These trial substructures were then rescoring using data from the various merged CysZ data sets and either likelihood- or correlation-based scoring. To evaluate the utility of each scoring method for differentiating correct from incorrect solutions, we converted the scores to Z scores showing how many s.d. each score is above the mean for solutions with zero or one correct site for the corresponding data set. The mean Z scores are shown (Supplementary Fig. 4 and Supplementary Data 5) as a function of the number of correct sites in solutions using CysZ merged data sets with anomalous signal less than or greater than 7.5. On average, the LLG-based Z scores are double the correlation-based Z scores, indicating a substantially greater utility in discrimination of correct from incorrect solutions.

Distribution of likelihood-based scores for CysZ merged data set. We tested whether it was possible to identify correct solutions on the basis of their LLG scores and numbers of sites added during the LLG completion process. We found that largely correct solutions to a merged CysZ data set (those containing at least half of the known sites) based on data from three crystals are readily identifiable from their high LLG scores and large numbers of sites added in the likelihood-based completion process (Supplementary Fig. 5 and Supplementary Data 6).

Map correlation as function of anomalous signal for SAD data sets from the PDB after parallel autobuilding. We tested our approach for following the initial structure determination procedure with randomly seeded parallel autobuilding and map averaging. This resulted in a total of 81% of the data sets with anomalous signal from 8 to 30 yielding a final map correlation of 0.50 or greater (Supplementary Fig. 6 and Supplementary Data 3).

Effects of optimizations on performance of Phenix structure determination. We carried out a series of tests to identify the effects of various optimizations on the overall performance of Phenix structure determination. We examined the utility of testing both uncorrected and anisotropy-corrected and sharpened data in structure determination (Supplementary Fig. 7a and Supplementary Data 3). We also examined using only a dual-space substructure search with using dual-space and likelihood-based searches in parallel (Supplementary Fig. 7b and Supplementary Data 3), and not using parameter testing or iteration of substructure searches with using both (Supplementary Fig. 7c and Supplementary Data 3). The optimizations are scored during the structure determination process primarily on the basis of an evaluation of the electron density map. We note that as this evaluation metric is not perfectly correlated with true map quality, there are some cases where optimization yields a poorer result than does using a simpler method.

Comparison of Phenix and Crank2 approaches using synthetic data sets. We applied the current Phenix algorithms and the Crank2 approaches to the synthetic data sets examined above (see Supplementary Fig. 3). In this comparison, the known anomalous substructure was used with the synthetic data to calculate phases and an anomalous difference Fourier. The highest peaks in this map were used as the starting substructure, and the map correlation obtained using each approach is plotted as a function of the anomalous signal in the synthetic data. We found that the Crank2 approaches (Supplementary Fig. 8 and Supplementary Data 7) yielded a largely correct solution (with a map correlation of at least 0.5) when the anomalous signal was at least 7.5. Structure determination was also carried out using the Phenix AutoSol and AutoBuild approaches described here. The initial structure determination with AutoSol was carried out once for each data set, and then this solution was improved with AutoBuild five separate times, each with a different random seed for the process of iterative automated model-building, density modification and refinement. Each of these individual AutoBuild analyses yielded a largely correct solution when the anomalous signal was about 7 or greater. Averaging of the five maps from automated model-building and iteration of the entire process of carrying out five model-building applications in parallel yielded largely correct solutions when the anomalous signal was as low as 6.5.

Display software. We used Coot47 for display and analysis of images of electron density as in Figure 3c.

Data availability. The CysZ data sets were generously provided by Q. Liu and W. Hendrickson and are available from them at http://x4.nsls.bnl.gov/native-SAD/CysZ_native-SAD_individual_plus_merged.tar.bz2. The rescaled and combined data sets used in Figure 2 (Supplementary Data 8) and the spreadsheets used to tabulate the data and prepare the figures (Supplementary Data 1–8) are available at http://www.phenix-online.org/phenix_data/terwilliger/. The synthetic data are available at http://bl831.als.lbl.gov/~jamesh/challenge/occ_scan/.

Software availability. All the Phenix tools and code described here are available from the Phenix website at http://www.phenix-online.org/. Version 1.9 of Phenix and closely related nightly builds were used for all the calculations in this work except for the brute-force substructure calculations, which were carried out with versions dev-1734 and later. All the features described here are available in versions dev-1801 and later of Phenix.
31. Xu, Q. et al. Proteins 69, 433–439 (2007).
32. Xu, Q. et al. Proteins 74, 1041–1049 (2009).
33. Hwang, W.C. et al. Proteins doi:10.1002/prot.24679 (1 September 2014).
34. Abendroth, J. et al. J. Struct. Funct. Genomics 12, 83–95 (2011).
35. Martinez-Rucobo, F.W., Sainsbury, S., Cheung, A.C. & Cramer, P. EMBO J. 30, 1302–1310 (2011).
36. Chattopadhyay, K., Ramagopal, U.A., Brenowitz, M., Nathenson, S.G. & Almo, S.C. Proc. Natl. Acad. Sci. USA 105, 635–640 (2008).
37. Cianci, M., Helliwell, J.R. & Suzuki, A. Acta Crystallogr. D Biol. Crystallogr. 64, 1196–1209 (2008).
38. Fabian, M.R. et al. Nat. Struct. Mol. Biol. 20, 735–739 (2013).
39. Tanaka, Y. et al. RNA 15, 1498–1506 (2009).
40. Thomas, S.R., McTamney, P.M., Adler, J.M., LaRonde-LeBlank, N. & Rokita, S.E. J. Biol. Chem. 284, 19659–19667 (2009).
41. de Graaff, R.A.G., Hilge, M., van der Plas, J.L. & Abrahams, J.P. Acta Crystallogr. D Biol. Crystallogr. 57, 1857–1862 (2001).
42. Winn, M.D. et al. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242 (2011).
43. Panjikar, S. et al. Acta Crystallogr. D Biol. Crystallogr. 61, 449–457 (2005).
44. Wang, J.W. et al. Acta Crystallogr. D Biol. Crystallogr. 60, 1991–1996 (2004).
45. Burla, M.C. et al. J. Appl. Cryst. 45, 357–361 (2012).
46. Terwilliger, T.C. & Berendzen, J. Acta Crystallogr. D Biol. Crystallogr. 55, 849–861 (1999).
47. Emsley, P. & Lohkamp, B. Acta Crystallogr. D Biol. Crystallogr. 66, 486–501 (2010).