HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 August 01.

Published in final edited form as:

Nat Methods. 2015 February ; 12(2): 127–130. doi:10.1038/nmeth.3212.

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

Gábor Bunkóczi[1], **Airlie J. McCoy**[1], **Nathaniel Echols**[2], **Ralf W. Grosse-Kunstleve**[2], **Paul D. Adams**[2], **James M. Holton**[2,3], **Randy J. Read**[1,*], and **Thomas C. Terwilliger**[4,*]

[1]Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720-8235, USA [3]Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158 [4]Los Alamos National Laboratory, Los Alamos NM 87545 USA

Abstract

A likelihood-based method for determining the sub-structure of anomalously-scattering atoms in macromolecular crystals can allow successful structure determination by single-wavelength anomalous diffraction (SAD) X-ray analysis with weak anomalous signal. Along with 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 for extending the applicability of the SAD method in challenging cases.

Single-wavelength anomalous diffraction (SAD phasing) is the dominant X-ray crystallographic method for determination of macromolecular structures by experimental phasing, accounting for 73% of such deposits in the Protein Data Bank[1] in 2013 (www.pdb.org). In the SAD method, the X-ray diffraction from anomalously-scattering atoms in a molecule provides X-ray phase information for the entire crystal structure[2,3]. 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 differences[4,5]. In a second step in structure determination, the sub-structure is used along with the X-ray data (including Bijvoet pairs) to estimate phases for the entire structure and to calculate an electron density map[2,6,7,8]. The phases can then be improved in a third step by an iterative process of phase improvement, model-building, and refinement[9], which can often yield an accurate electron density and a relatively complete model.

*Corresponding authors; rjr27@cam.ac.uk, terwilliger@lanl.gov.

Author contributions

RJR and TCT led the work, RJR, GB, and AJM developed the likelihood-based scoring and LLG maps in Phaser, RWG and PDA developed the HySS framework, JMH developed the synthetic data tests, TCT carried out the tests and optimized the Phenix autosol and autobuild algorithms.

Competing financial interests. The authors declare no competing financial interests.

The SAD phasing method can be challenging if the anomalous signal-to-noise ratio is low[10,11]. 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 their accuracy depends 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 recent experiments using X-ray lasers with SAD phasing, exceptional efforts may be necessary in order to obtain sufficient signal-to-noise[12,13].

A step that can be particularly difficult when the signal-to-noise is low is identifying the positions of the anomalously-scattering atoms in a structure. The sub-structure is often determined using “dual-space” algorithms based on the anomalous differences and alternating real-space peak-picking with reciprocal-space direct methods phase improvement[4,5]. Possible sub-structures generated by dual-space algorithms are scored based on agreement between the structure factors calculated from the sub-structure model and the measured anomalous differences.

Here we introduce the use of a likelihood function to find the sub-structure[7]. The SAD likelihood function describes the probability of measuring the observed data given a model of the sub-structure 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 substructure[7,8] and for finding missing sites in a nearly complete sub-structure[6,14,15]. We find candidate partial sub-structures from the anomalous difference Patterson function and use likelihood-based maps to complete the sub-structure 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 datasets with known structures from the Protein Data Bank[1] (PDB, www.pdb.org). The 162 datasets include anomalous signal from selenium, iodine, mercury, iron and zinc, contain from one to 74 anomalously-scattering atoms in the asymmetric unit, and have high-resolution limits from 1 Å to 3.3 Å. Most of these SAD datasets 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. The anomalous signal in these datasets was evaluated as the mean height of electron density (in units of the rms of the map) at positions of atoms in the known substructure in an anomalous difference Fourier map calculated with phases based on the deposited structure. Implementations of each approach within the same software (HySS[16] in the Phenix software suite[17]) were used in this comparison.

We determined the fraction of sites identified correctly for each dataset using dual-space completion and correlation-based scoring (Fig. 1**a**). For datasets with anomalous signal less than 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. 1**b**). With the likelihood-based approach, nearly all (96%) of those with signal over 7.5 along with some of the substructures (4%) for datasets 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 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 datasets that had been collected to test multi-crystal approaches for cases with a weak anomalous signal[10]. In that work, sulfur-SAD anomalous data was collected from 7 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 dataset. The anomalous substructure could be determined with the 7-crystal dataset and with at least some combinations of datasets assembled from three or more crystals[10]. We created a set of 28 merged datasets using from one to 7 crystals and tested likelihood-based substructure determination using each original or merged dataset. To check whether the likelihood-based methods are comparable to implementations of dual-space methods in other software packages, we carried out Shelxc/d dual-space substructure determination[5,18] with the same datasets. We examined the number of sites correctly identified as a function of the anomalous signal in these datasets (Fig. 2**a**). In our tests, dual-space substructure determination succeeded in at least some cases for merged datasets with an anomalous signal of about 8.4 or greater, while likelihood-based determination succeeded in cases with an anomalous signal as low as 7.4. The same data are plotted as a function of the number of crystals used in each merged dataset (Fig. 2**b**). The likelihood-based approach was successful in identifying the sulfur substructure in four of the eight two-crystal datasets examined. Beginning with the two-crystal merged dataset and sulfur-substructure (marked with an arrow in Fig. 2**b**) and the sequence of the protein, the number of sulfur atoms, and the wavelength of X-ray data collection, the automated structure determination algorithm described below yielded a high-quality electron density map (Fig. 2**c**). The resulting model produced a free R-value of 0.26 and had 435 of 453 residues assigned to sequence. Comparisons with algorithms implemented by other groups are difficult to carry out without bias. The developers of an algorithm are normally more expert at using their software than that of others, and comparison software is normally static while 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 re-analyzing all the datasets (Fig. 2**a**) with Phenix code developed prior to any use with this data (see Online Methods). 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). Overall we find that likelihood-based methods can be exceptionally powerful for sub-structure solution in a challenging case such as this *CysZ* membrane protein structure.

Once the sub-structure is identified, it is used along with the original data to estimate crystallographic phases[2,6,7,8]. 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 modification[19] for integrating information from density modification with phase information from the anomalous differences, optimization of parameters during the structure determination process, iteration[14,15] 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 datasets. We evaluated our methods by using the same software (Phenix[17]) 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 datasets (Fig. 3**a**). The map correlation, a standard metric of the quality of the structure determination process[20], 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 datasets 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. 3**a**). Applying the current algorithms (Fig. 3**b**) allows solution of 79% of these datasets. To place the capabilities of the Phenix algorithms in 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 datasets, 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 based on 73 of the most challenging datasets (from Fig. 3**a**, high-resolution limits ranging from 1.3 to 3.0 Å) with those obtained with the recently-improved algorithms in the Crank2 software[11] (Fig. 3**c**). To focus on the structure determination algorithms, each analyses started with the substructures determined by Phenix. Each point (Fig. 3**c**) has as its x-value the map correlation for the structure produced by the enhanced Crank2 pipeline and its y-value the map correlation from Phenix. We found that 6 of 73 these difficult SAD datasets could not be solved by Crank2 but could be analyzed with Phenix to obtain maps with high correlation to those calculated from the deposited structure (Fig. 3**c**). We also carried out a comparison of Phenix and Crank2, each carrying out the entire process of finding the substructure through phase improvement and model-building. Eight of these datasets could be determined by Phenix but not by Crank2 using sites determined by Shelxc/d (Supplementary Figure 1) and 41 could be determined by Phenix but not Crank2 using sites determined by Crunch2 (Supplementary Figure 2).

We conclude that likelihood-based determination of the anomalous sub-structure, combined with improvements in methodology for phase improvement, can be powerful approaches for structure determination using SAD phasing. It further seems possible that additional improvements in substructure determination may 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, identification of additional sites

based on a substructure, with the rapidity and extensive exploration of possible substructures possible with direct methods approaches.

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 function[7] 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 found[16]. A typical command for mixed dual-space/LLG substructure determination is:

```
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 based on 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 4-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:

```
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 AutoSol[20] and AutoBuild[21] were extended and used in this work.

Statistical density modification is carried out as described[19], 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 region[23].

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 modification[3], scored based on the agreement factor (R-value) obtained from density modification[20]. 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 obtained[20]. 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 identification[19].

Map-based iteration of anomalous substructure determination[14,15] 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-building[24] 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 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 Fig. 3.

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

```
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
```

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,

```
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=AutoSol_run_2_/
overall_best_refine_data.mtz seq_file=../1vlm.fa ha_file=AutoSol_run_2_/
overall_best_ha_pdb.pdb map_file=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* datasets and synthetic data were downloaded from the Protein Data Bank[1]. SAD datasets, 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 datasets used are listed in the spreadsheets supplied as supplementary material (Supplementary Data 1 and Supplementary Data 3). The model-phased anomalous difference maps and *2mFo-DFc exp(i*ϕ_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 re-refined 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 datasets used and results for each dataset see Supplementary Data 1 and Supplementary Data 3): 1VJN , 1VJR , 1VJZ , 1VK4 , 1VKM[26], 1VLM , 1VQR[27], 1Z82 , 1ZYB , 2A3N , 2A6B , 2AML , 2AVN , 2B8M , 2ETD , 2ETJ , 2ETS[28], 2ETV , 2EVR[29], 2F4P , 2FDN[30], 2FEA[31], 2FFJ , 2FG0[29], 2FG9 , 2FNA[32], 2FQP , 2FUR , 2FZT , 2G42 , 2GC9 , 2NLV[33], 2NUJ , 2NWV[33], 2O08 , 2O1Q , 2O2X , 2O2Z , 2O3L , 2O62 , 2O7T , 2O8Q , 2OBP , 2OC5 , 2OD5 , 2OD6 , 2OH3 , 2OKC , 2OKF[33], 2OOJ , 2OPK , 2OSD , 2OTM , 2OZG , 2OZJ , 2P10 , 2P4O , 2P7I , 2P97 , 2PG3 , 2PG4 , 2PGC , 2PIM , 2PN1 , 2PPV , 2PR7 , 2PRV , 2PRX , 2PV4 , 2PW4 , 3K9G[34], 3KM3[34], 3QQC[35], 2AZP, 2HZG, 2QDN[36], 2W1Y[37], 4J8S[38], 2I52, 2ZY6[39], 3GB5[40].

Re-analysis of *CysZ* merged datasets using Phenix code developed prior to any use of the *CysZ* datasets

In order to examine whether the availability of the *CysZ* datasets 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 prior to our first examination or use of any *CysZ* datasets. (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 are 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/. We used this version of Phenix to analyze each dataset in Fig. 2 (these data are available as Supplementary Data 8) using the commands:

```
phenix.hyss merge_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_llg=20 n_top_patt=30 nproc=64
```

where the name of the datafile was changed for each dataset 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 text this analysis yielded a number of correct sites for each dataset differing by at most two sites from the number found with the fully-developed Phenix version used in Fig. 2. For example, for the dataset merge_16.sca corresponding to the datapoint marked with an arrow in Fig. 2**a**, the solution obtained contained 25 sites, of which 17 were within 3 Å of a sulphur site in PDB entry 3tx3, and which had an rms difference from corresponding sulphur sites in 3tx3 of 0.50 Å.

Comparison of Phenix and Crank2 structure determination with substructure determination carried out by Crunch2 or Shelxc/d

We carried out comparisons of Phenix and Crank2, each carrying out 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 Fig. 3**c**. The Crank2 structure determinations began either with sites obtained by Crunch2[43] or by sites determined by Shelxc/d[5,18], in each case using default parameters in the CCP4i interface[41]. Eight of these datasets could be determined by Phenix but not by Crank2 using sites determined by Shelxc/d[5,18] (Supplementary Figure 1) and 41 could be determined by Phenix but not Crank2 using sites determined by Crunch2[43] (Supplementary Figure 2).

We note that there are many powerful software algorithms and suites for automatic or semi-automatic determination of macromolecular structures (for example, refs 9,11,18,44–47) 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 8 sulfur SAD datasets not used in that previous work). As most of these datasets were available for both algorithms tested, this choice reduced the bias that can be introduced by using the same datasets in testing and development. We re-analyzed all the data with Crank2 in the CCP4 suite[44] as the edge and remote datasets had not been analyzed previously and as the map correlation information for individual peak datasets 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[41] except for the wavelength and scattering factors which were taken from the Phenix analysis. For the dataset above for example, the Phenix analysis estimated that

the scattering factors were f'=–8.0 f"=4.5 based on the atom type of selenium and wavelength of 0.97936. Datasets for which no result was obtained using Crank2 (due to software crashes) are excluded from the analysis (these 7 datasets are listed in Supplementary Data 3).

Parameters for the Shelxc/d substructure determination for this dataset were:

TITL CRANK_fa.ins SAD in P21212

CELL 0.98000 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 datasets that have been used as challenging tests of the ability of crystallographic software to determine macromolecular structures using datasets with very low anomalous signal (http://bl831.als.lbl.gov/~jamesh/challenge/anom/). These datasets were created with varying simulated levels of substitution of sulfur with selenium at methionine residues and therefore varying levels of anomalous signal. The simulated datasets 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 Phenix[17] and CCP4[41] software packages.

The number of sites identified correctly by each of several approaches are shown (Supplementary Figure 3) as a function of the anomalous signal in the datasets. The dual-space approach implemented in HySS and the difference Fourier approach in SOLVE[42] 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 dataset with a signal of 10 but did not solve it in a dataset with a signal of 13.) The dual-space methods in Crunch2[43] and Shelxc/d[5,18] correctly identified most of the substructure in datasets 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 datasets 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 1000 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 datasets 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 Figure 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 datasets, while 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 datasets. 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 dataset with the lowest anomalous signal (8.9) that could be solved by Shelxc/d, 106 minutes were required for 100,000 tries using Shelxc/d on a 4-processor machine, and 49 minutes were required for the same dataset and machine for calculations using the brute-force likelihood-based approach. Taken together, our analyses (Fig. 1 and Supplementary Figure 1) indicate that LLG completion and likelihood scoring can be at least as effective for finding the anomalous substructure in these datasets 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. To verify 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 0.98000 52.65 52.65 217.04 90.00 90.00 120.00

LATT -1

SYMM -Y, X-Y, 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

Comparison of sensitivity of likelihood-based scoring with correlation scoring.

We used the *CysZ* sulfur-SAD datasets 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 one to 21 correct sites and generating substructures with 29 sites. After this process 2068 trial solutions were obtained containing 0 to 18 correct sites (within 3 Å of a corresponding sulfur position in the deposited model). These trial substructures were then rescored using data from the various merged *CysZ* datasets and either likelihood- or correlation-based scoring. To evaluate the utility of each scoring method for differentiating correct from incorrect solutions, the scores were converted to Z-scores showing how many standard deviations each score is above the mean for solutions with zero or one correct site for the corresponding dataset. The mean Z-scores are shown (Supplementary Figure 4) as a function

of the number of correct sites in solutions using *CysZ* merged datasets 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 dataset

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

Map correlation as function of anomalous signal for SAD datasets 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 datasets with anomalous signal from 8-30 yielding a final map correlation of 0.50 or greater (Supplementary Figure 6)

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 Figure 7**a**). We also examined using only a dual-space substructure search with using dual-space and likelihood-based searches in parallel (Supplementary Figure 7**b**), and not using parameter testing or iteration of substructure searches with using both (Supplementary Figure 7**c**). The optimizations are scored during the structure determination process based primarily on 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 using a simpler method.

Comparison of Phenix and Crank2 approaches using synthetic datasets

We applied the current Phenix algorithms and the Crank2 approaches to the synthetic datasets examined above (see Supplementary Figure 3). In this comparison, the known anomalous sub-structure 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 sub-structure, 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 Figure 8) 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 dataset, 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 used

We used Coot[48] for display and analysis of images of electron density as in Fig. 3**c**.

Data availability

The *CysZ* datasets 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 datasets used in Fig. 2 and the spreadsheets used to tabulate the data and prepare the figures 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 web site 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Q. Liu, W. Henderickson and colleagues for generously providing the *CysZ* sulfur SAD datasets. Support received from the NIH (grant P01GM063210 to PDA, TCT and RJR and grant GM073210 to JH) and the Wellcome Trust (Principal Research Fellowship to RJR, grant 082961) is gratefully acknowledged. This work was partially supported by the US Department of Energy under Contract DE-AC02-05CH11231.

References

1. Berman HM, et al. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
2. Hendrickson WA, Teeter M,M. Nature. 1981; 290:107–113.
3. Wang B-C. Methods Enzymol. 1985; 115:90–112. [PubMed: 4079800]
4. Weeks CM, DeTitta GT, Miller R, Hauptman HA. Acta Cryst. 1993; D49:179–181.
5. Schneider TR, Sheldrick GM. Acta Cryst. 2002; D58:1772–1779.
6. La Fortelle E, de & Bricogne G. Methods Enzymol. 1997; 276:472–494.
7. McCoy AJ, Storoni LC, Read RJ. Acta Cryst. 2004; D60:1220–1228.
8. Pannu NS, Read RJ. Acta Cryst. 2004; D60:22–27.
9. Langer G, Cohen SX, Lamzin VS, Perrakis A. Nature Protocols. 2008; 3:1171–1179. [PubMed: 18600222]

10. Liu Q, et al. Science. 2012; 336:1033–1037. [PubMed: 22628655]
11. Skubák P, Pannu R. Nature Communications. 2013; 4 Article number 2777, doi:10.1038/ncomms3777.
12. Weiss MS, Sicker T, Hilgenfeld R. Structure. 2001; 9:771–777. [PubMed: 11566127]
13. Barends TR, et al. Nature. 2014; 505:244–247. [PubMed: 24270807]
14. McCoy AJ, Read RJ. Acta Cryst. 2010; D66:458–469.
15. Read RJ, McCoy A. Acta Cryst. 2011; D67:338–344.
16. Grosse-Kunstleve RW, Adams PD. Acta Cryst. 2003; D59:1966–1973.
17. Adams PD, et al. Acta Cryst. 2010; D66:213–221.
18. Sheldrick GM. Acta Cryst. 2010; D66:479–485. (2010).
19. Terwilliger TC. Acta Cryst. 2000; D56:965–972. (2000).
20. Terwilliger TC, et al. Acta Cryst. 2009; D65:582–601.
21. Terwilliger TC, et al. Acta Cryst. 2008; D64:61–69.
22. Blanc E, Roversi P, Vonrhein C, Flensburg C, Lea SM, Bricogne G. Acta Cryst. 2004; D60:2210–2221.
23. Terwilliger TC. Acta Cryst. 2003; D59:1174–1182.
24. Perrakis A, Sixma TK, Wilson KS, Lamzin VS. Acta Cryst. 1997; D53:448–455.
25. Read RJ. Acta Cryst. 1986; A42:140–149.
26. Levin I, et al. Proteins. 2005; 59:864–868. [PubMed: 15822122]
27. Xu Q, et al. Proteins. 2006; 62:292–296. [PubMed: 16287129]
28. Kozbial P, et al. Proteins. 2008; 71:1589–1596. [PubMed: 18324683]
29. Xu Q, et al. Structure. 2009; 17:303–313. [PubMed: 19217401]
30. Dauter Z, Wilson KS, Sieker LC, Meyer J, Moulis JM. Biochemistry. 1997; 36:16065–16073. [PubMed: 9405040]
31. Xu Q, et al. Proteins. 2007; 69:433–439. [PubMed: 17654724]
32. Xu Q, et al. Proteins. 2009; 74:1041–1049. [PubMed: 19089981]
33. Hwang WC, et al. Proteins. 2014 (DOI: 10.1002/prot.24679).
34. Abendroth J, et al. J Struct Funct Genomics. 2011; 12:83–95. [PubMed: 21359836]
35. Martinez-Rucobo FW, Sainsbury S, Cheung AC, Cramer P. EMBO J. 2011; 30:1302–1310. [PubMed: 21386817]
36. Chattopadhyay K, Ramagopal UA, Brenowitz M, Nathenson SG, Almo SC. Proc. Natl. Acad. Sci. USA. 2008; 105:635–640. [PubMed: 18182486]
37. Cianci M, Helliwell JR, Suzuki A. Acta. Cryst. 2008; D64:1196–1209.
38. Fabian MR, et al. Nat. Struct. Mol. Biol. 2013; 20:735–739. [PubMed: 23644599]
39. Tanaka Y, et al. RNA. 2009; 15:1498–1506. [PubMed: 19509301]
40. Thomas SR, McTamney PM, Adler JM, Laronde-Leblank N, Rokita SE. J. Biol. Chem. 2009; 284:19659–19657. [PubMed: 19436071]
41. Winn MD, et al. Acta Cryst. 2011; D67:235–242.
42. Terwilliger TC, Berendzen J. Acta Cryst. 1999; D55:849–861.
43. Graaff RAG, de, Hilge M, van der Plas JL, Abrahams JP. Acta Cryst. 2001; D57:1857–1862.
44. Winn MD, et al. Acta. Cryst. 2011; D67:235–242.
45. Panjikar S, et al. Acta Cryst. 2005; D61:449–457.
46. Wang JW, et al. Acta Cryst. 2004; D60:1991–1996.
47. Burla MC, et al. J. Appl. Cryst. 2012; 45:357–361.
48. Emsley P, Lohkamp B. Acta Cryst. 2010; D66:486–501.

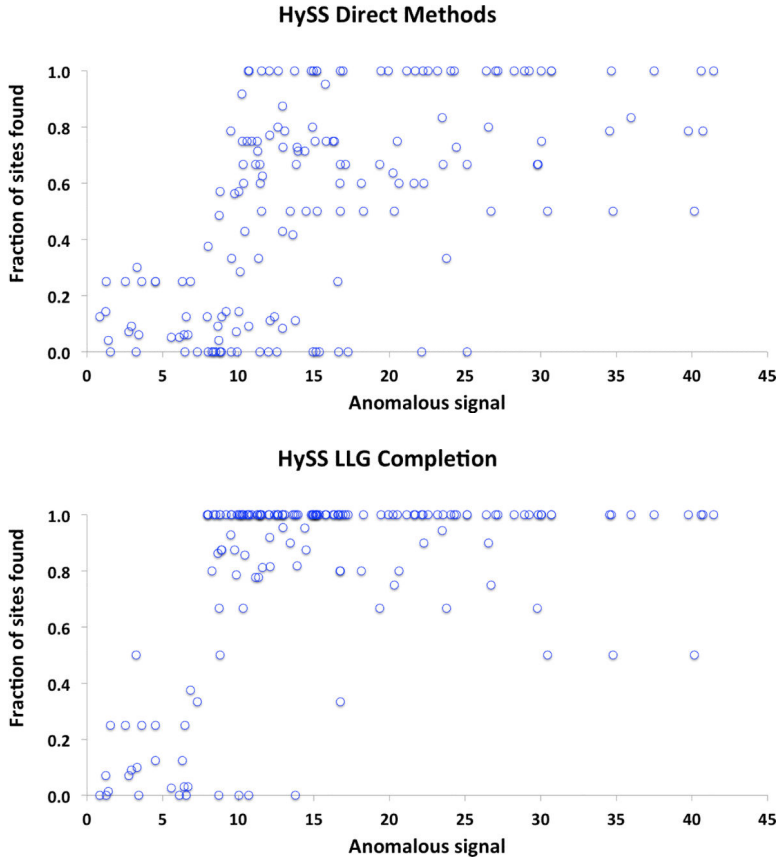
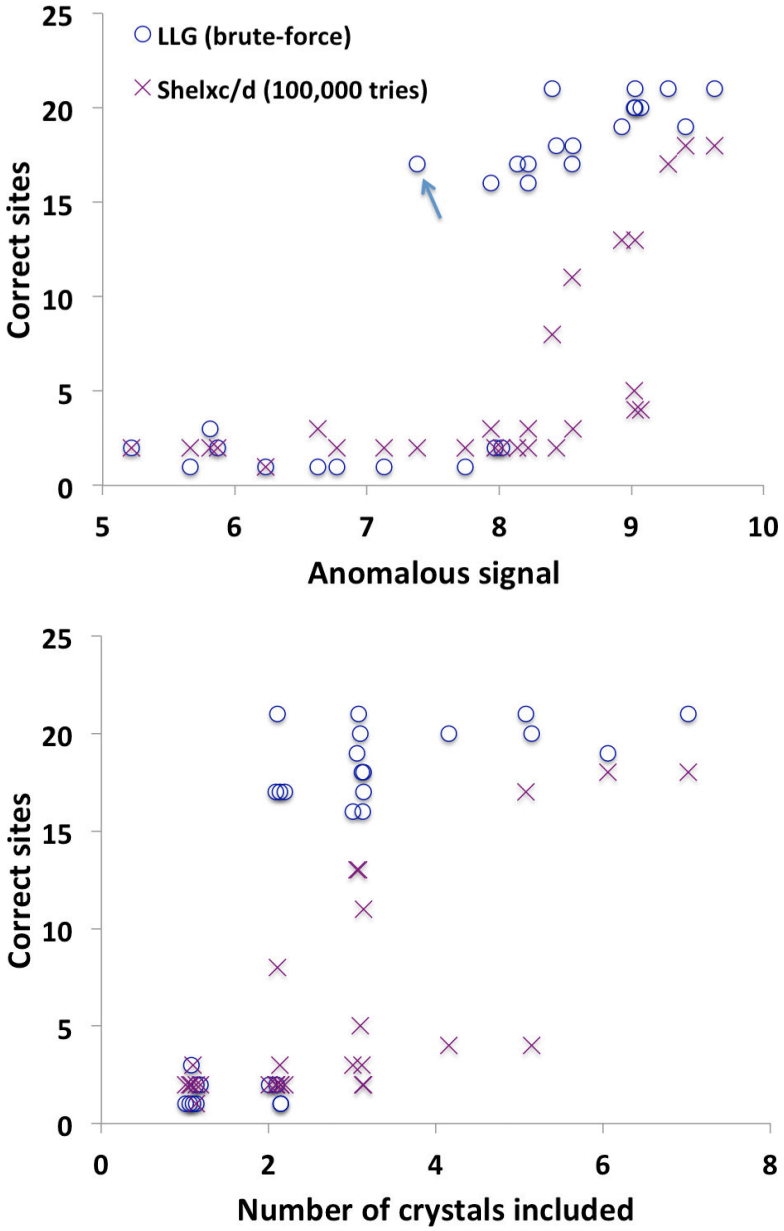


Figure 1.

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[17] software system is plotted as a function of the anomalous signal in each SAD dataset. 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 datasets where the deposited model did not correspond to the anomalous dataset), at the coordinates of the atoms in the anomalous substructure (see 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 is carried out using the SAD likelihood function.



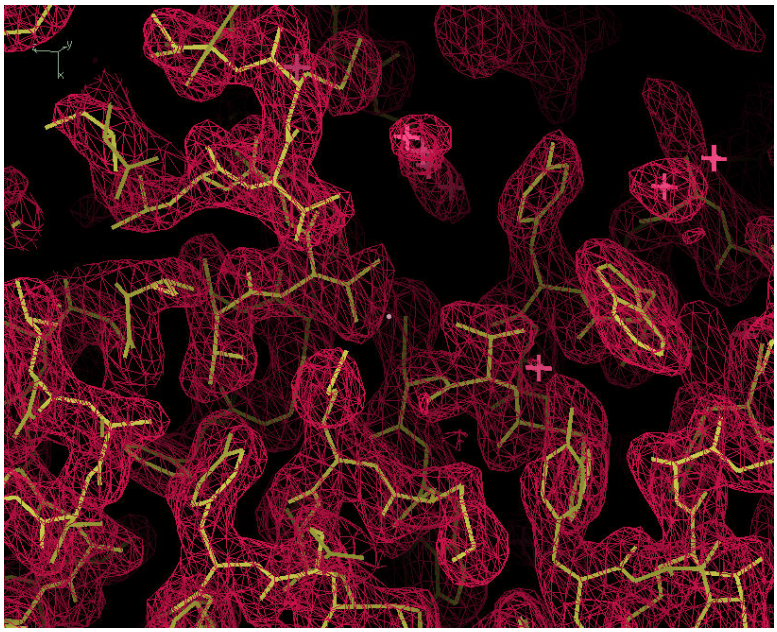
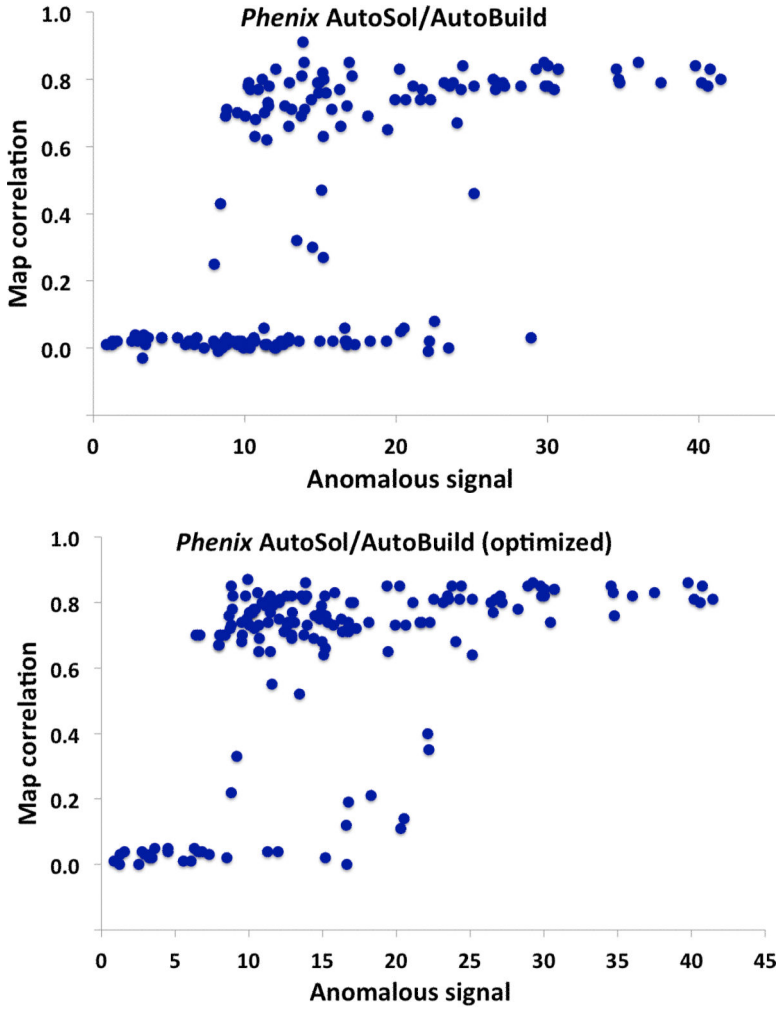


Figure 2.

Application of substructure completion algorithms to *CysZ* datasets merged from varying numbers of crystals. (**a**) Dual-space substructure determination with Shelxc/d [5,18] (100,000 tries, purple crosses) and brute-force likelihood-based completion (blue circles) are shown. A cutoff of 3.5 Å was used throughout. We also tested a cutoff of 2.8 Å for Shelxc/d and obtained similar results (At least 8 of 21 sites were found for 6 datasets using a cutoff of 2.8 Å and for 7 datasets 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 datasets (mean peak height at positions of atoms in the known substructure in model-phased anomalous difference Fourier map). (**b**) As 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[10]).



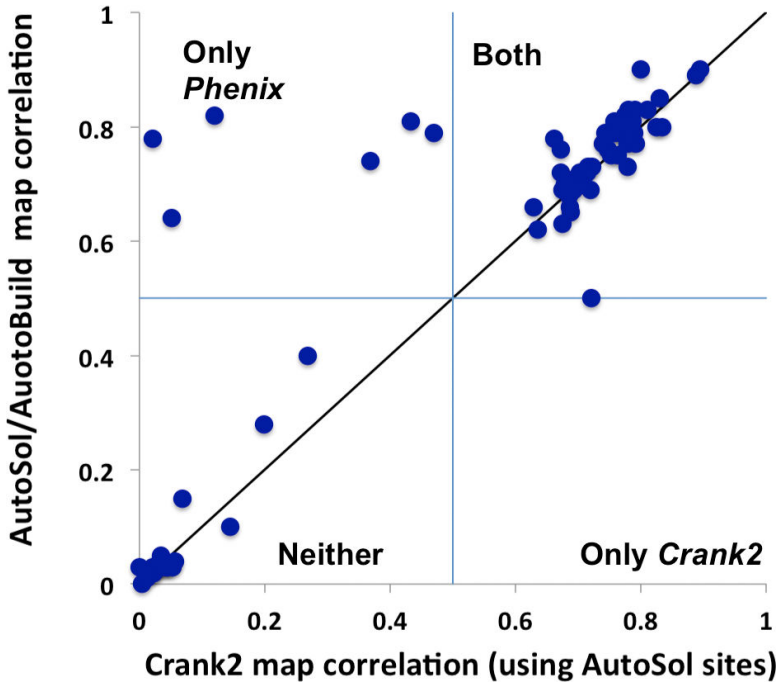


Figure 3.

Map correlation after structure determination. (**a**) Map correlation as function of anomalous signal for SAD datasets from the PDB. Phenix[17] structure determination without using the new features described here. (**b**) as in **a** with optimized procedures enabled using the parameter “thoroughness” set to “medium” (See Online Methods). (**c**) Comparison of structure determination using Phenix with structure determination using Crank2[11] starting with substructure determined with Phenix. The labels indicate whether Phenix and Crank2 succeed in obtaining an electron density map with correlation of 0.50 or greater.