

Decision-Making in Structure Solution using Bayesian Estimates of Map Quality: The PHENIX AutoSol Wizard

Thomas C. Terwilliger^a, Paul D. Adams^b, Randy J. Read^c, Airlie J. McCoy^c, Nigel W. Moriarty^b, Ralf W. Grosse-Kunstleve^b, Pavel V. Afonine^b, Peter H. Zwart^b, Li-Wei Hung^a

^aLos Alamos National Laboratory, Los Alamos, NM 87545, USA.

^bLawrence Berkeley National Laboratory, One Cyclotron Road, Bldg 64R0121, Berkeley, CA 94720, USA.

^cDepartment of Haematology, University of Cambridge, Cambridge CB2 0XY, UK

Running title: Bayesian estimates of map quality in the AutoSol Wizard

Abstract Estimates of the quality of experimental maps are important in many stages of structure determination of macromolecules. Map quality is defined here as the correlation between a map and the map calculated based on a final refined model. Here we examine 10 different measures of experimental map quality using a set of 1359 maps calculated by reanalysis of 246 solved MAD, SAD, and MIR datasets. A simple Bayesian approach to estimation of map quality from one or more measures is presented. We find that a Bayesian estimator based on the skew of histograms of electron density is the most accurate of the 10 individual Bayesian estimators of map quality examined, with a correlation between estimated and actual map quality of 0.90. A combination of the skew of electron density with the local correlation of *rms* density gives a further improvement in estimating map quality, with an overall correlation coefficient of 0.92. The PHENIX AutoSol Wizard carries out automated structure solution based on any combination of SAD, MAD, SIR, or MIR datasets. The Wizard is based on tools from the PHENIX package and uses the Bayesian estimates of map quality described here to choose the highest-quality solutions after experimental phasing.

Keywords: X-ray crystallography; structure solution; scoring; Protein Data Bank; phasing; decision-making; PHENIX; experimental electron-density maps

1. Introduction

Structure solution in macromolecular crystallography is a multi-step procedure in which more than one plausible possibility often exists at the conclusion of each step. At the start of the process one or more MAD, SAD, SIR or MIR datasets are collected and reduced to a list of indices and structure factor amplitudes (Leslie, 1992; Otwinowski & Minor, 1997; Pflugrath,

1999). Even at this stage there are often several possibilities for the space group that must be considered. For each possible space group, the process continues with finding a substructure containing heavy atoms or anomalously-scattering atoms (Grosse-Kunstleve, & Adams, 2003; Schneider & Sheldrick, 2002; Terwilliger & Berendzen, 1999; Weeks et al., 2003). There is often more than one plausible substructure at this stage. For example in space groups that are not chiral the two possible hands of the substructure cannot normally be distinguished. Furthermore for MAD datasets there may be alternative solutions found by searching for the substructure using different datasets (from various wavelengths or combining data from different wavelengths using F_A values; Terwilliger & Berendzen, 1994). Similarly, for MIR datasets there may also be substructures found for several different derivatives. In addition to these intrinsic possibilities, it is possible that more than one set of parameters or even more than one set of software might be used to generate possible solutions. The potential heavy-atom substructures found are then used to calculate phases of structure factors, which in turn are used as the starting point for density modification (Wang, 1985) and subsequent model-building (e.g., Perrakis et al., 1999; Terwilliger et al, 2007). Normally one of the best indications of map quality is that the map that can be interpreted in terms of an atomic model.

If every possibility at every stage were investigated fully by calculating maps, carrying out density modification and model-building, the process might take many hours or days to complete. To speed up the process, the possibilities at each stage are generally ranked, with only the highest-ranked possibilities being considered for the next step. This approach can be quite efficient, but if it is to yield the best solution at the end, it requires a reliable method for deciding which members of a set of solutions are of the highest quality.

The definition of “quality” when applied to electron density maps normally refers to the correlation between values of electron density in the map and the values of electron density in a hypothetical “true” map for the same structure. In this work, when tests are carried out to assess various measures of map quality, the “true” quality or map correlation is calculated between the map in question and a map calculated from a refined model of the corresponding structure. Maps that have a high map correlation as defined in this way are generally more useful for model-building and interpretation than those with a low map correlation. It should be noted however, that map correlation is not a perfect way to assess the utility of a map, as low-resolution terms are generally stronger and therefore have a higher relative contribution to the correlation than high-resolution terms, while the high-resolution terms are generally essential for interpretation of a map.

A number of methods for evaluating the quality of experimental macromolecular electron density maps have been developed. The methods can generally be grouped into real-space calculations and reciprocal-space calculations. Real-space methods are based on an examination of the electron density map and generally answer the question: “Does this map look like an electron density map of a macromolecule?” There are many distinctive features of macromolecular electron density maps that can be used to answer this question. A good map may be expected to have continuous chains of density (Baker et al., 1993). It may have local patterns of density that reflect shapes and interatomic spacings common to macromolecules (Colovos et al., 2000; Terwilliger, 2003). It may have a distribution of electron densities with a positive skew, reflecting the large number of points with moderate or low electron density, the lack of points with negative density, and the points with very positive electron density located near atoms in the structure (Podjarny, 1976; Lunin, 1993). There may be a large variation (contrast) in the local rmsd of electron density, reflecting regions of the structure containing the macromolecule (with high local variation) and solvent (with low local variation; Terwilliger, & Berendzen, 1999a, Schneider & Sheldrick, 2002). The contiguous nature of the regions of relatively flat solvent may be detected from the correlation of local rmsd at one point in a map with that at neighboring points (Terwilliger & Berendzen, 1999b). If non-crystallographic symmetry is present in the structure, then the correlation of NCS-related density can be detected (Cowtan & Main, 1998; Vellieux et al., 1999; Terwilliger, 2002a).

Reciprocal-space methods for evaluation of map quality generally address questions involving structure factors and expectations about the structure such as the model for the solvent region or for the heavy-atom substructure. One such question is simply, “Given the anomalously-scattering atom model and the observed data, what is the expected correlation between the experimental map and the true map?”. The value of the figure of merit of phasing (Blow & Crick, 1969; Terwilliger & Berendzen, 1999), when estimated correctly, is similar in magnitude to the correlation between the experimental and true maps and can be used as an estimate of this correlation. Another question addresses the data and the expectations about the electron density map: “Is the amplitude of each structure factor consistent with value expected based on the amplitudes and phases of all other reflections and the model of the solvent region?” This question can be answered based on the R-factor in the first cycle of density modification (which reflects the agreement between each measured amplitude and an estimate of that amplitude based on all other amplitudes and phases along with expectations about features in the map; Cowtan & Main, 1996; Terwilliger, 2001). A related question can be asked about the phases: “If a phase is estimated from the model of the solvent region, measured amplitudes of structure factors, and the

experimental values of all other phases, is this phase correlated with its experimentally-determined value?” This question can be answered using the correlation of experimental phases with map-probability phases obtained in statistical density modification (Terwilliger, 2001). A third question that might be asked is, “Do the phases calculated using only the highest peaks in the map match the experimental phases?” This question can be answered by truncating the density at a high level, calculating phases from the map, and comparing these with the experimental phases (Baker et al., 1993).

It is important to note that the measures of map quality are analyzed here for their utility in distinguishing quality of *experimental* electron density maps, as opposed to maps that have been calculated using a partially-correct model or maps that have had density modification applied. An important difference between experimental maps and those obtained using a model or based on density modification is that in the latter cases the maps have been specifically adjusted to maximize one or more of the properties that is being measured. For example, density modification typically flattens the solvent region of the map. Similarly, a map calculated from a model will tend to have a high skew of electron density and a high connectivity of high electron density. Some of these measures may also be useful in these two other important cases, but the values of each measure corresponding to a particular quality of map are likely to be substantially different.

In this work we implement 10 different measures of quality of experimental electron density maps, develop a simple Bayesian approach to estimating map quality from each, and show how the individual estimates can be combined to yield useful overall estimates of map quality. These map quality estimates are incorporated into the PHENIX AutoSol Wizard and are used to make decisions during automated structure solution.

2. Materials and Methods

2.1. Structure solution with the AutoSol Wizard

The AutoSol Wizard carries out structure solution for SAD/MAD or MIR/SIR/SIRAS data and any combination of these. If data representing more than one heavy-atom substructure is available, the data are grouped into “datasets” with common heavy-atom substructures.

Analysis with phenix.xtriage. Each available set of data is analyzed using *phenix.xtriage* (Zwart et al., 2005) for circumstances such as twinning, translational non-crystallographic symmetry, unexpectedly strong or weak reflections or groups of reflections, or anisotropic overall atomic displacement parameters that may complicate structure determination. The data are

corrected for anisotropy before structure solution is carried out if the overall anisotropy correction yields values that are highly anisotropic (by default, defined as greater than 1.5-fold ratio among the atomic displacement parameters' values along the three principal reciprocal axes and greater than 20 Å² difference between the highest and lowest values). If an anisotropy correction is applied, then the resulting corrected data are used for structure solution only and not for refinement (as an anisotropy correction is applied as part of the refinement process itself).

Substructure solution with HYSS. For each dataset (i.e. a MAD or SAD dataset or a SIR dataset) possible heavy-atom substructures are found using the hybrid substructure search (*HYSS*; Grosse-Kunstleve & Adams, 2003) from isomorphous, anomalous, or dispersive differences, or from F_A values (Terwilliger, 1994). The high-resolution limit used for the search is typically 3 Å.

Phasing with Phaser and SOLVE and map evaluation. Each potential heavy-atom substructure found above (along with their inverses) are used to calculate phases with Phaser (for SAD phasing; McCoy et al., 2004) or SOLVE (for MAD, SIR and MIR phasing; Terwilliger & Berendzen, 1996; Terwilliger & Berendzen, 1997; Terwilliger & Berendzen, 1999). The resulting phases and amplitudes of structure factors, along with weights (the figure of merit of phasing) are used to calculate experimental electron density maps using a high-resolution limit of 2.5 Å (or lower, if data are not available to this resolution). The high-resolution limit is applied to reduce the effects of resolution cutoffs on the features of electron density maps. These maps are evaluated with the measures of map quality described in this work and the overall Bayesian estimate of quality is used to rank solutions. In cases where two solutions have very similar heavy-atom parameters (*rmsd* among heavy-atom coordinates of less than 1/10 the high-resolution limit of the data) The estimate of uncertainty in the map quality is used to identify solutions that might plausibly (5% possibility or greater) be the best solution and normally all such solutions are considered at each step. By default up to 3 of the highest-ranking solutions (6 for MIR structures) for the heavy-atom substructure are used to calculate phases and weights at the full available resolution of the data and for density modification. In the structure determinations carried out below for development of the map evaluation criteria, rankings were done using a Z-score procedure (Terwilliger & Berendzen, 1999) based only on the skew of electron density (as defined below).

Statistical density modification with RESOLVE. The experimental phases obtained above are used as a starting point for statistical density modification using RESOLVE (Terwilliger, 2000). In statistical density modification with the AutoSol Wizard, a probabilistic estimate of the boundary between macromolecule and solvent is identified in two ways, and the one leading to the lower R-factor in density modification is used. The first method (Wang, 1985) is based on the

local *rms* density, smoothing the squared density using a sphere (Leslie, 1987) with a smoothing radius (r_{smooth}), given by an empirically-derived formula (chosen by optimizing parameters carrying out density modification using model data):

$$r_{smooth} (\text{\AA}) = 2.41 \text{\AA} (d_{\min}/1\text{\AA})^{0.9} \langle m \rangle^{-0.26} \quad , \quad (1)$$

where d_{\min} is the high-resolution limit of the data and $\langle m \rangle$ is the mean figure of merit of phasing. The second method for solvent boundary identification uses a comparison of histograms of density based on model maps calculated with partially-randomized phases with local histograms of density in the experimental map to assign a probability that each point in the map is part of the macromolecule or part of the solvent region. In both cases a probabilistic solvent boundary is obtained (Terwilliger, 1999).

Non-crystallographic symmetry is used in density modification if it is detected based on the heavy-atom substructure and the presence of correlated density at NCS-related positions in the electron density map (Terwilliger, 2002a; Terwilliger, 2002b). The value of r_{smooth} described above is used as a smoothing radius in a local correlation map to identify the region over which NCS holds (Vellieux et al., 1995).

Model-building with RESOLVE. After density modification, the AutoSol Wizard carries out automated model-building using a single cycle of building with the PHENIX AutoBuild Wizard (Terwilliger et al., 2007), or using rapid methods for building secondary structure of proteins and nucleic acids (TT, unpublished). Initially a secondary-structure-only model is built into each map. The correlation between a map calculated from the model and the density-modified map is then determined. If the value of the map-model correlation is less than a cutoff value (typically 0.35) then the building procedure is repeated with a standard cycle of building using the methods in the PHENIX AutoBuild Wizard. If a map-model correlation of a given cutoff (typically 0.20) or greater is obtained for at least one solution, then the top solution is identified as the one with the highest value of the map-model correlation. If a lower map-model correlation is obtained, then the top solution is identified (see below) based on the Bayesian estimates of quality using the skew of electron density (*skew*) and the correlation of local rms density (r_{RMS}^2).

2.2. Evaluation of measures of map quality

A set of measures of map quality were applied to a set of experimental maps (or structure factor amplitudes, phases and weights) obtained from real but re-enacted structure determinations. Each of the structures considered had been determined previously, so that a

refined model could be used to calculate a model map to use as a standard. The “true” quality of each map was taken to be the correlation with the corresponding standard map, calculated at the same nominal resolution. Each measure of quality was applied to each map and the resulting scores were saved along with the corresponding “true” quality. The structure solution process was automatically carried out by the PHENIX AutoSol Wizard, and each experimentally-phased map that was obtained during the structure solution process was examined in this way. To reduce the number of near-duplicate solutions considered, all solutions for a structure that had nearly-identical values of the map correlation to the standard map (within a range of ± 0.0005 in map correlation) were considered the same, and only the first was used in the analysis. For comparisons involving two possible enantiomers of a solution, the two enantiomers of a solution sometimes differed only slightly (i.e., the heavy-atom substructure was nearly centrosymmetric). In these analyses of enantiomeric pairs, only those that differed by an *rmsd* of at least 0.5 Å were considered.

For analysis of map quality, electron density maps and structure factors were calculated using a high-resolution limit of 2.5 Å (if data were available to that resolution), as described for the AutoSol Wizard above. Before applying each of the measures of map quality, the experimental maps were normalized to a mean of zero and a variance of unity. They were then adjusted in two steps to reduce the contribution from high density at the coordinates of heavy-atom sites. (The high density at heavy-atom sites might otherwise lead to high values for the skew, NCS correlation, contrast, and possibly other measures.) First, the electron density within a radius (r) of each heavy-atom site used in phasing (where r was given by twice the resolution of the data or 5 Å, whichever was greater) was limited to values less than or equal to twice the *rms* (2σ) of the map. Second, the electron density everywhere in the map was limited to values in the range of -5σ to $+5\sigma$. This modified map is referred to below as the normalized, truncated experimental electron density map.

Weighted electron density maps were calculated in the PHENIX environment (Adams et al., 2002) using RESOLVE (Terwilliger, 2000) on a grid with spacing of 1/3 the high-resolution limit of the data or finer. Map correlations were obtained by calculating the correlation coefficient of a pair of maps at all the grid points in the asymmetric unit of the unit cell. Model-map correlations were calculated in the same way, except that one map was calculated from the model and an overall B-factor (*b_{overall}*) was adjusted to maximize the correlation. For protein chains, an increment in B-factors (*beta_b*) for each bond between side-chain atoms and the C_β atom was also applied.

2.3. Real-space map-quality measures

The measures of map quality used in this work are described in this and the following section and are summarized in Table 1.

Skew of electron density: The *skew* of each normalized, truncated map (as described in section 2.1) was calculated using the relation,

$$skew = \langle \rho^3 \rangle / \langle \rho^2 \rangle^{3/2} \quad , \quad (2)$$

where the electron density (ρ) was calculated at all the grid points in the asymmetric unit.

Contrast of electron density. The contrast between the *rms* (root-mean-square) density in the solvent region and the *rms* density in the macromolecular region was calculated from the standard deviation of the local *rms* density over the entire asymmetric unit (Terwilliger, & Berendzen, 1999a; Schneider & Sheldrick, 2002). The normalized, truncated density described in section 2.1 was first squared. The squared density was then smoothed by averaging all values within a moving sphere with radius (r) given by the larger of 6 Å or twice the high-resolution limit of the data. The standard deviation (s) of the smoothed squared density was then calculated. To compensate for the effect of the solvent fraction in the crystal (f) on the resulting value, the standard deviation (s) calculated above was multiplied by the factor $[(1-f)/f]^{1/2}$ to yield the contrast c :

$$c = [(1-f)/f]^{1/2} s \quad (3)$$

The correction factor $[(1-f)/f]^{1/2}$ was chosen because it leads to a value of 1 for the contrast for a map for which the entire solvent region has a zero variance and the non-solvent region has a constant and non-zero variance.

Correlation of local rms density. The presence of contiguous flat solvent regions in a map was detected using the correlation coefficient of the smoothed squared electron density calculated as described above, with the same quantity calculated using half the value of the smoothing radius, yielding the correlation of *rms* density, r^2_{RMS} . In this way the local value of the *rms* density within a small local region (typically within a radius of 3 Å) is compared with the local *rms* density in a larger local region (typically within a radius of 6 Å). If there is a large, contiguous solvent region and another large contiguous region containing the macromolecule, the local *rms* density in the small region would be expected to be highly correlated with the *rms* density in the larger region.

On the other hand, if the “solvent” region is broken up into many small flat regions, then this correlation would be expected to be smaller.

Flatness of solvent region. A normalized, truncated electron density map was partitioned between regions of solvent and macromolecule as described in section 2.1. Then the rms electron density in the solvent region ($rms_{SOLVENT}$) and in the region of the macromolecule (rms_{PROT}) were calculated. The flatness (F) of the solvent region was expressed as the difference between the two:

$$F = rms_{PROT} - rms_{SOLVENT} \quad . \quad (4)$$

Number of regions enclosing high density. A threshold of density (t) was found such that 5% of the volume of the asymmetric unit of the crystal had a density greater than this threshold t . All the grid points in the map above the threshold t were marked. Then the number of discrete regions ($N_{regions}$) containing marked points was counted. For this purpose, a discrete region was defined as a set of all marked grid points that can be connected by tracing from one adjacent marked grid point to another (including symmetry-related marked grid points). To partially compensate for the fact that lower-resolution maps have fewer grid points, the number of regions is multiplied by the high-resolution limit of the data used to calculate the map (d_{min}). To further compensate for the volume of the asymmetric unit containing the macromolecule, the number of regions is then divided by the fraction of the asymmetric unit that contains macromolecule (f) and the volume of the asymmetric unit (V) to yield the normalized number of regions per unit volume (N_r):

$$N_r = N_{regions} / (fV) \quad . \quad (5)$$

Overlap of NCS-related density. If non-crystallographic symmetry is found in the heavy-atom substructure for a solution then the map is examined for the presence of correlated density at NCS-related locations in the map (Cowtan & Main, 1998; Vellieux et al., 1995). The overlap (O_{NCS}) between density at NCS-related locations is used to evaluate non-crystallographic symmetry:

$$O_{NCS} = \langle \rho_i \rho_j \rangle , \quad (6)$$

where ρ_i and ρ_j are density at NCS-related locations in the asymmetric unit and the average is either within a sphere with radius r_{smooth} (as described above for identifying the solvent boundary), or over a region within the asymmetric unit. The values of density ρ_i used are those from the normalized truncated map described above. The region where NCS applies is identified as a contiguous region where the local mean of the overlap is at least c_{MIN} , where this cutoff c_{MIN} is selected to yield a total volume occupied by all NCS copies approximately the same as the total volume (f) occupied by the macromolecule in the asymmetric unit (Terwilliger, 2002a). For purposes of evaluating a map, the mean value of the overlap of NCS density, O_{NCS} , is calculated over this entire NCS region. If the value of the overlap found is less than O_{MIN} (typically $O_{MIN}=0.3$), the NCS is ignored.

2.4. Reciprocal-space map-quality measures

R-factor and phase correlation from statistical density modification. The amplitudes and phases of structure factors calculated using statistical density modification, but without including the experimental phase probabilities, can be compared with the observed amplitudes and experimental phases (Cowtan & Main, 1996; Terwilliger, 2001). These comparisons yield an R-value (R_{DENMOD}) for the amplitudes and a mean cosine of the phase difference (m_{DENMOD}) for the phases.

Figure of merit of phasing. The mean figure of merit of phasing ($\langle m \rangle$) was used directly from Phaser (for SAD phasing calculations; McCoy et al., 2004) or SOLVE (for MIR and MAD phasing calculations; Terwilliger & Berendzen, 1999) as an estimate of the quality of a map.

Density truncation (peak-picking). The number of non-hydrogen atoms (n) in the asymmetric unit is roughly estimated from the fraction of the asymmetric unit that contains macromolecule (f) and the volume of the asymmetric unit (V) using an approximate average atomic volume of $V_o=26 \text{ \AA}^3$ ($n=fV/V_o$). Then the highest n grid points in the asymmetric unit of the electron density map are identified and C atoms are placed at these grid points. A map is calculated from these C atoms and the correlation ($r^2_{TRUNCATION}$) with the original map is obtained, after adjusting an overall thermal factor to maximize this correlation

2.5. Bayesian estimates of map quality

A simple Bayesian approach was used to create estimators of map quality based on one or more of the measures of map quality described in sections 2.3 and 2.4. For each measure (e.g.,

skew) the analysis of maps corresponding to solved structures yielded a list of values of “true” map correlation (r^2_{MODEL}) and the measure of quality (e.g., *skew*). A two-dimensional histogram was created to represent the joint distribution $p(r^2_{MODEL}, skew)$. The distributions were sampled with 30 bins for each variable, with a range of allowed values of each ranging from -0.1 to 1.1. Any values obtained outside this range were put in the closest available bin. To compensate for the fact that insufficient data (1359) were present to generate an accurate value for all 900 bins, the values of $p(r^2_{MODEL}, skew)$ were smoothed using a Gaussian smoothing algorithm in which $p(r^2_{MODEL}, skew)$ was convoluted with a Gaussian function $G(r)$ with a radius (σ) of 3 bins ($G(r) \propto \exp\{-(u^2+v^2)/(2\sigma^2)\}$), reducing the effective number of bins to about 100.

To estimate the value of map quality (r^2_{MODEL}) from a new observation of the quality measure (*skew*), Bayes’ rule (Hamilton, 1964) was used:

$$p(r^2_{MODEL} | skew) = A p_o(r^2_{MODEL}) p(skew | r^2_{MODEL}) , \quad (7a)$$

where the normalization factor A assures that the integrated probability for r^2_{MODEL} is unity and is given by,

$$A = 1 / \int_{r^2} [p_o(r^2) p(skew | r^2)] dr^2 . \quad (7b)$$

Eq. (7a) says that the (posterior) probability of a particular value of r^2_{MODEL} , given the measurement *skew*, is the prior probability of r^2_{MODEL} ($p_o(r^2_{MODEL})$) multiplied by the conditional probability ($p(skew | r^2_{MODEL})$) of measuring this value of *skew* given that r^2_{MODEL} is the correct value, divided by a normalization factor. We calculated the conditional probability $p(skew | r^2_{MODEL})$ in Eq. 7a from the joint probability distribution $p(r^2_{MODEL}, skew)$ using the relation,

$$p(skew | r^2_{MODEL}) = p(r^2_{MODEL}, skew) / p_o(r^2_{MODEL}) . \quad (7c)$$

For the present work we assume the prior probability distribution $p_o(r^2_{MODEL})$ is uniform on [0,1].

If several measures of map quality (e.g., *skew* and contrast c) have been measured, then the estimates can be combined using the same approach:

$$p(r^2_{MODEL} | skew, c) = A p_o(r^2_{MODEL}) p(skew, c | r^2_{MODEL}), \quad (8a)$$

$$A = 1 / \int_{r^2} [p_o(r^2) p(skew, c | r^2)] dr^2 . \quad (8b)$$

We approximate the probability distribution $p(skew, c | r^2_{MODEL})$ as the product of the two 2-dimensional conditional probabilities that we have estimated above:

$$p(skew, c | r^2_{MODEL}) \propto p(skew | r^2_{MODEL}) p(c | r^2_{MODEL}) , \quad (9)$$

which amounts to assuming that the skew and contrast c are conditionally independent for a given fixed r^2_{MODEL} value.

To obtain the estimated value and variance of r^2_{MODEL} given a set of observations of predictor variables (e.g., $skew$, c) we used the probability distribution given by Eq. 8a and calculated the expectation value of $\langle r^2_{MODEL} \rangle$:

$$\langle r^2_{MODEL} \rangle = \int_{r^2} p(r^2, | skew, c) r^2, d r^2, \quad (10a)$$

$$\langle \sigma^2 \rangle = \int_{r^2} p(r^2, | skew, c) [r^2, -\langle r^2_{MODEL} \rangle]^2 d r^2, \quad (10b)$$

An improved estimate of the conditional probability distributions such as $p(skew, c | r^2_{MODEL})$ could potentially be obtained by calculating the covariance of the variables $skew$ and c for each fixed value of r^2_{MODEL} and assuming a normal distribution of $skew$ and c for this fixed value of r^2_{MODEL} . This formulation differs from that in Eq. 9 by including correlations between $skew$ and c instead of assuming that they are zero, and also through the assumption of normality in the distributions of $skew$ and c for fixed r^2_{MODEL} . Leaving out the fixed value of r^2_{MODEL} for clarity, representing the two-dimensional vector $(skew, c)$ as $\mathbf{x}=(skew, c)$ and the mean values of $skew$ and c for this value of r^2_{MODEL} as $\mathbf{u}=(\langle skew \rangle, \langle c \rangle)$, we can write (Hamilton, 1964):

$$p(skew, c) \sim \exp \{ -1/2 [\mathbf{x}-\mathbf{u}] \Sigma^{-1} [\mathbf{x}-\mathbf{u}]^T \} / [2\pi \det(\Sigma)] , \quad (10a)$$

where Σ is the covariance matrix with elements σ_{ij} representing the variation of $skew$ and c around their means $\langle skew \rangle$ and $\langle c \rangle$:

$$\sigma_{12} = \sigma_{21} = \langle (skew - \langle skew \rangle) (c - \langle c \rangle) \rangle = cov(skew, c) , \quad (10b)$$

$$\sigma_{11} = \langle (skew - \langle skew \rangle)^2 \rangle = \sigma^2_{skew} , \quad (10c)$$

$$\sigma_{22} = \langle (c - \langle c \rangle)^2 \rangle = \sigma^2_c . \quad (10d)$$

To test this approach we used the data described above, but grouped in bins of r^2_{MODEL} . The observations in each bin of r^2_{MODEL} were analyzed using Eqs. 10a-10d based on the values of the N predictor variables ($skew, c, \dots$) for all the observations in that bin to obtain an approximation of the conditional probability distribution $p(skew, c | r^2_{MODEL})$ for that bin. This set of approximations (one for each bin of r^2_{MODEL}) was then used in Eq. 8 to estimate r^2_{MODEL} for individual sets of observations of the N predictor variables. This approach gave correlations that were at most marginally improved over those obtained using estimates of the conditional probability distribution $p(skew, c | r^2_{MODEL})$ based on Eq. 9. For example, using $skew$ and correlation of local rms density (r^2_{RMS}) as predictor variables, and analyzing the same data shown in Table 3 (but without cross-validation), the overall correlation coefficient between true values of r^2_{MODEL} and estimates using Eq. 9 (in which independence of $skew$ and r^2_{RMS} is assumed) was 0.925. Using Eq. 10 (assuming Gaussian distributions for $skew$ and r^2_{RMS}) and setting the covariance terms to zero (assuming independence of $skew$ and r^2_{RMS}), yielded a value of 0.926, and the same analysis, but

including the covariance terms, yielded a value of 0.927. As this approach did not significantly improve the correlation, it was not used. Fig. 1c suggests that the assumption of normality in the distributions of the predictor variables (e.g., *skew* and r^2_{RMS}) for fixed r^2_{MODEL} is not well-justified. This may partially explain the poor performance of this approach.

2.6. Structures and data used

Data from 47 structures in the PHENIX library of MAD, SAD and MIR datasets were used along with 246 MAD and SAD structures from the Joint Center for Structural Genomics (JCSG, www.jcs.org). The structures from the PHENIX library included 1029B (1N0E, Chen et al., 2004), 1038B (1LQL, Choi et al., 2003), 1063B (1LFP, Shin et al., 2002), 1071B (1NF2, Shin et al., 2003), 1102B (1L2F, Shin et al., 2003b), 1167B (1S12, Shin et al., 2005), aep-transaminase (1M32, Chen et al., 2002), armadillo (3BCT, Huber et al., 1997), calmodulin (1EXR, Wilson & Brunger, 2000), cobd (1KUS, Cheong et al., 2002), cp-synthase (1L1E, Huang et al., 2002), cyanase (1DW9, Walsh et al., 2000), epsin (1EDU, Hyman et al., 2000), flr (1BKJ, Tanner et al., 1996), fusion-complex (1SFC, Sutton et al., 1998), gene-5 (1VQB, Skinner et al., 1994), gere (1FSE, Ducros et al., 2001), gpatase (1ECF, Muchmore et al., 1998), granulocyte (2GMF, Rozwarski et al., 1996), groEL (1OEL, Braig et al., 1995), group2-intron (1KXK, Zhang & Doudna, 2002), hn-rnp (1HA1, Shamo et al., 1997), ic-lyase (1F61, Sharma et al., 2000), insulin (2BN3, Nanao et al., 2005), lysozyme (unpublished results; CSHL Macromolecular Crystallography Course), mbp (1YTT, Burling et al., 1996), mev-kinase (1KKH, Yang et al., 2002), myoglobin (Ana Gonzales, personal communication), nsf-d2 (1NSF, Yu et al., 1998), nsf-n (1QCS, Yu et al., 1999), p32 (1P32, Jiang et al., 1999), p9 (1BKB, Peat et al., 1998), pdz (1KWA, Daniels et al., 1998), penicillopepsin (3APP, James & Sielecki, 1983), psd-95 (1JXM, Tavares et al., 2001), qaprtase (1QPO, Sharma et al., 1998), rab3a (1ZBD, Ostermeier & Brunger, 1999), rh-dehalogenase (1BN7, Newman et al., 1999), rnase-p (1NZ0, Kazantsev et al., 2003), rnase-s (1RGE, Sevcik et al., 1996), rop (1F4N, Willis et al., 2000), s-hydrolase (1A7A, Turner et al., 1998), sec17 (1QQE, Rice & Brunger, 1999), synapsin (1AUV, Esser et al., 1998), synaptotagmin (1DQV, Sutton et al., 1999), tryparedoxin (1QK8, Alphey et al., 1999), ut-synthase (1E8C, Gordon et al., 2001), vmp (1L8W, Eicken et al., 2002).

The structures from the JCSG included PDB entries 1O1X (Xu et al., 2004), 1VJF, 1VJR, 1VK4, 1VK8, 1VK9, 1VKD, 1VKN, 1VL0, 1VL5, 1VLI, 1VLO, 1VLY, 1VM8, 1VMG, 1VMI, 1VP8, 1VPM, 1VPZ (Rife et al., 2005), 1VQR (Xu et al., 2006), 1VQS, 1VQY, 1VQZ, 1VR0 (DiDonato et al., 2006), 1VR3 (Xu et al., 2006), 1VR5, 1VR8 (Xu et al., 2006), 1VRM (Han et

al., 2006), 1Z82, 1Z85, 1ZBT, 1ZEJ, 1ZH8, 1ZKO, 1ZTC, 1ZX8 (Jin et al., 2006), 1ZY9, 1ZYB, 2A3N, 2AAM, 2AML, 2AX3, 2B8N (Schwarzenbacher et al., 2006), 2ETD, 2ETS, 2EVR, 2F4I, 2F4L, 2FG0, 2FG9, 2FNA, 2FTR, 2FUP, 2FUR, 2G0W, 2GB5, 2GC9, 2GF6, 2GFG, 2GHR (Zubieta et al., 2007), 2GNO, 2GO7, 2GPI, 2GPJ, 2GRJ, 2GVH, 2H1Q, 2H1T, 2H9F, 2HCF, 2HH6, 2HHZ, 2HI0, 2HQ7, 2HQ9, 2HR2, 2HSZ, 2HUH, 2HX1, 2HX5, 2HXV, 2I02, 2I8D, 2I9W, 2IG6, 2II1, 2ILB, 2ISB, 2IT9, 2ITB, 2NUJ, 2O08, 2O2G, 2O2X, 2O2Z, 2O3L, 2O62, 2OA2, 2OAF, 2OC6, 2OD5, 2OGI, 2OH1, 2OH3, 2OIK, 2OOJ, 2OOK, 2OP5, 2OPL, 2OQM, 2ORD, 2OSD, 2OTM, 2OU3, 2OU5, 2OU6, 2OWN, 2OYO, 2OZG, 2OZJ, 2P10, 2P1A, 2P7I, 2P8J, 2PBL, 2PEB, 2PFW, 2PG4, 2PGC, 2PKE, 2PN1, 2PQ7, 2PR7, 2PRR, 2PRV, 2PV4, 2PV7, 2PWN, 2PY6, 2PYQ, 2PYX, 2Q02, 2Q04, 2Q0T, 2Q14, 2Q3L, 2Q78, 2Q7X, 2Q9K, 2Q9R, 2QE6, 2QE9, 2QEZ, 2QG3, 2QHP, 2QJ8, 2QL8, 2QML, 2QPX, 2QR6, 2QTP, 2QTQ, 2QW5, 2QWW, 2QWZ, 2QYV, 2R01, 2R0X, 2R1I, 2R3B, 2R44, 2R4I, 2R9V, 2RA9, 2RAS, 2RCC, 2RCD, 2RD9, 2RDC, 2RE3, 2RE7, 2RFP, 2RGQ, 2RHA, 2RHM, 2RIJ, 2RIL, 2RKH, 3B5E, 3B5O, 3B77, 3B7F, 3B81, 3B8L, 3BB5, 3BB9, 3BCW, 3BDD, and 3BDE.

3. Results and Discussion

3.1. Measures of map quality

A key goal of this work was to identify one or more quality measures of maps or of structure factors that are simple to calculate and that can yield accurate estimates of the qualities of the corresponding electron density maps. Table 1 lists 6 measures of map quality examined here that are based on the features in the maps (real-space measures), and Table 2 lists 4 additional measures we have examined that depend on the structure factors and phases used to calculate maps. The measures we have examined were chosen to represent a range of possible measures that cover many important features of electron density maps and structure factors.

To evaluate possible measures of map quality, we carried out a re-analysis of data for 246 previously-solved MAD, SAD and MIR structures, creating electron density maps during the structure-determination process and analyzing them with each of the measures in Tables 1 and 2. As the structures are all known, the “true” map quality for each map could be calculated as the correlation coefficient r^2_{MODEL} between each map and the corresponding map obtained from the refined model of the structure (after any necessary origin shifts are applied) using the *PHENIX* tool *phenix.get_cc_mtz_mtz*. Figures 1A through 1J show the values of each measure plotted against r^2_{MODEL} for 1359 maps based on structures calculated from the MAD, SAD, and MIR data listed in section 2.6. The maps represent phases obtained at several stages in structure determination. Some are calculated using heavy-atom solutions found from anomalous or

isomorphous differences or from F_A values with HYSS (Grosse-Kunstleve & Adams, 2003). Others are calculated using the corresponding substructures with inverted hand. Others are obtained from difference Fourier (MIR) and anomalous difference Fourier (MAD) analyses. In the case of MIR, a large number of additional solutions are obtained by combinations of partial solutions from different derivatives.

The general features of the plots in Fig. 1 are illustrated by a discussion of Fig. 1A, which shows the *skew* of electron density in experimental maps as a function of the true map quality, r^2_{MODEL} . In Fig. 1A the purple squares correspond to datasets with a nominal resolution lower than 2 Å, and the black diamonds to datasets with resolutions of 2 Å or higher. (Note that the data for all these calculations are truncated at a resolution of 2.5 Å, so that most resolution-dependent differences are likely to be due to dataset-dependent decreases of intensities with resolution, rather than the resolution of the data.)

Fig. 1A shows that the skew of the electron density depends strongly on the map quality, as represented by the correlation of the density in the map with that of a model map (r^2_{MODEL}). The skew is approximately zero for maps with a correlation in the range of $0.0 < r^2_{MODEL} < 0.2$. It increases slightly for maps with correlations in the range of $0.2 < r^2_{MODEL} < 0.4$, and then it increases substantially for maps with higher correlations ($r^2_{MODEL} > 0.4$). The standard deviation of values of the skew is about 0.05-0.10 over most ranges of map correlation. For example, for values of map correlation with $r^2_{MODEL} < 0.2$, the mean skew is -0.02 and the standard deviation is 0.07, and for values of map correlation with $0.4 < r^2_{MODEL} < 0.5$ the mean skew is 0.14 with standard deviation of 0.06. For values of map correlation with $0.6 < r^2_{MODEL} < 0.7$, the mean skew is 0.38 with standard deviation of 0.10. Another way to view these relationships is to note that the difference (0.16) in mean values of the skew between values of map correlation of $r^2_{MODEL} < 0.2$ and values of map correlation in the range of $0.4 < r^2_{MODEL} < 0.5$ is about twice the standard deviation of the skew in either range. This means that the skew can be expected to differentiate between maps with model correlations r^2_{MODEL} of zero and 0.4, but that cannot differentiate them correctly all of the time. This can also be seen directly from Fig. 1A, in which some of the values of skew for maps with model correlations r^2_{MODEL} near 0.4 are lower than values for maps with near-zero values of r^2_{MODEL} .

The maps represented in Fig 1A that are based on high-resolution datasets (< 2 Å) have values of skew that are similar to those of lower-resolution datasets. This similarity most likely reflects the fact that all the data in these calculations are truncated at a resolution of 2.5 Å.

Several of the other 9 measures of map quality examined have relationships to model map correlation similar to those of the skew described above. The contrast (c , Fig. 1B), correlation of

local *rms* density (r^2_{RMS} , Fig. 1C), and flatness of the solvent region (F , Fig. 1D) in particular show very similar behaviour, except that neither discriminates as well as the skew between maps of moderate quality (correlations r^2_{MODEL} near 0.4) and those of very low quality with correlations near zero. These three measures are all related as they all are based on the presence of solvent and non-solvent regions in the crystal. The calculations differ, however, in that the contrast (c) does not require knowledge of the solvent boundary while the flatness (F) does. Additionally the correlation of local *rms* density reflects the contiguous nature of the solvent region while the contrast (c) and flatness (F) reflect the presence of a solvent region, whether contiguous or not.

A somewhat different behaviour is shown by the number of contiguous regions (N_r) required to enclose the highest 5% of density in a map (Fig. 1E). This measure decreases with increasing map quality, but only slightly, so that it is not a strong discriminator between maps of low and moderate quality.

The overlap of NCS-related density (Fig. 1F) is a measure which, as implemented here, only applies to maps where NCS can be identified from the symmetry present in the heavy-atom sites. It is therefore different from the measures discussed so far and cannot be used as a general measure of map quality. It is nevertheless useful in differentiating between maps of very high model map correlations (r^2_{MODEL}) and those that have lower model map correlation.

Figs. 1G and 1H show the phase correlations (m_{DENMOD}) and R-factors (R_{DENMOD}) obtained from the first cycle of statistical density modification using the same structure factors, phases, and weights that are used to calculate electron density maps analyzed in Figs. 1A-1F. In the first cycle of statistical density modification (Terwilliger, 2000) estimates of the phase and amplitude of a reflection k are obtained using only information from all the *other* reflections in the dataset. The amplitude and phase for reflection k from the density modification procedure can then be compared with the experimentally observed amplitude and the “experimental” phase (derived using isomorphous or anomalous differences) to yield an R-factor for density modification (R_{DENMOD}) and a mean cosine of the phase difference (m_{DENMOD}). Figure 1G shows that, as expected, the R-factor for density modification decreases with increasing map quality, while Fig. 1H shows that the phase correlation increases over the same range.

Fig. 1I shows that the correlation of pseudo-maps calculated using dummy atoms placed at the highest peaks in a map with their corresponding original maps ($r^2_{TRUNCATION}$) is weakly related to the quality of the map. It seems possible that more sophisticated methods of map skeletonization (Baker et al., 1993) might be more useful in map evaluation than our simple measure.

Finally, Fig. 1J shows that the mean figure of merit of phasing ($\langle m \rangle$) is related to the quality of the map, but that there are many maps with very low correlation to the corresponding model

maps that nevertheless have high mean figures of merit. This complex relationship can be understood by considering that the figures of merit of phasing of two maps that are calculated using the same data, but opposite enantiomers of the heavy-atom substructure, are normally identical for SAD phasing if all the anomalous scatterers are of the same type. Typically one of these maps may have a high correlation to the model map, while the other may have a very low correlation.

Overall, Fig. 1 shows that several measures of map quality based on different features of the map and of structure factors and phases leading to the map have strong relationships to the quality of the electron density map, with the skew of electron density clearly being one of the best indicators of map quality.

3.2. Estimation of map quality using features of the map and of structure factors used to calculate the map

Figure 1 showed that each of the 6 different features of electron density maps and 4 characteristics of structure factors we examined depend in some way on the quality of the corresponding map. We used the Bayesian approach described in section 2.5 to use this information to estimate map quality from these 10 features. The general idea of this approach is very simple. Imagine that a particular map has been examined, yielding a skew of 0.20. Based on Fig. 1A, it is reasonable to conclude that this map is very likely to have a correlation (r^2_{MODEL}) with the corresponding model map in the range of $0.4 < r^2_{MODEL} < 0.6$, because nearly all examples in Fig. 1A with a skew of about 0.20 are in this range. Equation 7a is simply a mathematical way to make this statement. Eq. 8a is a similar statement, except that it includes more than one measure of map quality. As described in section 2.5, we assume here that the various measures of map quality (skew, contrast, etc.) are independent. This allows a very simple calculation (Eq. 8a) to be used to estimate r^2_{MODEL} from several measures of map quality.

Fig. 2A shows the results of using Eq. 7a to estimate r^2_{MODEL} from the skew of electron density. In Fig. 2A the abscissa is the Bayesian estimate of r^2_{MODEL} using the skew of electron density, and the ordinate is the true value of r^2_{MODEL} . To ensure that the parameters in the Bayesian estimator did not contain information on the specific cases being tested, a jackknife procedure was used in which all solutions for the structure being examined were excluded when constructing the Bayesian estimators. Fig. 2A shows that in cases where the true value of r^2_{MODEL} is in the range of $0.0 < r^2_{MODEL} < 0.2$, the estimates of r^2_{MODEL} all have very similar values of about 0.1. This can be understood from Fig. 1A, in which the skew is seen to be insensitive to values of r^2_{MODEL} in this range. The Bayesian estimates of r^2_{MODEL} for low values of skew are all close to

the midpoint of this range, as they are simply the average of plausible values of r^2_{MODEL} , given the observation of the value of the skew. For higher values of r^2_{MODEL} , the estimates of r^2_{MODEL} are closer to the true values. Overall, the correlation coefficient between the Bayesian estimates and true values of r^2_{MODEL} is 0.90 and the *rms* error in prediction of r^2_{MODEL} is 0.10. As a check on our procedures, we note that the mean uncertainty estimates for r^2_{MODEL} obtained from the Bayesian procedure was 0.11, quite similar to the actual *rms* error in prediction of r^2_{MODEL} of 0.10.

Table 3 summarizes the accuracy of the Bayesian estimates of map quality based on each of the measures described in Tables 1 and 2 (with the exception that the overlap of NCS density is not included because it does not apply to most of the maps in our tests). For each measure, Table 3 lists the values of the correlation coefficient of the Bayesian estimates and the true map quality (r^2_{MODEL}) along with the *rms* prediction error in r^2_{MODEL} . Overall, the skew of electron density, having a correlation coefficient between Bayesian estimates and true values of r^2_{MODEL} is 0.90, is the most reliable indicator of map quality, with the correlation of local *rms* density next best (correlation of 0.85), and with contrast, flatness of solvent region, and density-modification phase correlations and R-factor giving only slightly poorer predictions of r^2_{MODEL} with correlations in the range of 0.75-0.80.

To identify an optimal combination of measures for estimation of map quality, we began with the best single measure (skew) and used Eq. 9 to combine information from each of the other measures. The measure giving the best prediction of r^2_{MODEL} in combination with skew was the correlation of local *rms* density (r^2_{RMS} , Table 3). Figure 2B shows how the estimates of map quality obtained using just the correlation of *rms* electron density compare with actual map quality, and Fig. 2C shows estimates based on both skew and correlation of *rms* electron density. The correlation of *rms* density was the next-best single predictor after *skew* and in addition the correlation of prediction errors from these two variables was relatively low (0.61, Table 4). The assumptions in Eq. 9 are therefore relatively well-justified and it is not surprising that the resulting estimator is improved over the one using just the skew of electron density. This process was continued but no further improvement was obtained in the Bayesian estimator. The optimized combination of measures based on skew and correlation of local *rms* density yielded a correlation coefficient between the Bayesian estimates and true values of r^2_{MODEL} of 0.92 and an *rms* prediction error of 0.09 (Table 3 and Fig. 2C).

3.3. Identification of the hand of heavy-atom substructures using measures of map quality

A particularly important application of measures of map quality is the identification of the hand of heavy-atom substructures. In space groups that are not enantiomorphic, the hand of the heavy-atom substructure can normally not be identified directly during substructure determination by direct methods such as the HYSS procedure (Grosse-Kunstleve & Adams, 2003) used here. Consequently some procedure is needed for identifying which hand of the heavy-atom substructure is correct. Figures 3A through Fig. 3I compare the values obtained for 9 measures of map quality based on 353 pairs of heavy-atom substructures with correct and inverted handedness from the 186 datasets in this work for which the space group was not chiral. The mean figure of merit of phasing is not shown because it is essentially identical for the two hands of the substructure in all the cases examined. The 706 maps represented by these 353 pairs are a subset of the 1359 maps used in the calculations shown in Fig. 1.

It is somewhat remarkable that these 9 measures of map quality all give very good discrimination between the correct and incorrect hands of heavy-atom substructures (Fig. 3 and Table 5), even though they are not all so useful in estimating the absolute quality of maps (Table 3). The best discrimination between correct and incorrect hands is obtained with the skew of electron density (Fig. 3A), as expected from the high correlation of estimates of map quality based on skew with actual map quality (Table 3). Using the skew of electron density to make decisions on handedness (Fig. 3A), 98% of decisions (in cases where the quality of the maps for the two hands differs by at least 0.05) would lead to a map with higher quality than that of the opposite hand (Table 5). Note that for SIR or MIR data without anomalous differences, none of these techniques can identify the correct hand because the inverse hand of the heavy atoms leads to a map that has inverse chirality but is otherwise identical. A similar argument would partially apply in cases where the anomalous signal is weak. This situation is presumably the cause of the large number of MIR-derived points along the diagonal of the panels in Fig. 3.

3.4. Identification of the highest-quality density modified map for a structure

The scoring procedures described above are based on an analysis of the phases and structure factor amplitudes corresponding to an experimental electron density map. Prior to final map interpretation, however, the experimentally-determined phases of structure factors are normally optimized by density modification (Wang, 1985). It seemed possible that the best experimental maps would not always lead to the best density-modified maps, and consequently that some additional method of scoring the density modified maps might be useful.

To investigate this possibility, we carried out automated structure determination using the datasets used in Fig. 1, this time with default parameters in the AutoSol Wizard, including Bayesian estimates of experimental map quality based on the skew of electron density (*skew*) and the correlation of local *rms* density (r^2_{RMS}). For each structure, the final steps were to carry out density modification on the top-ranked solution or solutions and then to build a preliminary atomic model. In cases where there was one solution that was much better than all others (see Methods), then only that solution was used in density modification. However in most cases there were multiple solutions with similar Bayesian estimates of quality and up to 3 (MAD, SAD) or 6 (MIR) of these were used in density modification.

Figure 4A shows the relationship between qualities of experimental maps and the qualities of the corresponding density-modified maps for 545 experimental maps for 240 datasets. For experimental maps of high quality (correlation with model map over 0.6), the quality of the density-modified map is generally (but not always) very high, typically ranging from 0.75 to 0.90. On the other hand, for experimental maps of low or moderate quality (map correlation of less than about 0.5), there is remarkably little correspondence between the quality of the experimental map and that of the density-modified map.

Part of the variability in density modification illustrated in Fig. 4A could be due to the differences in solvent content, non-crystallographic symmetry, type of experiment and resolution between the different structures. To examine this we have plotted in Fig. 4B the true map qualities of density-modified maps for all 206 pairs of solutions from Fig 4A that are from the same structure and that have values of true experimental map correlation within 0.05 of each other. In Fig. 4B each point corresponds to one pair of solutions. The abscissa is the value of density-modified map quality for the solution with the higher value of experimental map quality and the ordinate is the density-modified map quality for the solution with lower experimental map quality. Each member of such a pair has identical solvent content, resolution, non-crystallographic symmetry and experiment type, and differs only slightly in true experimental map quality. Fig. 4B shows that even when all these factors are controlled there is considerable variability in the quality of the density-modified map. Furthermore, for pairs of solutions with similar experimental map qualities, the solution with higher experimental map quality does not necessarily lead to the better density-modified map. For example, the point in Fig. 4B at (0.55, 0.89) corresponds to a pair of solutions from the MAD structure 2QML, at a resolution of 1.55 Å, with no non-crystallographic symmetry and a solvent content of 0.55. These solutions have true experimental map qualities of 0.42 and 0.37, respectively, where the solution with the slightly

lower experimental map quality (map correlation of 0.37) has led to the better density-modified map (map correlation of 0.90).

The variation in effects of density modification illustrated in Fig. 4B suggests that it might be useful to carry out a final ranking of solutions based on a measure of quality of the corresponding density-modified maps. We used the map-model correlation between density-modified maps and the preliminary atomic models built with the AutoSol Wizard as such a measure of quality. Table 6 shows the utility of this map-model correlation in identifying the solution with the best density-modified map for each of the 134 structures used in Fig. 4A in which there was more than one solution tested by density modification and model-building, and in which the model-building process yielded a model with a model-map correlation of at least 0.20. The first row in Table 6 provides a background for this analysis by considering the use of our Bayesian estimates of experimental map quality to identify the best solutions. Using the Bayesian estimates (based on the experimental maps) the best experimental map for a particular structure could be identified 92% of the time. Furthermore the worst error in identification of the best map corresponded to a difference in map correlation of only 0.16. On the other hand, the solution with the highest Bayesian estimate of experimental map quality led to the best density-modified map only 57% of the time, and this density modified map had a true map correlation as much as 0.64 lower than the best density modified map for the corresponding structure.

Using the map-model correlation for the model built into the density-modified maps, the situation is reversed, with the best experimental map identified only 61% of the time and the best density-modified map identified 70% of the time. Most importantly, the density-modified map yielding the highest map-model correlation was never worse than the very best density-modified map obtained by more than a difference in correlation of 0.09, indicating that the model-map correlation is a useful criterion for final ranking of solutions.

3.5. Using the AutoSol Wizard to redetermine structures from the PHENIX structure library

To test the utility of the Bayesian estimates of map quality obtained using the skew and correlation of local *rms* density as described in section 3.2, we carried out structure determinations on all 48 MAD, SAD, and MIR structures in the PHENIX structure library and used these quality estimates to make decisions about which solutions to pursue. The structures in this library range from relatively straightforward cases of SAD and MAD structure determination to considerably more complex cases that involved combinations of SAD or MAD with MIR. In the automated tests carried out here, only one source of phase information was used for each

structure (i.e., MAD, SAD, or MIR) except in the case of the fusion-complex structure (1SFC, Sutton et al., 1998) in which SAD and SIR data were combined. We compared the qualities of the maps obtained after density-modification from this automated procedure using two methods of making decisions. The first method was to use the Bayesian estimates based on the combination of the skew of electron density and the correlation of local *rms* density, as described above. The second method was to use a decision-making process using perfect scores in which the actual correlation coefficient of each map with that of the corresponding model map was used to decide which map was best. Figure 5A illustrates these comparisons for MAD structure determinations, Fig. 5B illustrates them for SAD structure determinations, and Fig. 5C for MIR structure determinations.

For MAD and SAD structure determinations the decision-making procedure using Bayesian estimates based on the combination of the skew of electron density and the correlation of local *rms* density led to density-modified electron density maps that were of comparable quality to those obtained using a perfect decision-making process (Fig. 5). In the case of fusion-complex, the Bayesian decision-making procedure led to a slightly better density-modified map than a procedure using the actual quality of experimental maps for decision-making. This occurred because a solution with the best experimental map led to a density-modified map that was not quite the best. For MIR structure determinations the decision-making process was not as good. In several MIR cases the final maps obtained using the Bayesian estimates were substantially poorer than obtained using perfect map correlation. The AutoSol Wizard failed, using either method of decision-making, to find a solution in one difficult case (groEL; Braig et al., 1995) that was previously solved by MIR. In this case heavy-atom solutions could not be automatically found for any of the derivatives.

4. Conclusions

Each of the 10 measures of quality of experimental electron density maps evaluated here has some utility in estimating the true quality of these maps. These measures of map quality have a wide range of bases (Tables 1 and 2) ranging from the flatness of the solvent region typically found in macromolecular structures to the connectivity of regions of high electron density corresponding to the chains of polymers in these structures. Overall, however, the skew of electron density stands out as the best of these measures (Table 3 and Fig. 2). Used in a simple Bayesian estimator, the correlation between map quality estimated with the skew of electron density with true map quality is about 0.90, while the next-best estimator (correlation of local *rms* density) gives a correlation of only 0.85. Combining the two yields the most useful estimator we

have developed, with a correlation between estimated and actual map quality of 0.92 and an *rms* prediction error in map quality of 0.09.

With the exception of mean figure of merit of phasing, which does not depend on the hand of the heavy-atom substructure, all the measures of map quality analyzed are remarkably good discriminators between maps calculated using the correct and inverse hands of the heavy-atom substructure (Fig. 3). Using the combination of skew of electron density and correlation of local *rms* density in a Bayesian estimator of map quality, the AutoSol Wizard is able to carry out automated structure solution. The AutoSol Wizard makes decisions about the heavy-atom substructures to pursue based on these map quality estimates. This process yields density-modified electron density maps of approximately the same overall quality as those obtainable with a perfect decision-making system (Fig. 5).

Our Bayesian estimates of map quality, while highly useful in evaluating experimental maps, are nevertheless not the best indicators of the quality of the corresponding density modified maps. The map-model correlation obtained after preliminary model-building is a considerably better indicator of the quality of density modified maps (Fig. 4 and Table 6).

In this work we have ignored the resolution-dependence of the measures of map quality. This is made possible in part by the use of a high-resolution cutoff of 2.5 Å for all the calculations of map quality and is generally justified by the relatively small remaining resolution dependence of most of the measures of map quality (Fig. 1). Nevertheless it seems possible that some improvement in estimation of map quality might be obtained by including the resolution dependence (or the effective overall isotropic displacement factor) of the data in the analysis. Additionally, we have assumed independence of the various measures of map quality in Eq. 8a. We were not able to improve the estimates of map quality using a simple covariance-matrix approach to combining estimates of map quality, but other more sophisticated approaches, along with a much greater set of sample data, might also lead to improved estimates of map quality.

Figure 1 Measures of quality of electron-density maps and structure factors. Each measure of quality is calculated as described in the text for 1359 sets of structure factors and associated maps. Each measure is plotted with an abscissa based on the correlation of density of the map with a map calculated from a final model (r^2_{MODEL}). Measures based on structures determined at resolutions of 2 Å or higher resolution are shown as black diamonds and those at lower resolution than 2 Å are shown as purple squares. All measures of quality and the correlation with model density (r^2_{MODEL}) are calculated at a resolution of 2.5 Å or the nominal resolution of the data, whichever is the lower resolution. A. Skew of electron density. B. Contrast of electron density. C. Correlation of local rms density. D. Flatness of solvent region. E. Number of

regions enclosing high density. F. Overlap of NCS-related density. G. Phase correlation from statistical density modification. H. R-factor from statistical density modification. I. Density truncation. J. Figure of merit of phasing.

Figure 2 Comparisons of jackknifed estimates of map quality with actual map quality. Measures of map quality as shown in Fig. 1 were used in Eqs. 7a and 8a to estimate overall map quality. The calculations were carried out one dataset at a time. For each dataset, joint probability distributions of each measure of quality and true quality (e.g., $p(\textit{skew}, r^2_{MODEL})$) were calculated excluding data from all solutions for that structure. Then these jackknifed joint probability distributions were used in Eqs. 7a and 8a to estimate map quality using the measures of quality for each map associated with that dataset. In each case true map quality (r^2_{MODEL}) is plotted as a function of the Bayesian estimates of map quality. A. Estimates of map quality using the skew of electron density in Eq. 7a. B. Estimates using the correlation of local *rms* density in Eq. 7a. C. Estimates using the skew and correlation of local *rms* density in Eq. 8a.

Figure 3 Comparisons of measures of map quality for pairs of maps based on enantiomorphic heavy-atom substructures. For structures in non-chiral space groups, all pairs of solutions derived from enantiomorphic pairs of heavy-atom substructures were selected. The member of the pair leading to the map with the higher correlation coefficient to the corresponding model map was identified as the “correct” hand, and the other as the “inverse” hand. The value of each measure of map quality for the correct hand is plotted as the abscissa in each plot, and the value of the measure for the corresponding inverse hand is the ordinate. Maps based on MAD data are represented as black diamonds, those from MIR data (note that all the pairs are actually single derivatives) are red triangles, and those from SAD data are blue squares. A. Skew of electron density. B. Contrast of electron density. C. Correlation of local *rms* density. D. Flatness of solvent region. E. Number of regions enclosing high density. F. Overlap of NCS-related density. G. Phase correlation from statistical density modification. H. R-factor from statistical density modification. I. Density truncation.

Figure 4 Map qualities of density-modified maps. A. Qualities of density-modified maps as a function of the qualities of the corresponding experimental maps. B. Comparison of qualities of pairs of density-modified maps for the same structure (see text).

Figure 5 Comparison of quality of density-modified maps obtained using the skew of electron density and correlation of local *rms* density for scoring with those obtained using the true map quality (correlation to the corresponding model map) for scoring. See text for details. The light blue bars labelled “Perfect scoring” correspond to running the AutoSol Wizard and using the actual map quality to make decisions at each step. The dark maroon bars labelled “Bayesian scoring” correspond to using the Bayesian scores based on the skew of electron density and correlation of local *rms* density. A. Structures determined using MAD. Structures shown are: aep-transaminase (1M32, Chen et al., 2002), armadillo (3BCT, Huber et al., 1997), cobd (1KUS, Cheong et al., 2002), cp-synthase (1L1E, Huang et al., 2002), cyanase (1DW9, Walsh et al., 2000), epsin (1EDU, Hyman et al., 2000), gene-5 (1VQB, Skinner et al., 1994), gere (1FSE, Ducros

et al., 2001), gpatase (1ECF, Muchmore et al., 1998), group2-intron (1KXK, Zhang & Doudna, 2002), ic-lyase (1F61, Sharma et al., 2000), lysozyme (unpublished results; CSHL Macromolecular Crystallography Course), mbp (1YTT, Burling et al., 1996), mev-kinase (1KKH, Yang et al., 2002), nsf-d2 (1NSF, Yu et al., 1998), p32 (1P32, Jiang et al., 1999), p9 (1BKB, Peat et al., 1998), pdz (1KWA, Daniels et al., 1998), psd-95 (1JXM, Tavares et al., 2001), rab3a (1ZBD, Ostermeier & Brunger, 1999), s-hydrolase (1A7A, Turner et al., 1998), synapsin (1AUV, Esser et al., 1998), tryparedoxin (1QK8, Alphey et al., 1999), vmp (1L8W, Eicken et al., 2002) B. Structures determined using SAD: 1029B (1N0E, Chen et al., 2004), 1038B (1LQL, Choi et al., 2003), 1063B (1LFP, Shin et al., 2002), 1071B (1NF2, Shin et al., 2003), 1102B (1L2F, Shin et al., 2003b), 1167B (1S12, Shin et al., 2005), mase-p (1NZ0, Kazantsev et al., 2003), calmodulin (1EXR, Wilson & Brunger, 2000), fusion-complex (1SFC, Sutton et al., 1998), insulin (2BN3, Nanao et al., 2005), myoglobin (Ana Gonzales, personal communication), nsf-n (1QCS, Yu et al., 1999), sec17 (1QQE, Rice & Brunger, 1999), ut-synthase (1E8C, Gordon et al., 2001). Note that fusion-complex was solved with SAD plus SIR. C. Structures determined using MIR: flr (1BKJ, Tanner et al., 1996), granulocyte (2GMF, Rozwarski et al., 1996), groEL (1OEL, Braig et al., 1995), hn-rnp (1HA1, Shamoo et al., 1997), penicillopepsin (3APP, James & Sielecki, 1983), qaprtase (1QPO, Sharma et al., 1998), rh-dehalogenase (1BN7, Newman et al., 1999), mase-s (1RGE, Sevcik et al., 1996), rop (1F4N, Willis et al., 2000), synaptotagmin (1DQV, Sutton et al., 1999).

Table 1 Real-space measures of map quality tested in this work

Method	Symbol	Basis	Expected properties	
			Perfect map	Random map
Skew of electron density	$skew$	High positive density and no negative density in a good map	Positive skew	Near-zero skew
Contrast of electron density	c	Solvent and macromolecule have different <i>rms</i> densities	High contrast	Low contrast
Correlation of local <i>rms</i> density	r_{RMS}^2	Solvent region is contiguous so local <i>rms</i> is correlated with neighboring local <i>rms</i>	Low correlation	High correlation
Flatness of electron density	F	Solvent region has nearly-flat electron density	High value of flatness	Low value of flatness
Number of regions enclosing high density	N_r	Chains of a macromolecule can be represented by a few connected regions of density	Few (but extended) connected regions	Many short connected regions
Overlap of NCS-related density	O_{NCS}	If NCS is present, NCS-related density is similar	High overlap	Low overlap

Table 2 Reciprocal-space measures of map quality tested in this work

Method	Symbol	Basis	Expected properties	
			Perfect map	Random map
Phase correlation from statistical density modification	m_{DENMOD}	Phases from first cycle of density modification are unbiased and are correlated with experimental phases	High m_{DENMOD}	Low m_{DENMOD}
R-factor from statistical density	R_{DENMOD}	Amplitudes for a reflection can be	Low R_{DENMOD}	High R_{DENMOD}

modification		calculated from phases and amplitudes of all other reflections and expected features of the map		
Density truncation	$r^2_{TRUNCATION}$	Much of the information in a map of a macromolecule consists of the density at points in the map near atomic positions	High $r^2_{TRUNCATION}$	Low $r^2_{TRUNCATION}$
Mean figure of merit of phasing	$\langle m \rangle$	Estimates of accuracy of experimental phases are an approximate upper bound on quality of the map	High $\langle m \rangle$	Low $\langle m \rangle$

Table 3 Cross-validated prediction correlation

Quality measure(s)	Prediction correlation coefficient	Rms prediction error
<i>skew</i>	0.90	0.10
<i>c</i>	0.78	0.15
r^2_{RMS}	0.85	0.12
<i>F</i>	0.80	0.14
N_r	0.42	0.20
m_{DENMOD}	0.80	0.10
R_{DENMOD}	0.77	0.14
$r^2_{TRUNCATION}$	0.48	0.21
$\langle m \rangle$	0.42	0.21
<i>skew</i> and r^2_{RMS}	0.92	0.09

Table 4 Correlation of prediction errors*

	<i>skew</i>	<i>c</i>	r^2_{RMS}	<i>F</i>	N_r	m_{DENMOD}	R_{DENMOD}	$r^2_{TRUNCATION}$	$\langle m \rangle$
<i>skew</i>	1								
<i>c</i>	0.69	1							
r^2_{RMS}	0.60	0.82	1						
<i>F</i>	0.73	0.95	0.84	1					
N_r	0.61	0.86	0.61	0.79	1				
m_{DENMOD}	0.63	0.81	0.79	0.88	0.66	1			
R_{DENMOD}	0.66	0.79	0.74	0.79	0.77	0.84	1		
$r^2_{TRUNCATION}$	0.54	0.82	0.63	0.71	0.88	0.61	0.76	1	
$\langle m \rangle$	0.55	0.73	0.61	0.68	0.69	0.64	0.70	0.85	1

*Values of r^2_{MODEL} were estimated for each measure of map quality using Eq. 7a as in Fig. 3. Then the true values of r^2_{MODEL} were subtracted, yielding prediction errors for each map for each measure of map quality. The correlation coefficients (r^2) of prediction errors among the various measures of map quality are listed.

Table 5 Decision-making accuracy* for enantiomeric pairs

Quality measure(s)	Percentage of correct predictions
<i>skew</i>	0.98
<i>c</i>	0.94
r^2_{RMS}	0.95
<i>F</i>	0.94
N_r	0.95
O_{NCS}	0.90
m_{DENMOD}	0.93
R_{DENMOD}	0.94
$r^2_{TRUNCATION}$	0.97

* The percentage of cases in which the higher (or lower, as appropriate) value of the quality measure is associated with the higher value of the actual map correlation coefficient with the corresponding model map. Only cases in which the actual map correlations differ by at least 0.05 are considered.

Table 6 Decision-making accuracies in choosing the solution with the best experimental or density-modified map*

Quality measure	Percentage of correct predictions of best maps		Worst error in identification of best maps	
	Experimental maps	Density-modified maps	Experimental maps	Density-modified maps
<i>Bayesian estimate using skew and r^2_{RMS} of experimental map</i>	92	57	0.16	0.64
<i>Map-model correlation for model built into density-modified map</i>	61	70	0.31	0.09

* The percentage of correct predictions of best maps is the percentage of cases in which the highest value of the quality measure is associated with the highest value of the actual map correlation coefficient with the corresponding model map. The analysis is based on 331 sets of structure factors and associated maps obtained from 134 datasets as in Fig. 1, selecting the top-ranked 2 to 6 solutions and carrying out density modification with RESOLVE (Terwilliger, 2002) to yield density-modified maps. A model was built into each density-modified map using a rapid method for building helices and strands. If the value of the map-model correlation was less than 0.35 then the building procedure was repeated with a standard cycle of building using the methods in the PHENIX AutoBuild Wizard (Terwilliger et al., 2007) and the value map-model correlation from the full standard procedure was used. Only structures for which at least one model-map correlation was at least 0.20 are included in the analysis. The worst error in identification of best maps is the largest value of the difference between the correlation coefficient of the best map with the corresponding model map and that of the map with the highest value of the quality measure.

Acknowledgements The authors would like to thank the NIH Protein Structure Initiative for generous support of the PHENIX project (1P01 GM063210). This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231. RJR is supported by a Principal Research Fellowship from the Wellcome Trust (UK). The authors are grateful to the Joint Center for Structural Genomics for making raw data available at www.jcsg.org and to the many researchers who contributed their data to the PHENIX structure library.

References

- Adams, P. D., McCoy, A. J., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C (2002). *Acta Cryst. D* *Acta Cryst. D58*, 1948-1954.
- Alphey, M.S., Leonard, G.A., Gourley, D.G., Tetaud, E., Fairlamb, A.H. & Hunter, W.N. (1999). *J. Biol. Chem.* *274*, 25613-25622.
- Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* (1993). *D49*, 429-439.
- Blow D. M. & Crick. F. H. C. (1969). *Acta Cryst.* *12*, 794-802.
- Braig, K., Adams, P.D. & Brunger, A.T. (1995). *Nat. Struct. Biol.* *2*, 1083-1094.
- Burling, F.T., Weis, W.I., Flaherty, K.M. & Brunger, A.T. (1996). *Science* *271*, 72-77.
- Chen, C.C.H., Zhang, H., Kim, A.D., Howard, A., Sheldrick, G.M., Mariano-Dunnaway, D. & Herzberg, O. (2002). *Biochemistry* *41*, 13162-13169.
- Chen, S., Jancrick, J., Yokota, H., Kim, R. & Kim, S.-H. (2004). *Proteins* *55*, 785-791.
- Cheong, C.G., Bauer, C.B., Brushaber, K.R., Escalante-Semerena, J.C. & Rayment, I. (2002). *Biochemistry* *41*, 4798-4808.
- Choi, I.-G., Shin, D.H., Brandsen, J., Jancarik, J., Busso, D., Yokota, H., Kim, R. & Kim, S.-H. (2003). *J. Struct. Funct. Genomics* *4*, 31-34.
- Colovos, C., Toth, E. A. & Yeates, T. O. (2000). *Acta Cryst. D56*, 1421-1429.
- Cowtan, K. & Main, P. (1996). *Acta Cryst. D52*, 43-48.
- Cowtan, K. & Main, P. (1998). *Acta Cryst. D54*, 487-493.
- Daniels, D.L., Cohen, A.R., Anderson, J.M. & Brunger, A.T. (1998). *Nat. Struct. Biol.* *5*, 317-325.
- DiDonato, M., Krishna, S.S., Schwarzenbacher, R., McMullan, D., Agarwalla, S., Brittain, S.M., Miller, M.D., Abdubek, P., Ambing, E., Axelrod, H.L., Canaves, J.M., Chiu, H.J., Deacon, A.M., Duan, L., Elsliger, M.A., Godzik, A., Grzechnik, S.K., Hale, J., Hampton, E., Haugen, J., Jaroszewski, L., Jin, K.K., Klock, H.E., Knuth, M.W., Koesema, E., Kreuzsch, A., Kuhn, P., Lesley, S.A., Levin, I., Morse, A.T., Nigoghossian, E., Okach, L., Oommachen, S., Paulsen, J., Quijano, K., Reyes, R., Rife, C.L., Spraggon, G., Stevens, R.C., van den Bedem, H., White, A., Wolf, G., Xu, Q., Hodgson, K.O., Wooley, J. & Wilson, I.A. (2006). *Proteins* *65*, 771-776.
- Ducros, V.M., Lewis, R.J., Verma, C.S., Dodson, E.J., Leonard, G., Turkenburg, J.P., Murshudov, G.N., Wilkinson, A.J. & Brannigan, J.A. (2001). *J. Mol. Biol.* *306*, 759-771.

- Eicken, C., Sharma, V., Klabunde, T., Lawrenz, M.B., Hardham, J.M., Norris, S.J. & Sacchettini, J.C. (2002). *J. Biol. Chem.* 277, 21691-21696.
- Esser, L., Wang, C.R., Hosaka, M., Smagula, C.S., Sudhof, T.C. & Deisenhofer, J. (1998). *EMBO J.* 17, 977-984.
- Gordon, E., Flouret, B., Chantalat, L., van Heijenoort, J., Mengin-Lecreulx, D. & Dideberg, O. (2001). *J. Biol. Chem.* 276, 10999-11006.
- Grosse-Kunstleve, R.W. & Adams, P.D. (2003). *Acta Cryst.* D59, 1966-1973.
- Hamilton, W.C. (1964). *Statistics in Physical Science*. The Ronald Press Company: New York
- Han, G.W., Sri Krishna, S., Schwarzenbacher, R., McMullan, D., Ginalski, K., Elsliger, M.A., Brittain, S.M., Abdubek, P., Agarwalla, S., Ambing, E., Astakhova, T., Axelrod, H., Canaves, J.M., Chiu, H.J., DiDonato, M., Grzechnik, S.K., Hale, J., Hampton, E., Haugen, J., Jaroszewski, L., Jin, K.K., Klock, H.E., Knuth, M.W., Koesema, E., Kreuzsch, A., Kuhn, P., Miller, M.D., Morse, A.T., Moy, K., Nigoghossian, E., Oommachen, S., Ouyang, J., Paulsen, J., Quijano, K., Reyes, R., Rife, C., Spraggon, G., Stevens, R.C., van den Bedem, H., Velasquez, J., Wang, X., West, B., White, A., Wolf, G., Xu, Q., Hodgson, K.O., Wooley, J., Deacon, A.M., Godzik, A., Lesley, S.A. & Wilson, I.A. (2006). *Proteins* 64, 1083-1090.
- Huang, C.-C., Smith, C.V., Glickman, M.S., Jacobs Jr., W.R. & Sacchettini, J.C. (2002). *J. Biol. Chem.* 277, 11559-11569.
- Huber, A.H., Nelson, W.J. & Weis, W.I. (1997). *Cell* 90, 871-882.
- Hyman, J., Chen, H., Di Fiore, P.P., De Camilli, P. & Brunger, A.T. (2000). *J. Cell. Biol.* 149, 537-546.
- James, M.N. & Sielecki, A.R. (1983). *J. Mol. Biol.* 163, 299-361.
- Jiang, J., Zhang, Y., Krainer, A.R. & Xu, R.M. (1999). *Proc. Natl. Acad. Sci. USA* 96, 3572-3577.
- Jin, K.K., Krishna, S.S., Schwarzenbacher, R., McMullan, D., Abdubek, P., Agarwalla, S., Ambing, E., Axelrod, H., Canaves, J.M., Chiu, H.J., Deacon, A.M., DiDonato, M., Elsliger, M.A., Feuerhelm, J., Godzik, A., Grittini, C., Grzechnik, S.K., Hale, J., Hampton, E., Haugen, J., Hornsby, M., Jaroszewski, L., Klock, H.E., Knuth, M.W., Koesema, E., Kreuzsch, A., Kuhn, P., Lesley, S.A., Miller, M.D., Moy, K., Nigoghossian, E., Okach, L., Oommachen, S., Paulsen, J., Quijano, K., Reyes, R., Rife, C., Stevens, R.C., Spraggon, G., van den Bedem, H., Velasquez, J., White, A., Wolf, G., Han, G.W., Xu, Q., Hodgson, K.O., Wooley, J. & Wilson, I.A. (2006). *Proteins* 63, 1112-1118.
- Kazantsev, A.V., Krivenko, A.A., Harrington, D.J., Carter, R.J., Holbrook, S.R., Adams, P.D. & Pace, N.R. (2003). *Proc. Natl. Acad. Sci. USA* 100, 7497-7502.
- Leslie, A. G. W. (1987). *Acta Cryst.* A43, 134-136.
- Leslie, A. G. W. (1992) *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography*, No. 26.
- Lunin, V. Y. (1993). *Acta Cryst.* D49, 90-99.
- McCoy, A. J., Storoni, L. C. & Read, R. J. (2004). *Acta Cryst.* D60, 1220-1228.

- Muchmore, C.R., Krahn, J.M., Kim, J.H., Zalkin, H. & Smith, J.L. (1998). *Protein Sci.* 7, 39-51.
- Nanao, M.H., Sheldrick, G.M. & Ravelli, R.B. (2005). *Acta Crystallogr. Sect. D* 61, 1227-1237.
- Newman, J., Peat, T.S., Richard, R., Kan, L., Swanson, P.E., Affholter, J.A., Holmes, I.H., Schindler, J.F., Unkefer, C.J. & Terwilliger, T.C. (1999). *Biochemistry* 38, 16105-16114.
- Ostermeier, C. & Brunger, A.T. (1999). *Cell* 96, 363-374.
- Otwinowski, Z. & Minor, W. (1997). *Methods in Enzymology*, Vol. 276: *Macromolecular Crystallography*, part A, 307-326. (C.W. Carter, Jr. & R. M. Sweet, Eds.). Academic Press (New York).
- Peat, T.S., Newman, J., Waldo, G.S., Berendzen, J. & Terwilliger, T.C. (1998). *Structure* 6, 1207-1214.
- Perrakis, A., Morris, R., Lamzin, V.S. (1999). *Nature Struct. Biol.* 6, 458-463.
- Pflugrath, J. W. (1999). *Acta Cryst. D*55, 1718-1725.
- Podjarny, A.D. (1976) X-ray studies of crystalline proteins. PhD. Thesis, Weizmann Institute of Science.
- Rice, L.M. & Brunger, A.T. (1999). *Mol. Cell* 4, 85-95.
- Rife, C., Schwarzenbacher, R., McMullan, D., Abdubek, P., Ambing, E., Axelrod, H., Biorac, T., Canaves, J.M., Chiu, H.J., Deacon, A.M., DiDonato, M., Elsliger, M.A., Godzik, A., Grittini, C., Grzechnik, S.K., Hale, J., Hampton, E., Han, G.W., Haugen, J., Hornsby, M., Jaroszewski, L., Klock, H.E., Koesema, E., Kreuzsch, A., Kuhn, P., Lesley, S.A., Miller, M.D., Moy, K., Nigoghossian, E., Paulsen, J., Quijano, K., Reyes, R., Sims, E., Spraggon, G., Stevens, R.C., van den Bedem, H., Velasquez, J., Vincent, J., White, A., Wolf, G., Xu, Q., Hodgson, K.O., Wooley, J. & Wilson, I.A. (2005). *Proteins* 61, 449-453.
- Rozwarski, D.A., Diederichs, K., Hecht, R., Boone, T. & Karplus, P.A. (1996). *Proteins* 26, 304-313.
- Schneider, T. T. & Sheldrick, G. M. (2002). *Acta Cryst. D*58, 1772-1779.
- Schwarzenbacher, R., McMullan, D., Krishna, S.S., Xu, Q., Miller, M.D., Canaves, J.M., Elsliger, M.A., Floyd, R., Grzechnik, S.K., Jaroszewski, L., Klock, H.E., Koesema, E., Kovarik, J.S., Kreuzsch, A., Kuhn, P., McPhillips, T.M., Morse, A.T., Quijano, K., Spraggon, G., Stevens, R.C., van den Bedem, H., Wolf, G., Hodgson, K.O., Wooley, J., Deacon, A.M., Godzik, A., Lesley, S.A. & Wilson, I.A. (2006). *Proteins* 65, 243-248.
- Sevcik, J., Dauter, Z., Lamzin, V.S. & Wilson, K.S. (1996). *Acta Crystallogr. Sect. D* 52, 327-344.
- Shamoo, Y., Krueger, U., Rice, L.M., Williams, K.R. & Steitz, T.A. (1997). *Nat. Struct. Biol.* 4, 215-222.
- Sharma, V., Grubmeyer, C. & Sacchettini, J.C. (1998). *Structure* 6, 1587-1599.
- Sharma, V., Sharma, S., Hoener zu Bentrup, K., McKinney, J.D., Russell, D.G., Jacobs Jr., W.R. & Sacchettini, J.C. (2000). *Nat. Struct. Biol.* 7, 663-668.
- Shin, D.H., Lou, Y., Jancarik, J., Yokota, H., Kim, R. & Kim, S.H. (2005). *J. Struct. Biol.* 152, 113-117.

- Shin, D.H., Nguyen, H.H., Jancarik, J., Yokota, H., Kim, R. & Kim, S.H. (2003b). *Biochemistry* 42, 13429-13437.
- Shin, D.H., Roberts, A., Jancarik, J., Yokota, H., Kim, R., Wemmer, D.E. & Kim, S.H. (2003). *Protein Sci.* 12, 1464-1472.
- Shin, D.H., Yokota, H., Kim, R. & Kim, S.H. (2002). *Proc. Natl. Acad. Sci. USA* 99, 7980-7985.
- Skinner, M.M., Zhang, H., Leschnitzer, D.H., Guan, Y., Bellamy, H., Sweet, R.M., Gray, C.W., Konings, R.N., Wang, A.H. & Terwilliger, T.C. (1994). *Proc. Natl. Acad. Sci. USA* 91, 2071-2075.
- Sutton, R.B., Ernst, J.A. & Brunger, A.T. (1999). *J. Cell. Biol.* 147, 589-598.
- Sutton, R.B., Fasshauer, D., Jahn, R. & Brunger, A.T. (1998). *Nature* 395, 347-353.
- Tanner, J.J., Lei, B., Tu, S.C. & Krause, K.L. (1996). *Biochemistry* 35, 13531-13539.
- Tavares, G.A., Panepucci, E.H. & Brunger, A.T. (2001). *Mol. Cell* 8, 1313-1325.
- Terwilliger, T. C. (1994). *Acta Crystallographica D*50, 11-16.
- Terwilliger, T. C. (1999). *Acta Crystallographica*, D55, 1863-1871.
- Terwilliger, T. C. (2000). *Acta Cryst.* D56, 965-972.
- Terwilliger, T. C. (2001). *Acta Crystallographica D*57, 1763-1775.
- Terwilliger, T. C. (2002a). *Acta Cryst.* D58, 2082-2086.
- Terwilliger, T. C. (2002b). *Acta Cryst.* D58, 2213-2215.
- Terwilliger, T. C. (2003). *Acta Cryst.* D59, 1688-1701.
- Terwilliger, T. C. and Berendzen, J. (1996). *Acta Crystallographica D*52, 749-757.
- Terwilliger, T. C. and Berendzen, J. (1997). *Acta Crystallographica*, D53, 571-579.
- Terwilliger, T. C. and Berendzen, J. (1999a). *Acta Crystallographica*, D55, 501-505
- Terwilliger, T. C. and Berendzen, J. (1999b). *Acta Crystallographica*, D55, 1872-1877
- Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, .PH., Hung, L.W., Read, R.J., Adams, P.D. (2007). *Acta Cryst D*, 64, 61-69.

- Turner, M.A., Yuan, C.S., Borchardt, R.T., Hershfield, M.S., Smith, G.D. & Howell, P.L. (1998). *Nat. Struct. Biol.* 5, 369-376.
- Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *J. Appl. Cryst.* (1995). 28, 347-351.
- Walsh, M.A., Otwinowski, Z., Perrakis, A., Anderson, P.M. & Joachimiak, A. (2000). *Structure* 8, 505-514.
- Wang, B.-C. (1985). *Methods Enzymol.* 115, 90-112.
- Weeks, C.M., Adams, P.D., Berendzen, J., Brunger, A.T., Dodson, E.J., Grosse-Kunstleve, R.W., Schneider, T.R., Sheldrick, G.M., Terwilliger, T.C., Turkenburg, M.G., Uson, I. (2003). *Methods Enzymol.* 374, 37-82.
- Willis, M.A., Bishop, B., Regan, L. & Brunger, A.T. (2000). *Structure Fold. Des.* 8, 1319-1328.
- Wilson, M.A. & Brunger, A.T. (2000). *J. Mol. Biol.* 301, 1237-1256.
- Xu, Q., Krishna, S.S., McMullan, D., Schwarzenbacher, R., Miller, M.D., Abdubek, P., Agarwalla, S., Ambing, E., Astakhova, T., Axelrod, H.L., Canaves, J.M., Carlton, D., Chiu, H.J., Clayton, T., DiDonato, M., Duan, L., Elsliger, M.A., Feuerhelm, J., Grzechnik, S.K., Hale, J., Hampton, E., Han, G.W., Haugen, J., Jaroszewski, L., Jin, K.K., Klock, H.E., Knuth, M.W., Koesema, E., Kreusch, A., Kuhn, P., Morse, A.T., Nigoghossian, E., Okach, L., Oommachen, S., Paulsen, J., Quijano, K., Reyes, R., Rife, C.L., Spraggon, G., Stevens, R.C., van den Bedem, H., White, A., Wolf, G., Hodgson, K.O., Wooley, J., Deacon, A.M., Godzik, A., Lesley, S.A. & Wilson, I.A. (2006). *Proteins* 65, 777-782.
- Xu, Q., Schwarzenbacher, R., Krishna, S.S., McMullan, D., Agarwalla, S., Quijano, K., Abdubek, P., Ambing, E., Axelrod, H., Biorac, T., Canaves, J.M., Chiu, H.J., Elsliger, M.A., Grittini, C., Grzechnik, S.K., DiDonato, M., Hale, J., Hampton, E., Han, G.W., Haugen, J., Hornsby, M., Jaroszewski, L., Klock, H.E., Knuth, M.W., Koesema, E., Kreusch, A., Kuhn, P., Miller, M.D., Moy, K., Nigoghossian, E., Paulsen, J., Reyes, R., Rife, C., Spraggon, G., Stevens, R.C., van den Bedem, H., Velasquez, J., White, A., Wolf, G., Hodgson, K.O., Wooley, J., Deacon, A.M., Godzik, A., Lesley, S.A. & Wilson, I.A. (2006). *Proteins* 64, 808-813.
- Xu, Q., Schwarzenbacher, R., McMullan, D., Abdubek, P., Agarwalla, S., Ambing, E., Axelrod, H., Biorac, T., Canaves, J.M., Chiu, H.J., Deacon, A.M., DiDonato, M., Elsliger, M.A., Godzik, A., Grittini, C., Grzechnik, S.K., Hale, J., Hampton, E., Han, G.W., Haugen, J., Hornsby, M., Jaroszewski, L., Klock, H.E., Koesema, E., Kreusch, A., Kuhn, P., Lesley, S.A., Miller, M.D., Moy, K., Nigoghossian, E., Paulsen, J., Quijano, K., Reyes, R., Rife, C., Spraggon, G., Stevens, R.C., van den Bedem, H., Velasquez, J., White, A., Wolf, G., Hodgson, K.O., Wooley, J. & Wilson, I.A. (2006). *Proteins* 62, 292-296.
- Xu, Q., Schwarzenbacher, R., McMullan, D., von Delft, F., Brinen, L.S., Canaves, J.M., Dai, X., Deacon, A.M., Elsliger, M.A., Eshagi, S., Floyd, R., Godzik, A., Grittini, C., Grzechnik, S.K., Jaroszewski, L., Karlak, C., Klock, H.E., Koesema, E., Kovarik, J.S., Kreusch, A., Kuhn, P., Lesley, S.A., Levin, I., McPhillips, T.M., Miller, M.D., Morse, A., Moy, K., Ouyang, J., Page, R., Quijano, K., Robb, A., Spraggon, G., Stevens, F., van den Bedem, H., Velasquez, J., Vincent, J., Wang, X., West, B., Wolf, G., Hodgson, K.O., Wooley, J. & Wilson, I.A. (2004). *Proteins* 56, 171-175.
- Yang, D., Shipman, L.W., Roessner, C.A., Scott, A.I. & Sacchettini, J.C. (2002). *J. Biol. Chem.* 277, 9462-9467.

Yu, R.C., Hanson, P.I., Jahn, R. & Brunger, A.T. (1998). *Nat. Struct. Biol.* 5, 803-811.

Yu, R.C., Jahn, R. & Brunger, A.T. (1999). *Mol. Cell* 4, 97-107.

Zhang, L. & Doudna, J.A. (2002). *Science* 295, 2084-2088.

Zubieta, C., Krishna, S.S., McMullan, D., Miller, M.D., Abdubek, P., Agarwalla, S., Ambing, E., Astakhova, T., Axelrod, H.L., Carlton, D., Chiu, H.J., Clayton, T., Deller, M., Didonato, M., Duan, L., Elsliger, M.A., Grzechnik, S.K., Hale, J., Hampton, E., Han, G.W., Haugen, J., Jaroszewski, L., Jin, K.K., Klock, H.E., Knuth, M.W., Koesema, E., Kumar, A., Marciano, D., Morse, A.T., Nigoghossian, E., Oommachen, S., Reyes, R., Rife, C.L., Bedem, H.V., Weekes, D., White, A., Xu, Q., Hodgson, K.O., Wooley, J., Deacon, A.M., Godzik, A., Lesley, S.A. & Wilson, I.A. (2007). *Proteins* 68, 999-1005.

Zwart, P.H., Grosse-Kunstleve, R.W., & Adams, P.D. (2005). *CCP4 newsletter Winter, Contribution 7.*

Figure 1A

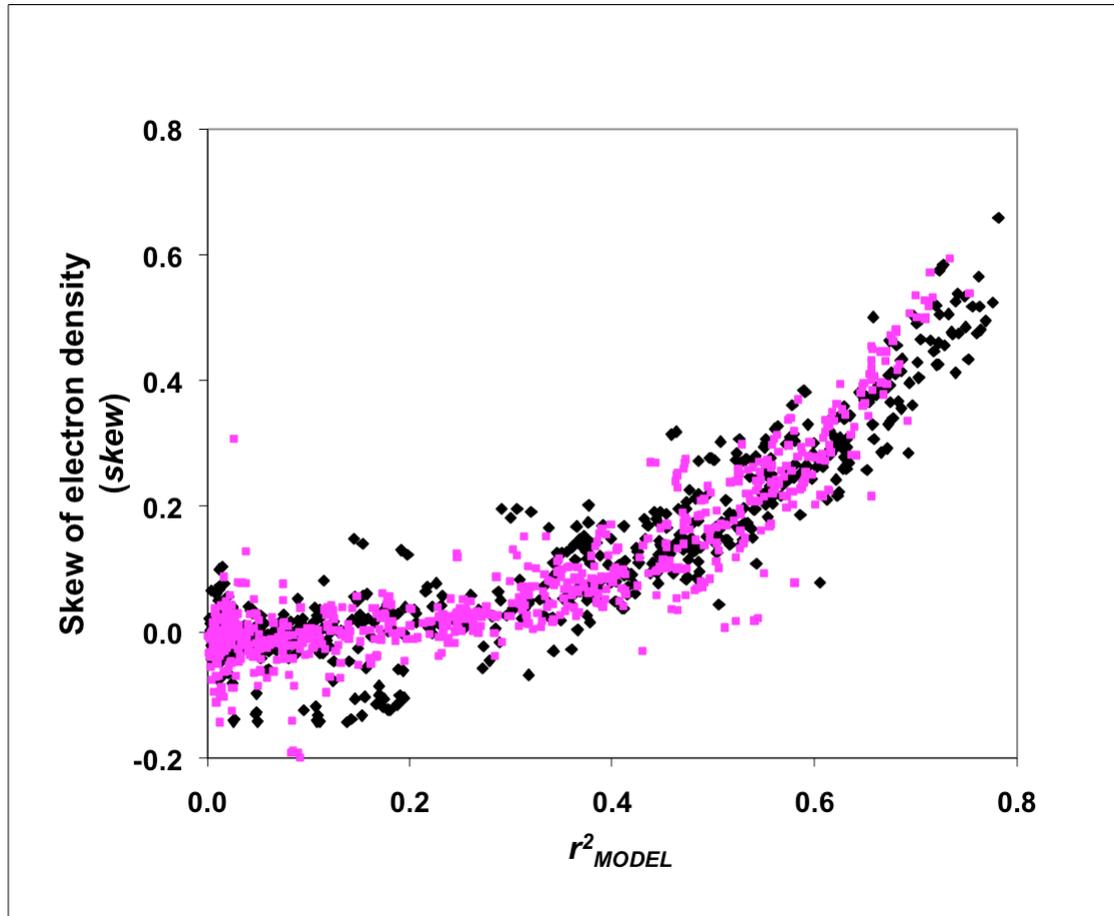


Figure 1B

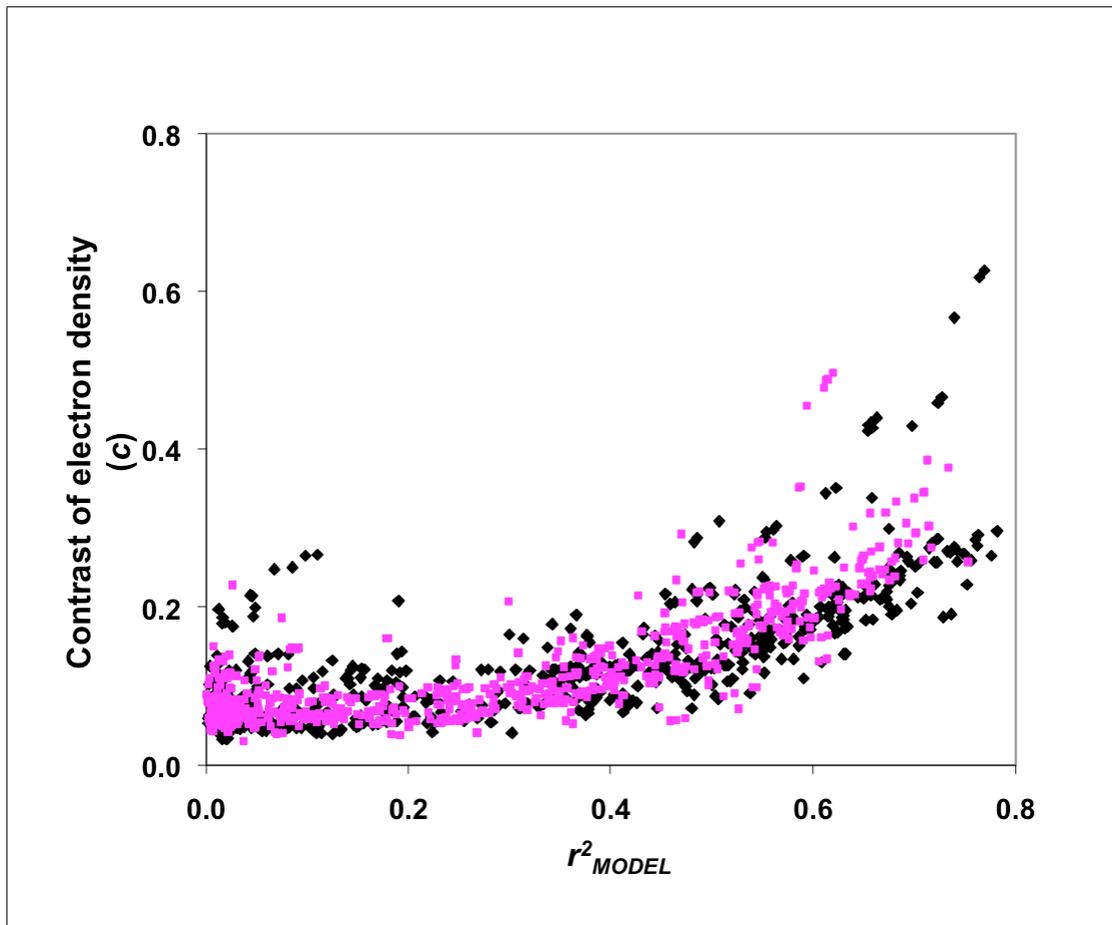


Figure 1C

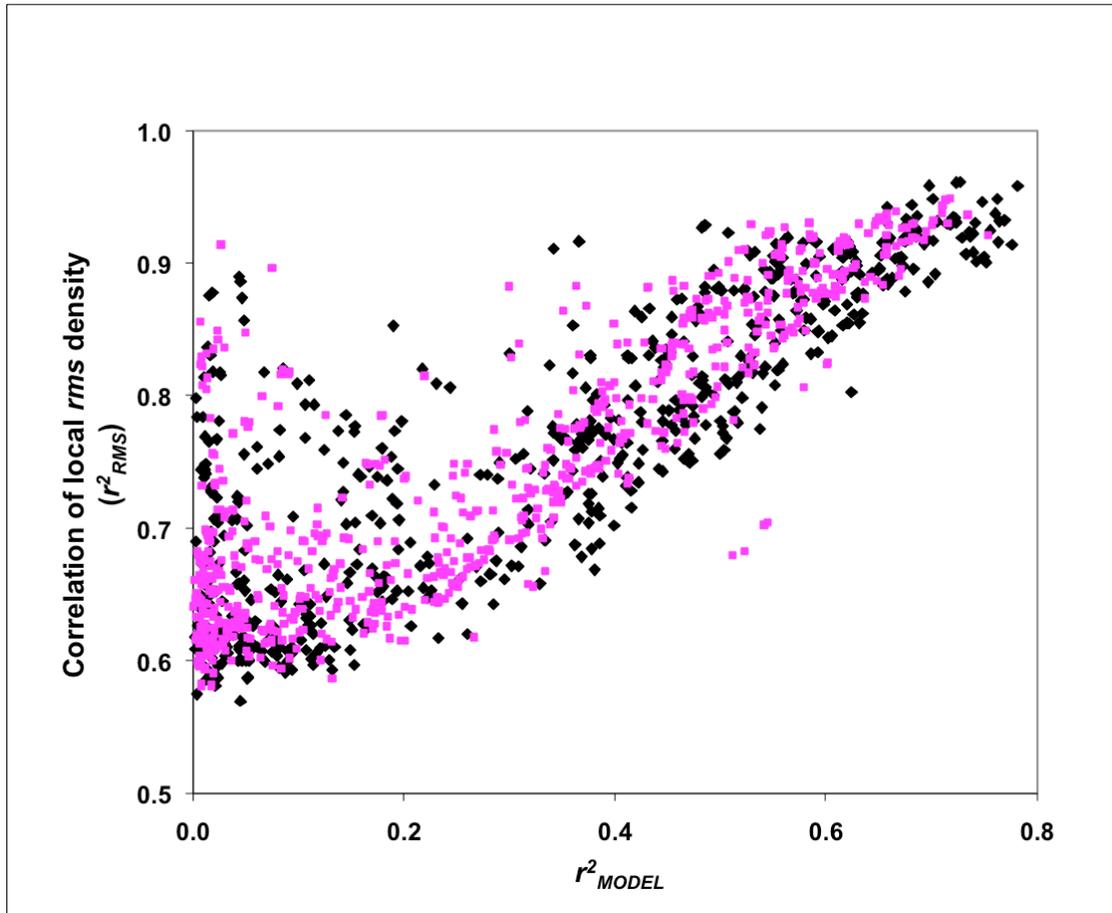


Figure 1D

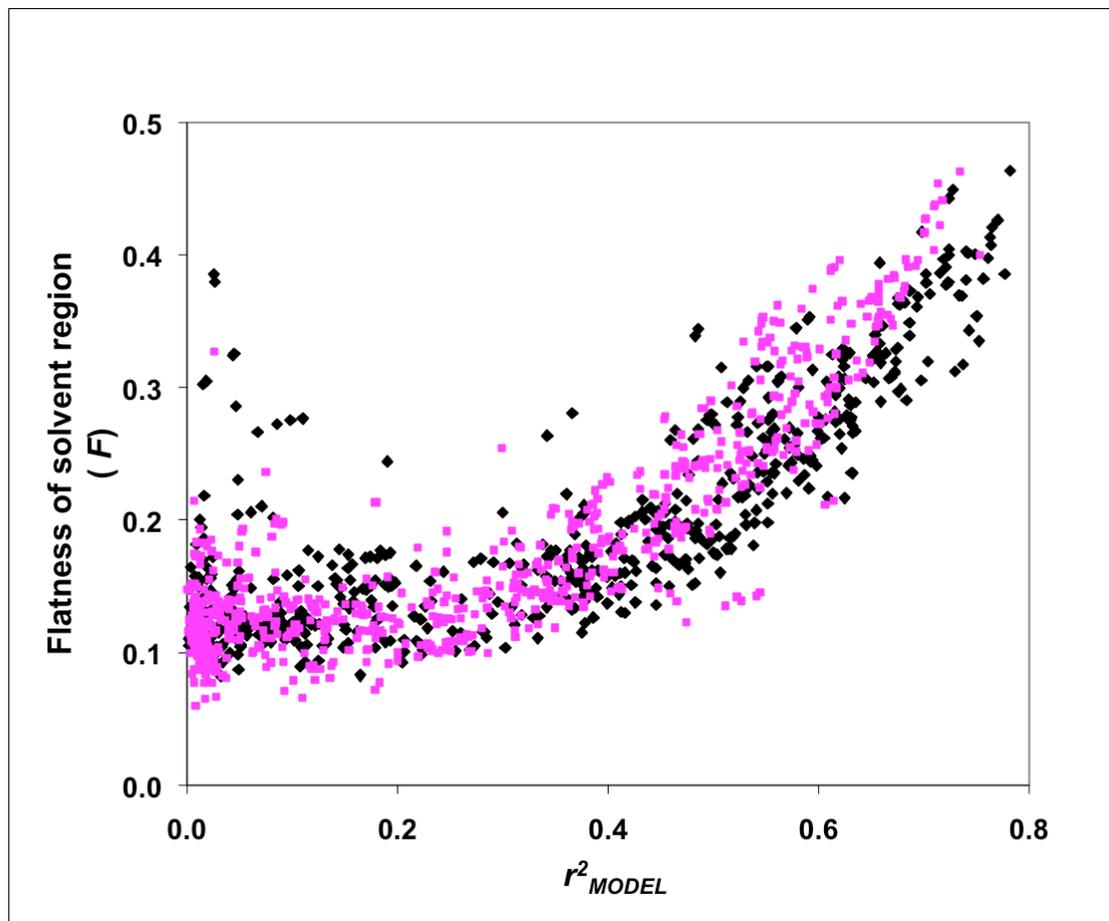


Figure 1E

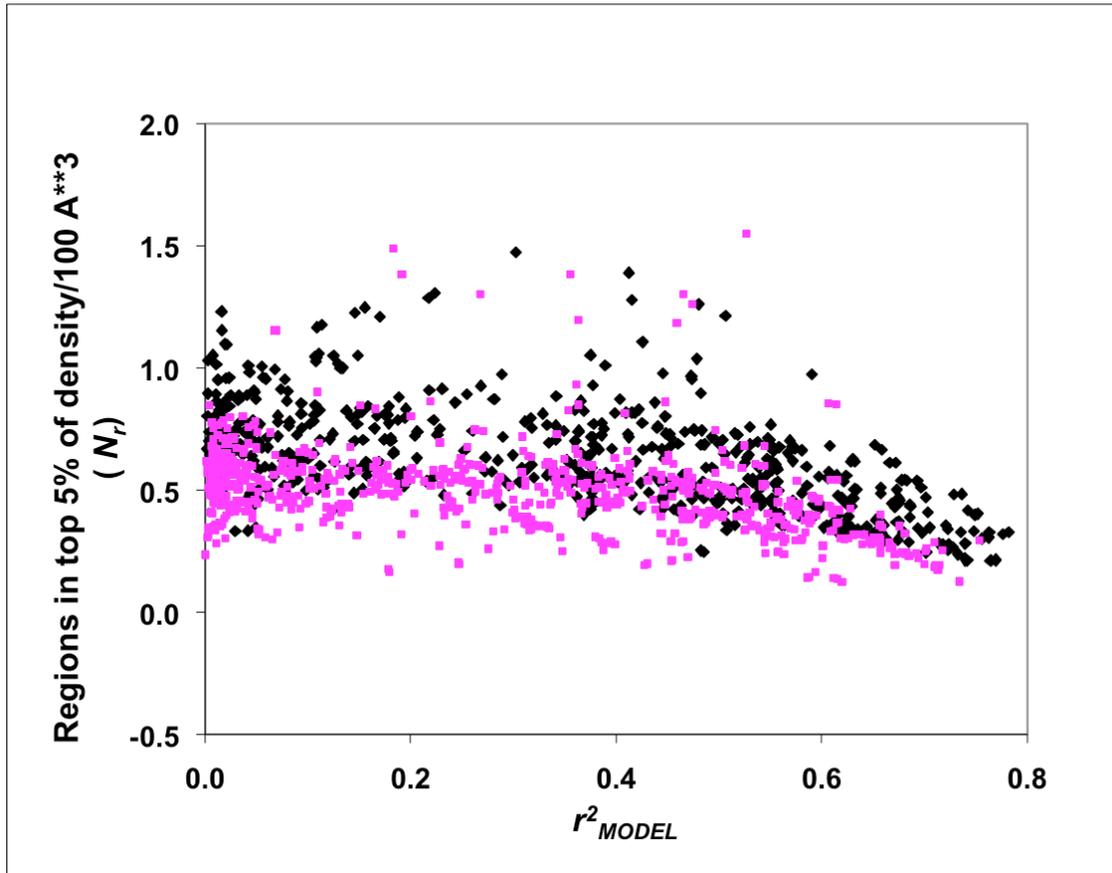


Figure 1F

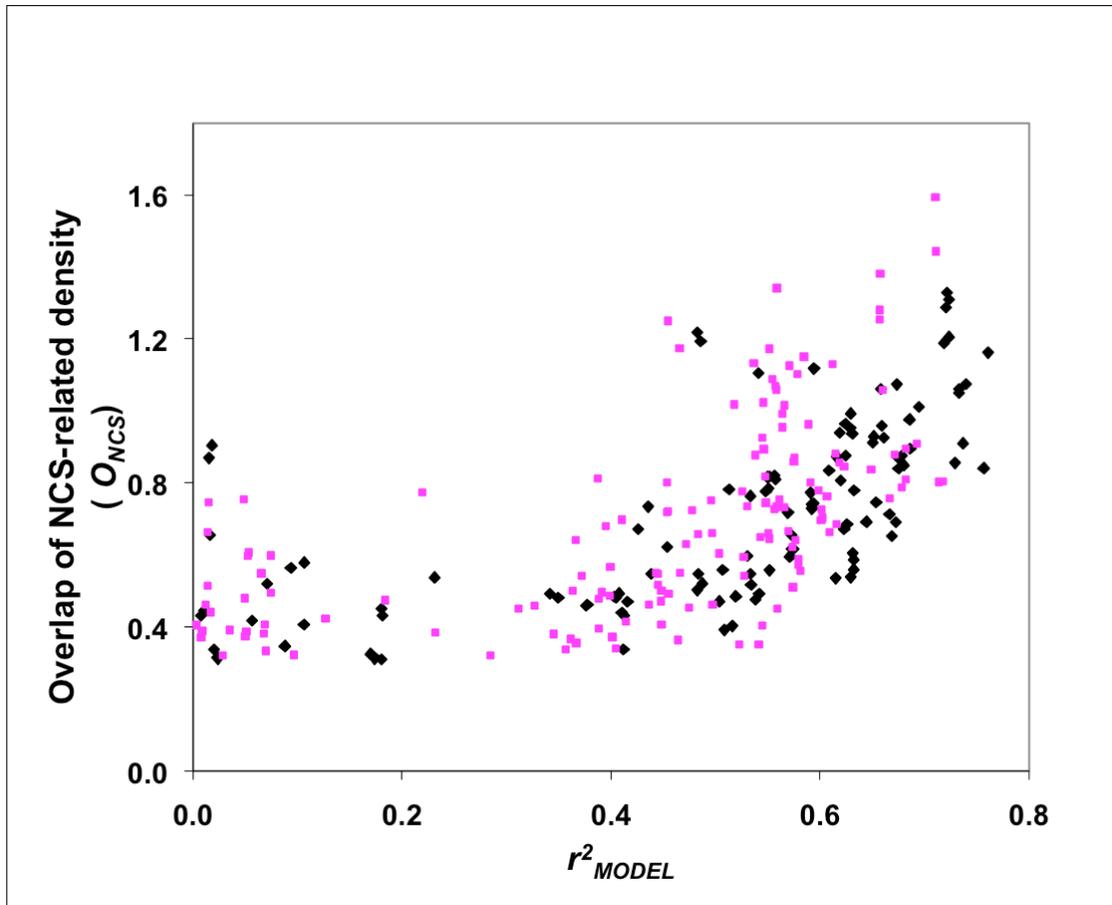


Figure 1G

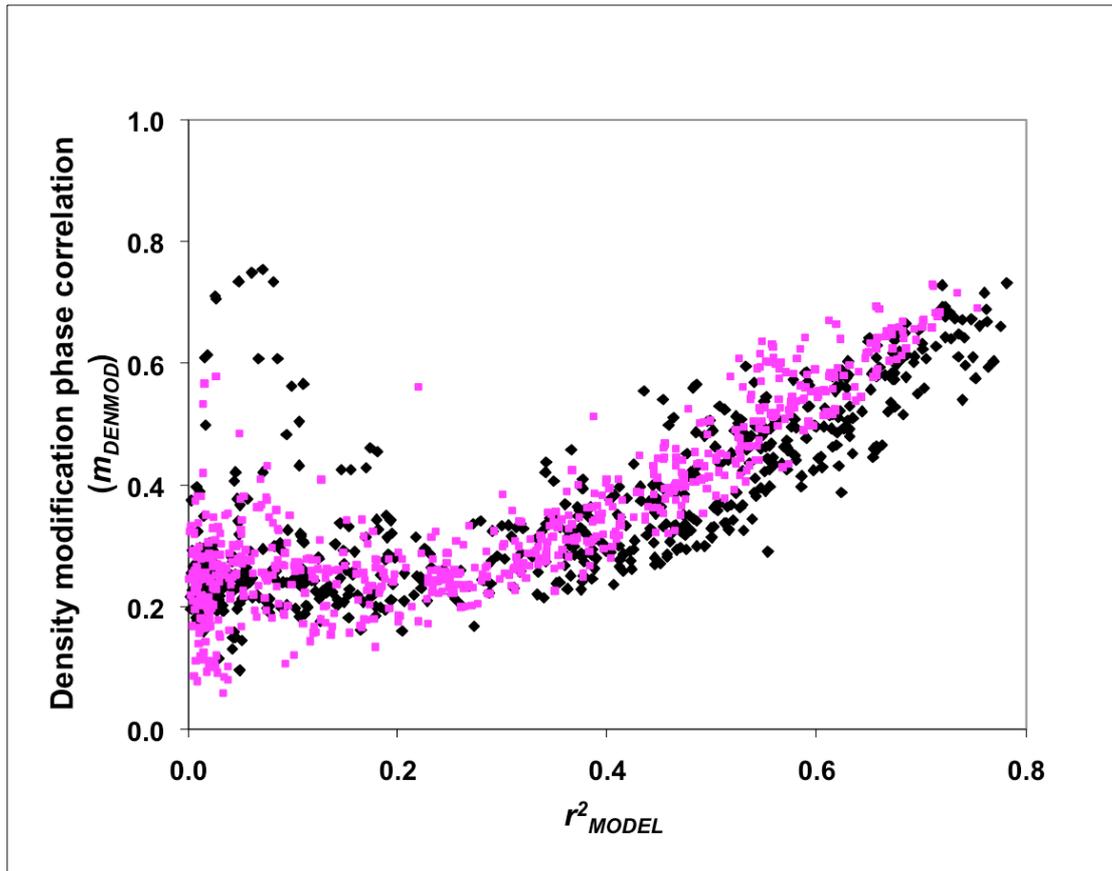


Figure 1H

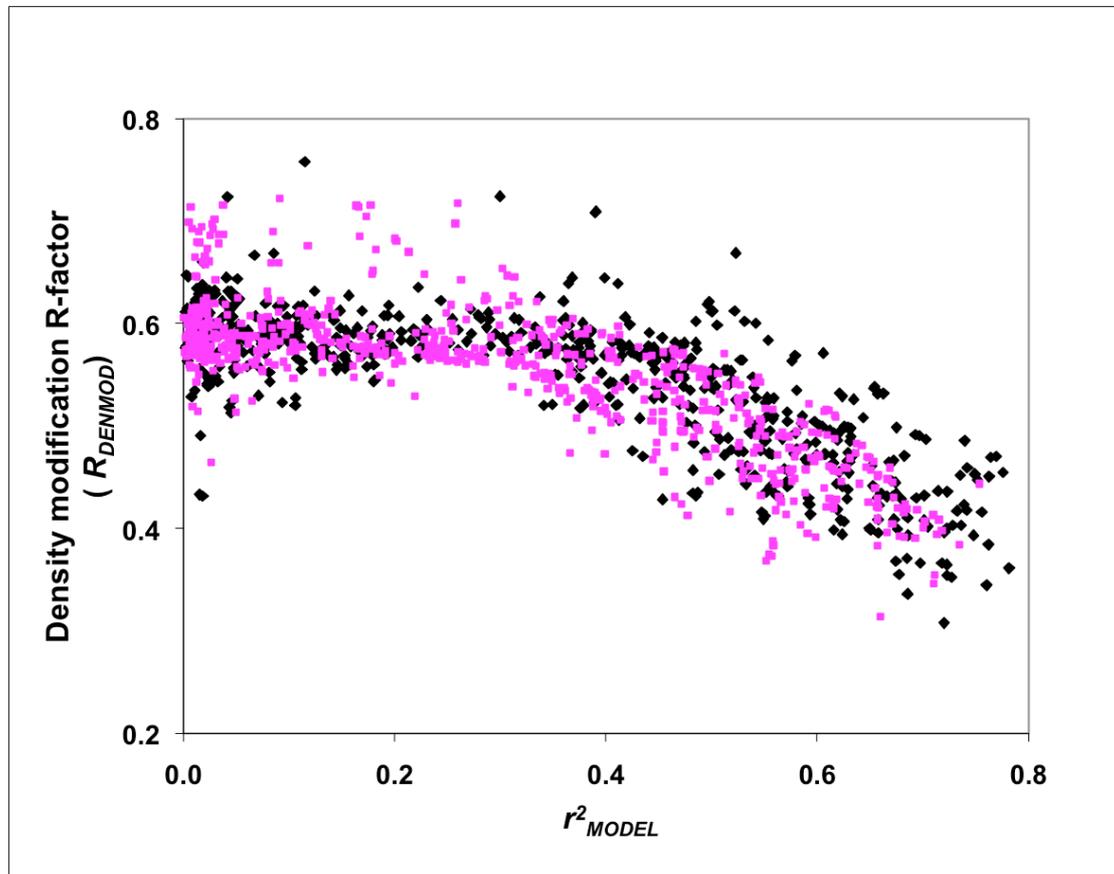


Figure 11

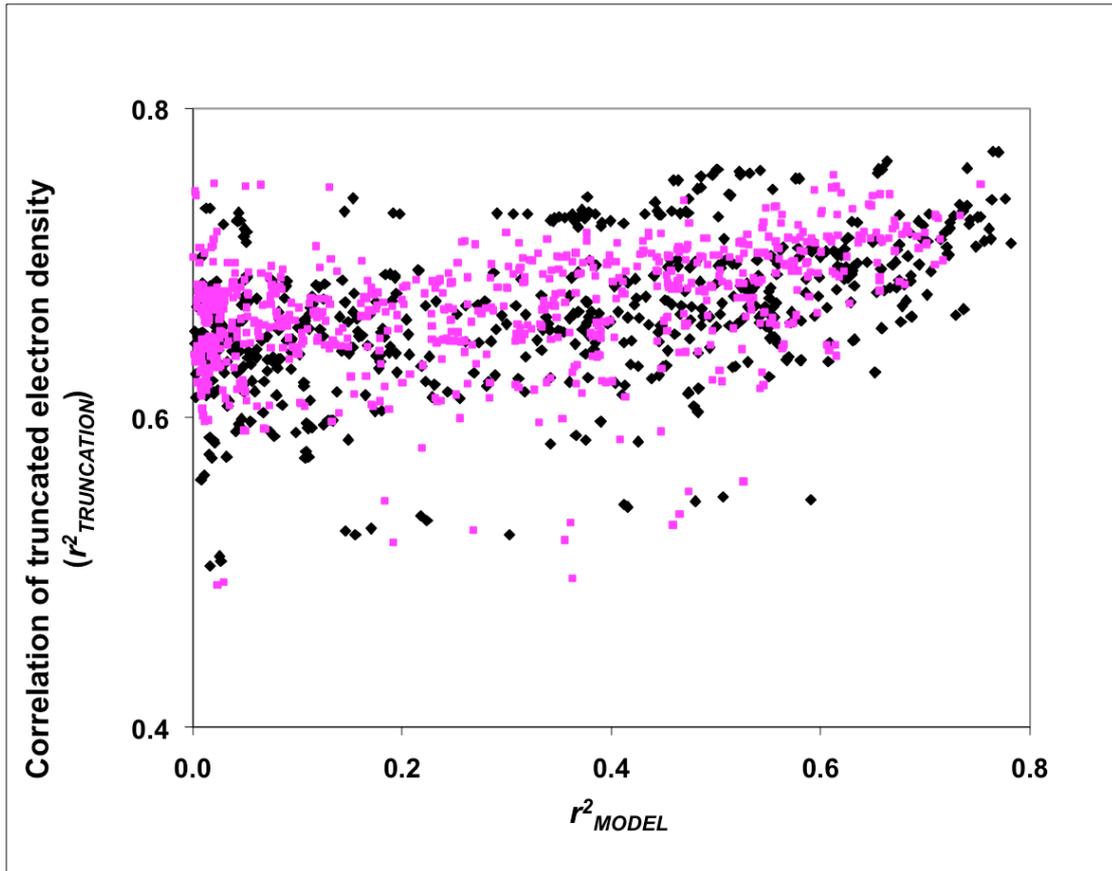


Figure 1J

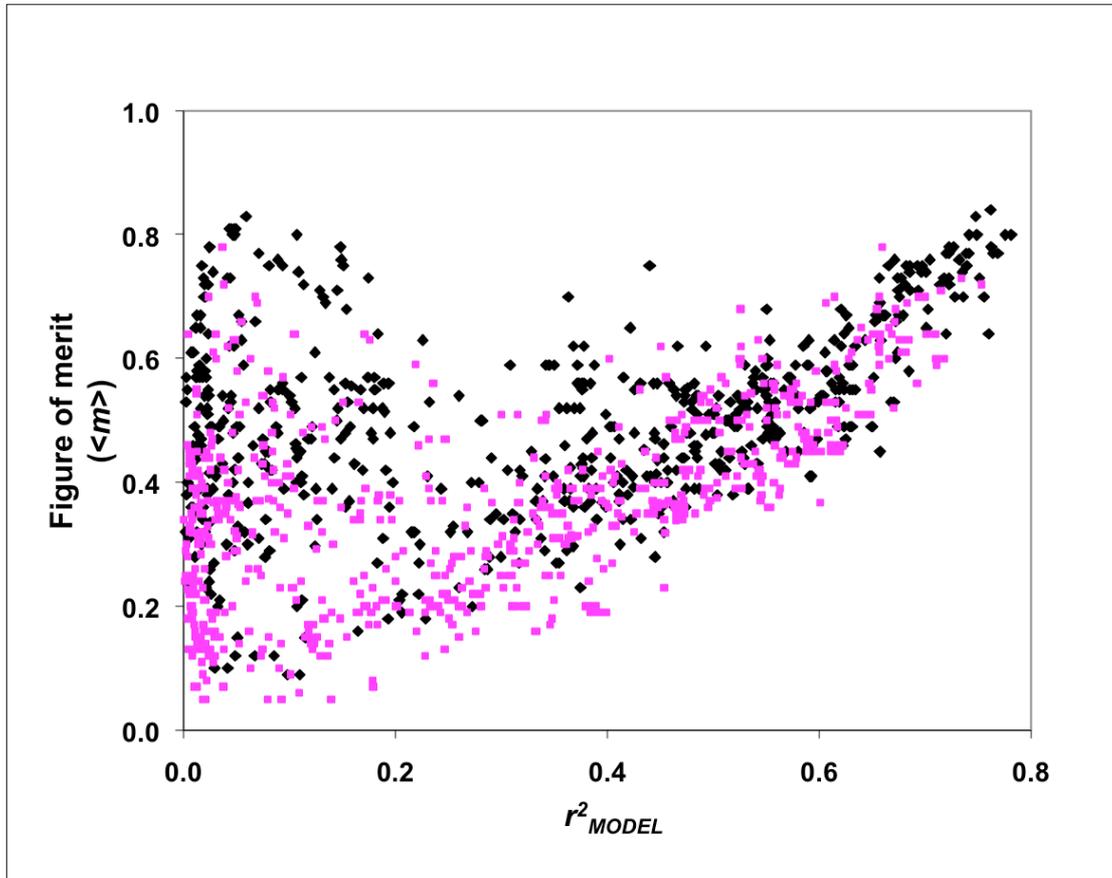


Figure 2A

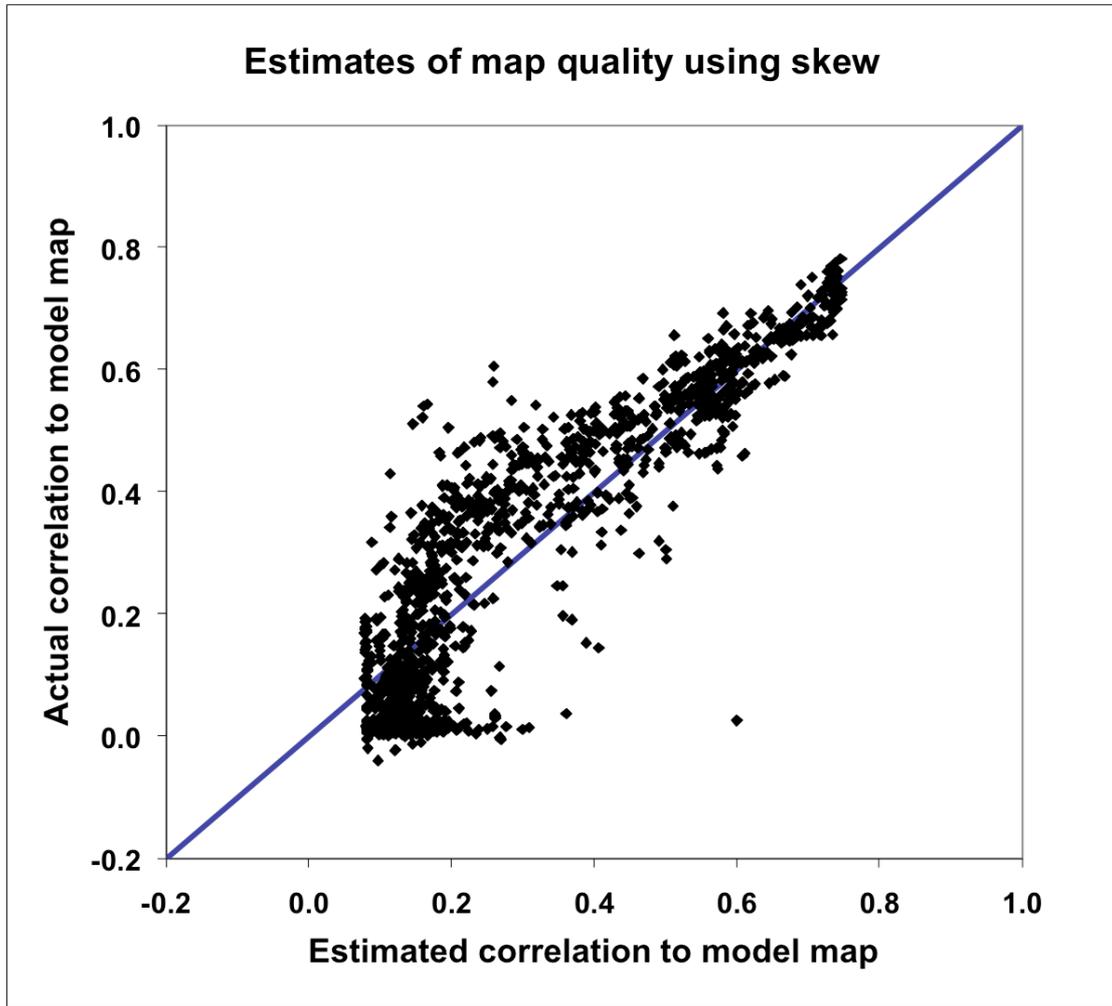


Figure 2B

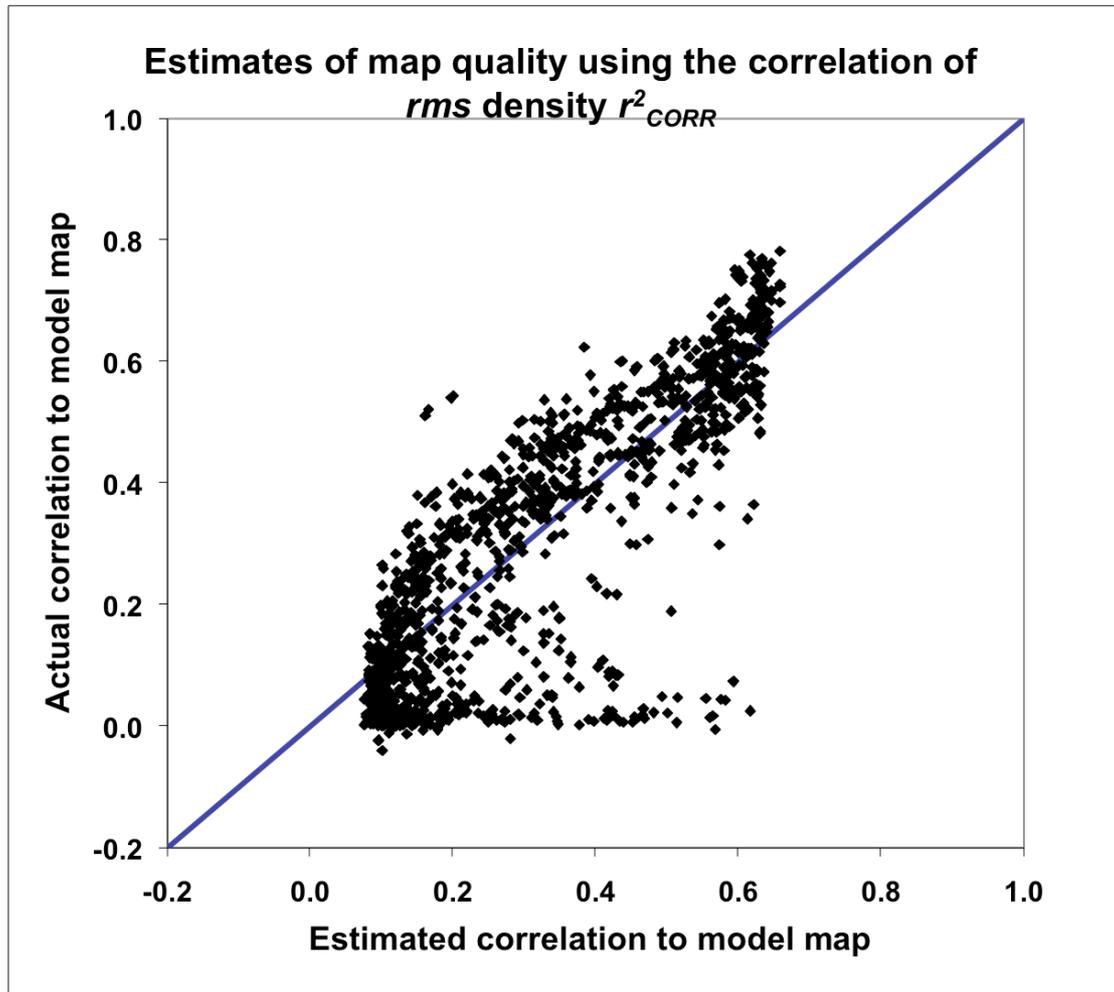


Figure 2C

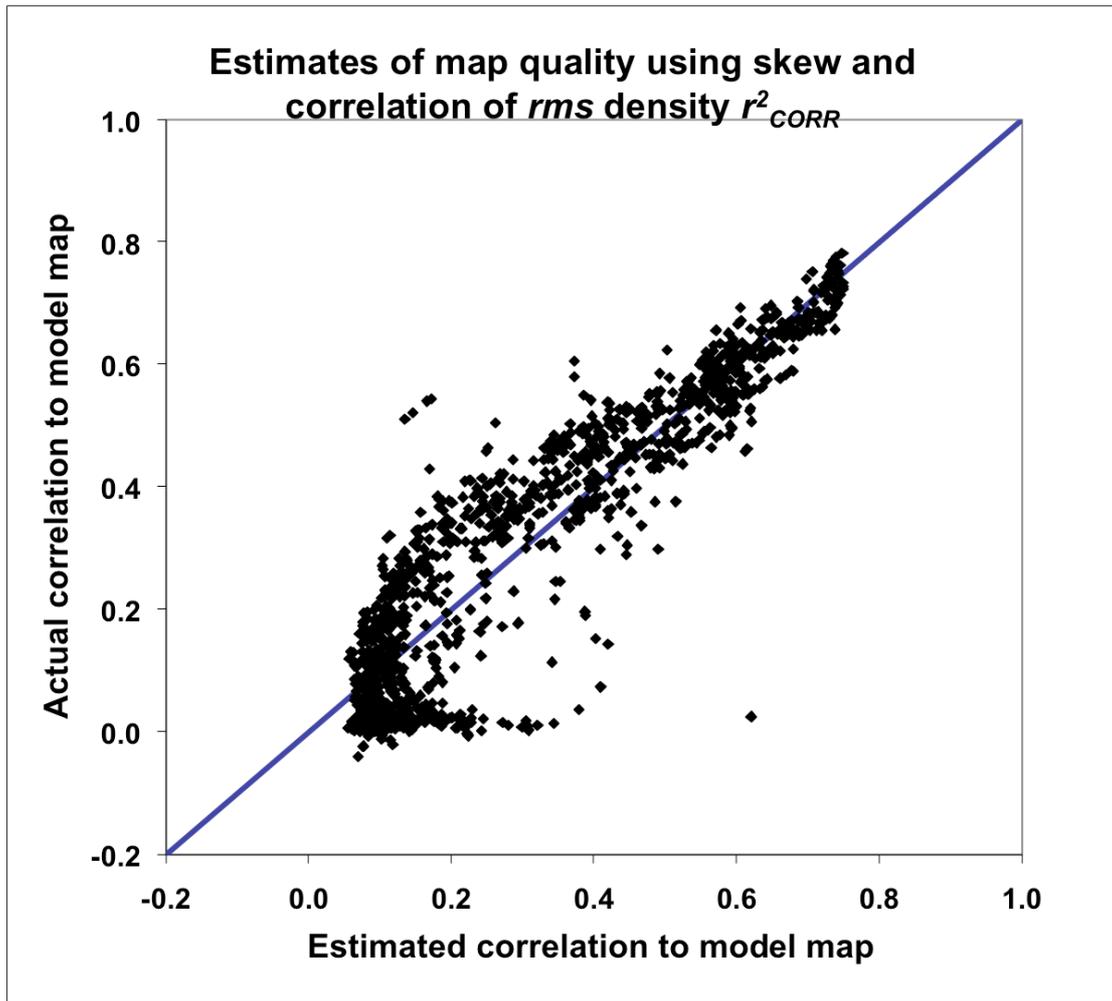


Fig. 3 A

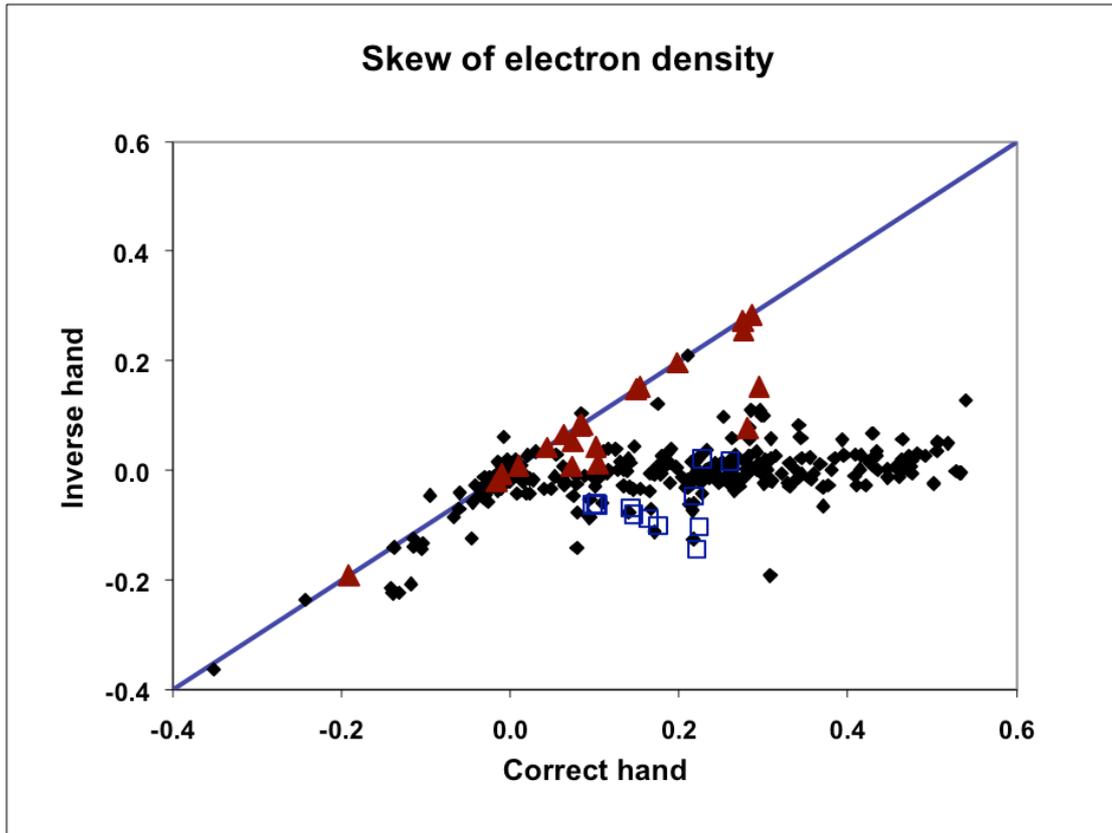


Fig. 3 B

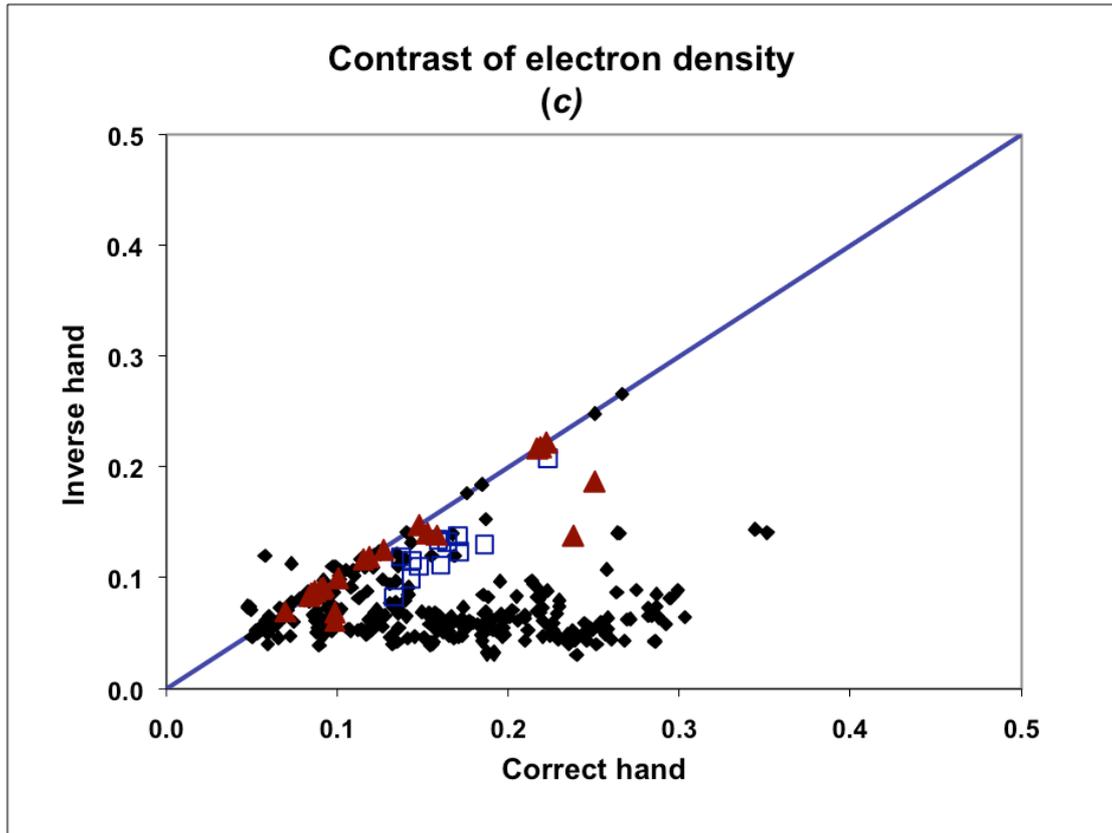


Fig. 3 C

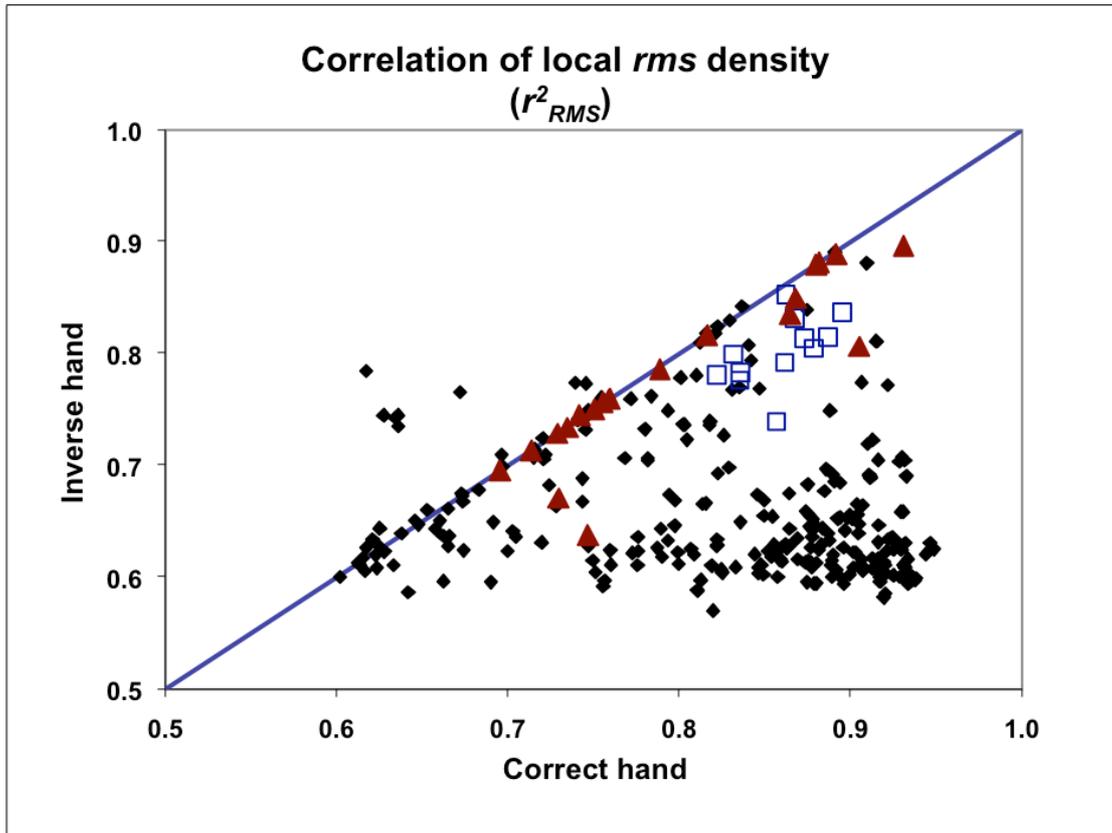


Fig. 3 D

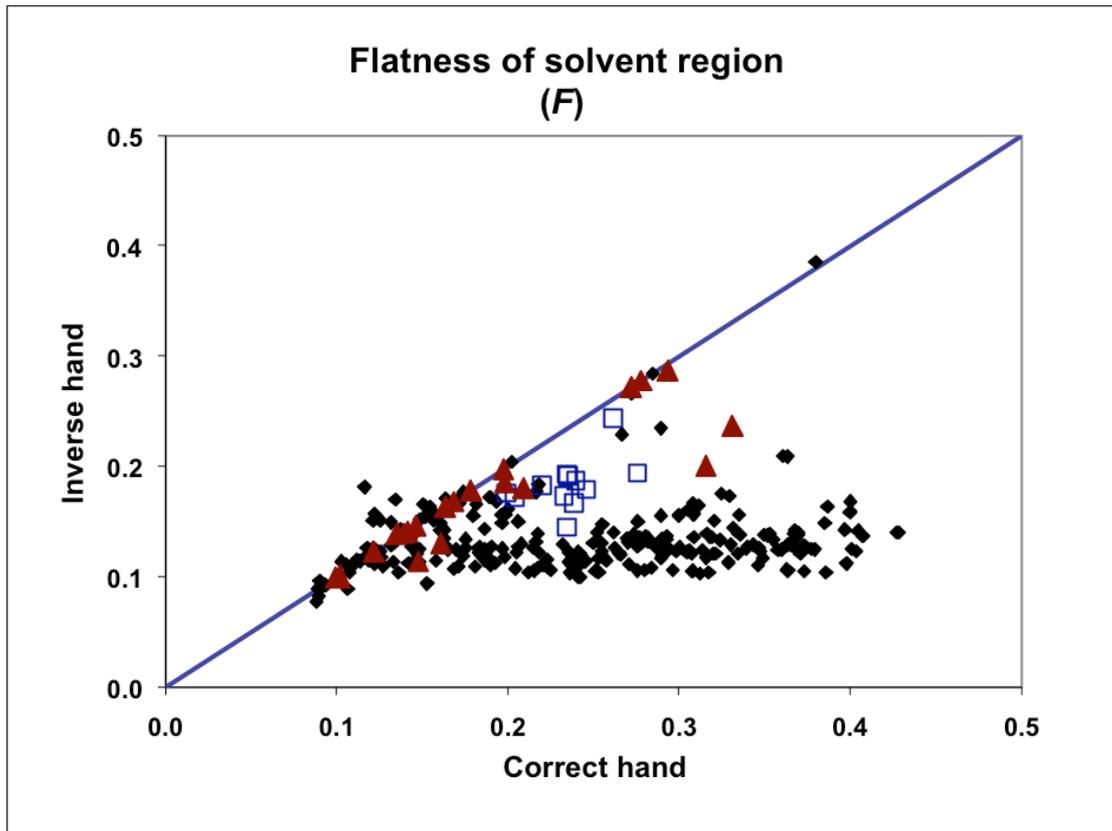


Fig. 3 E

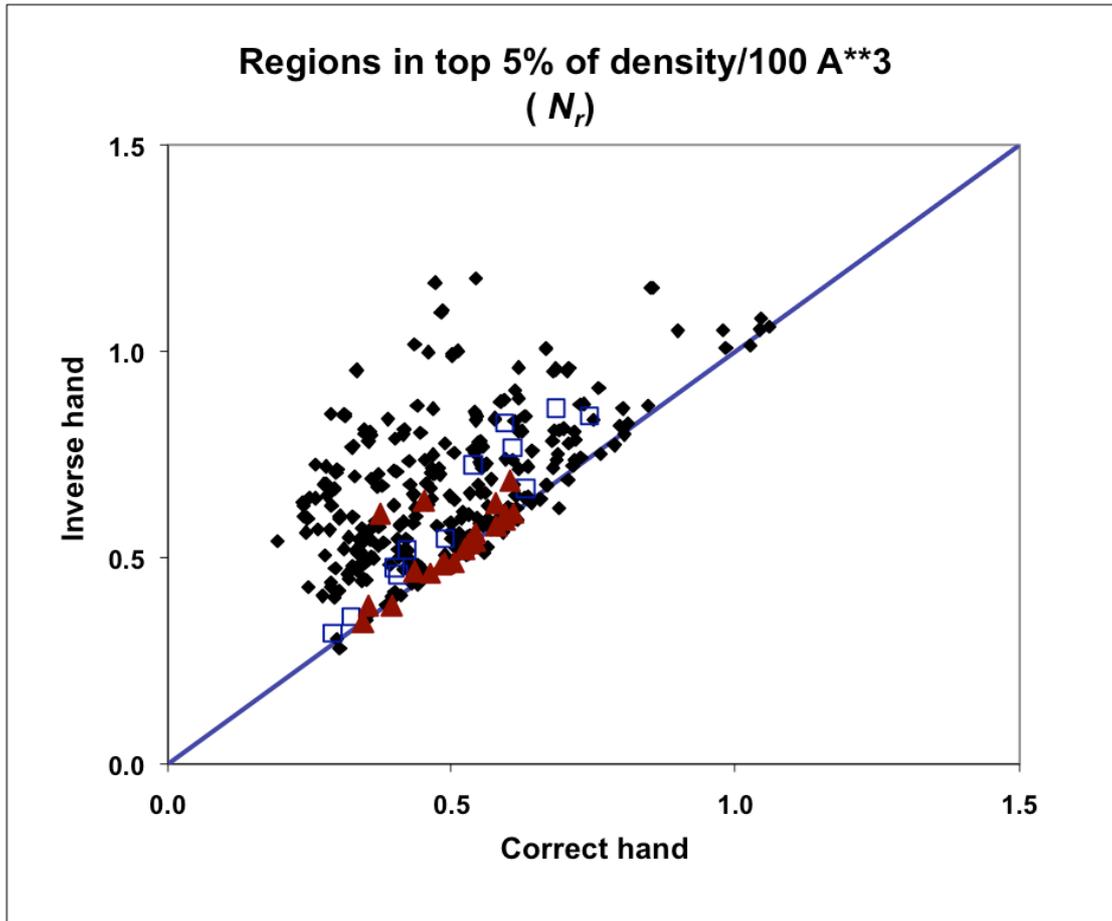


Fig. 3 F

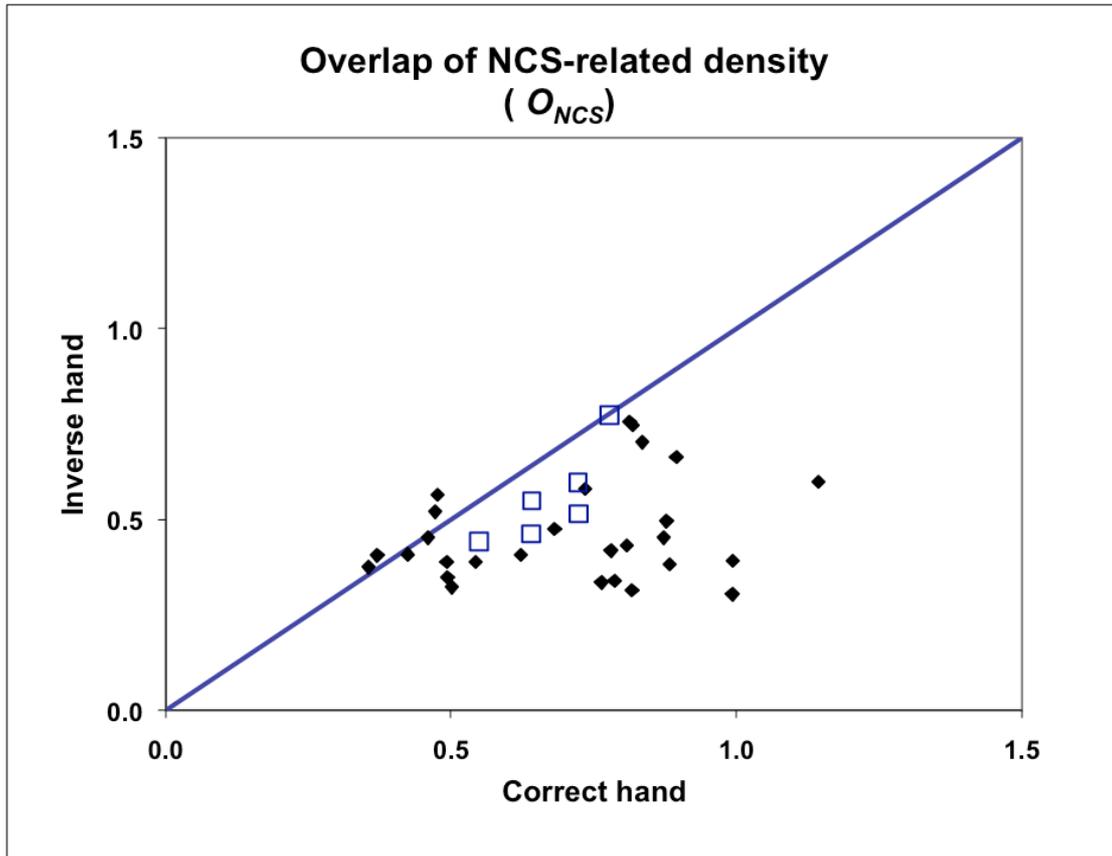


Fig. 3 G

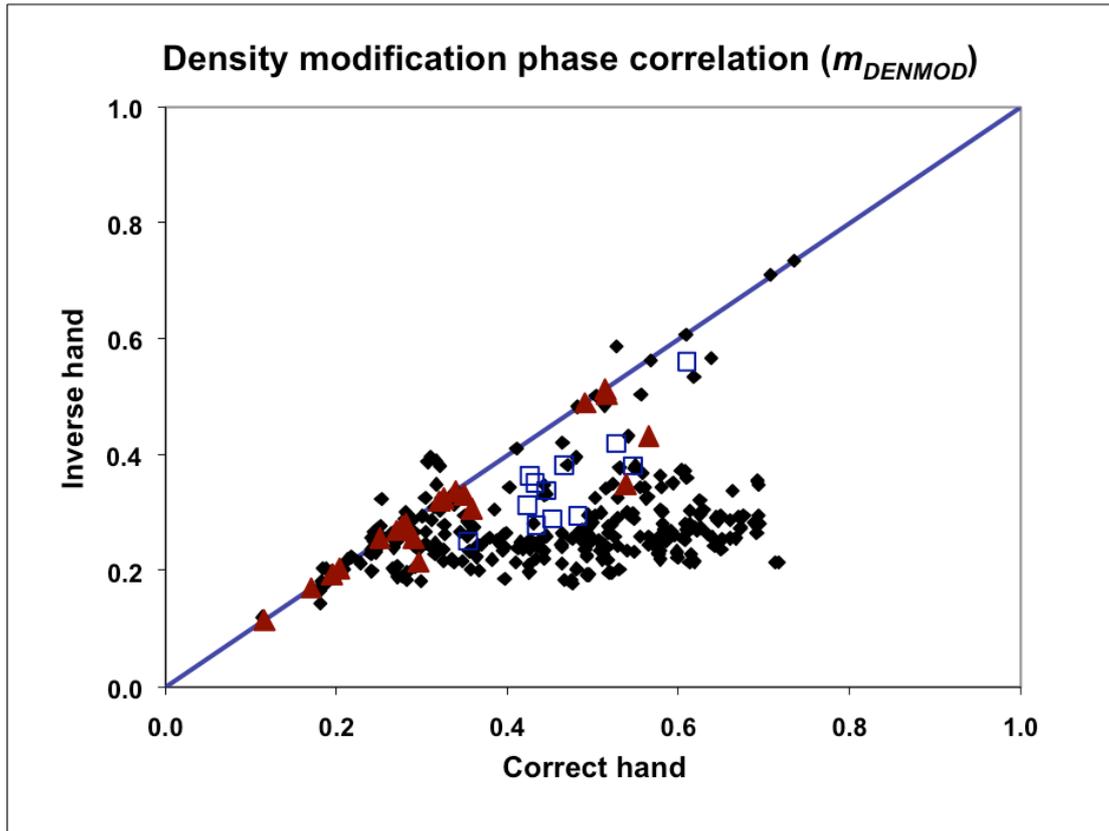


Fig. 3 H

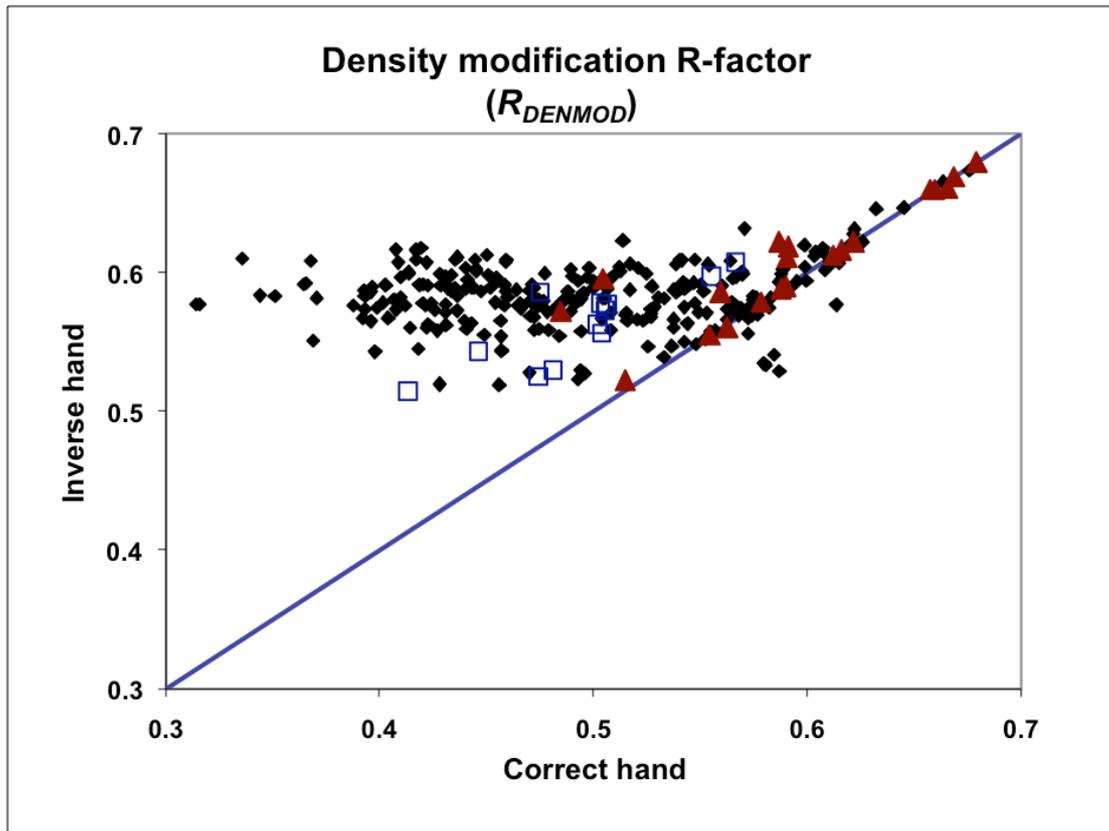


Fig. 3 I

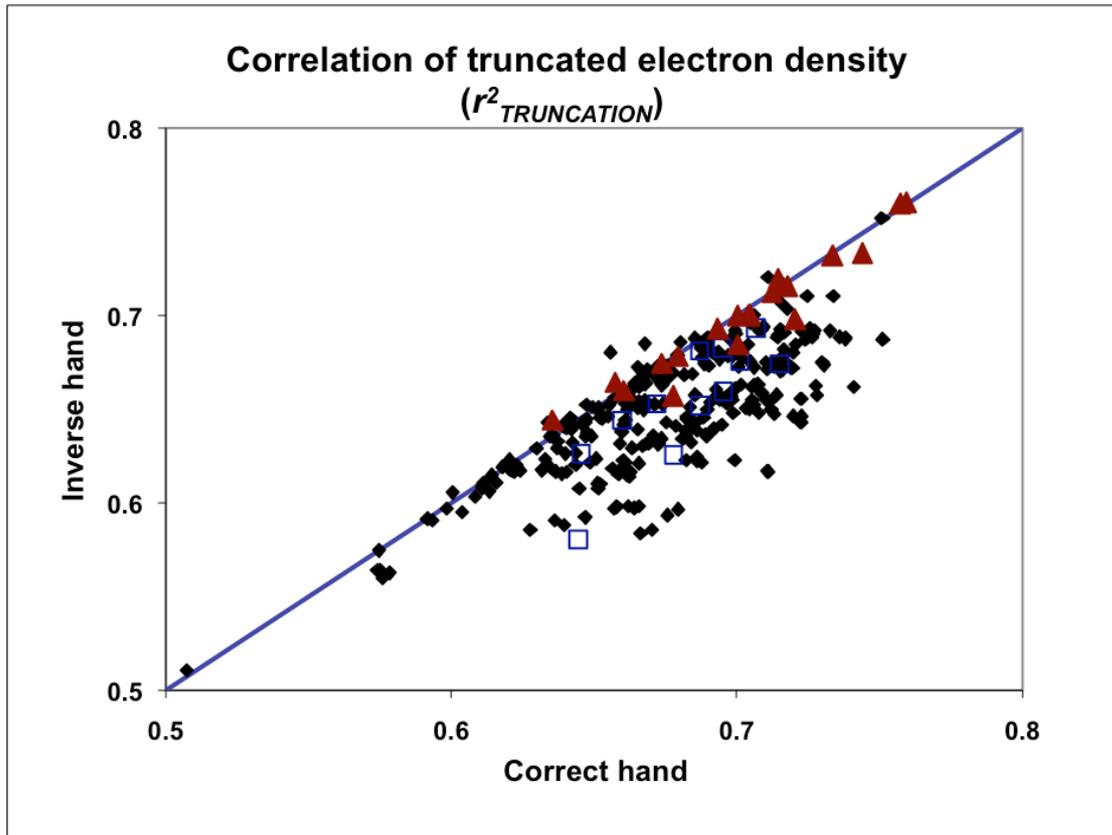


Fig. 4A

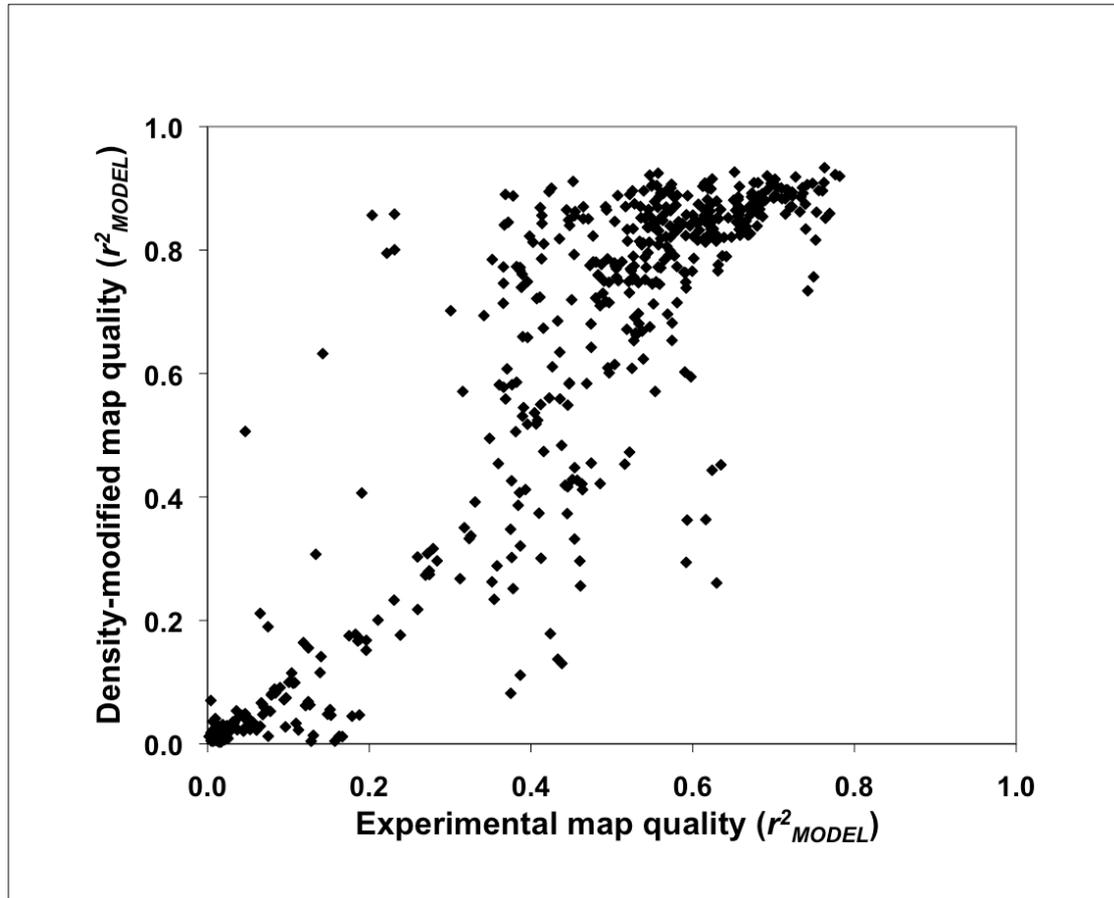


Fig. 4B

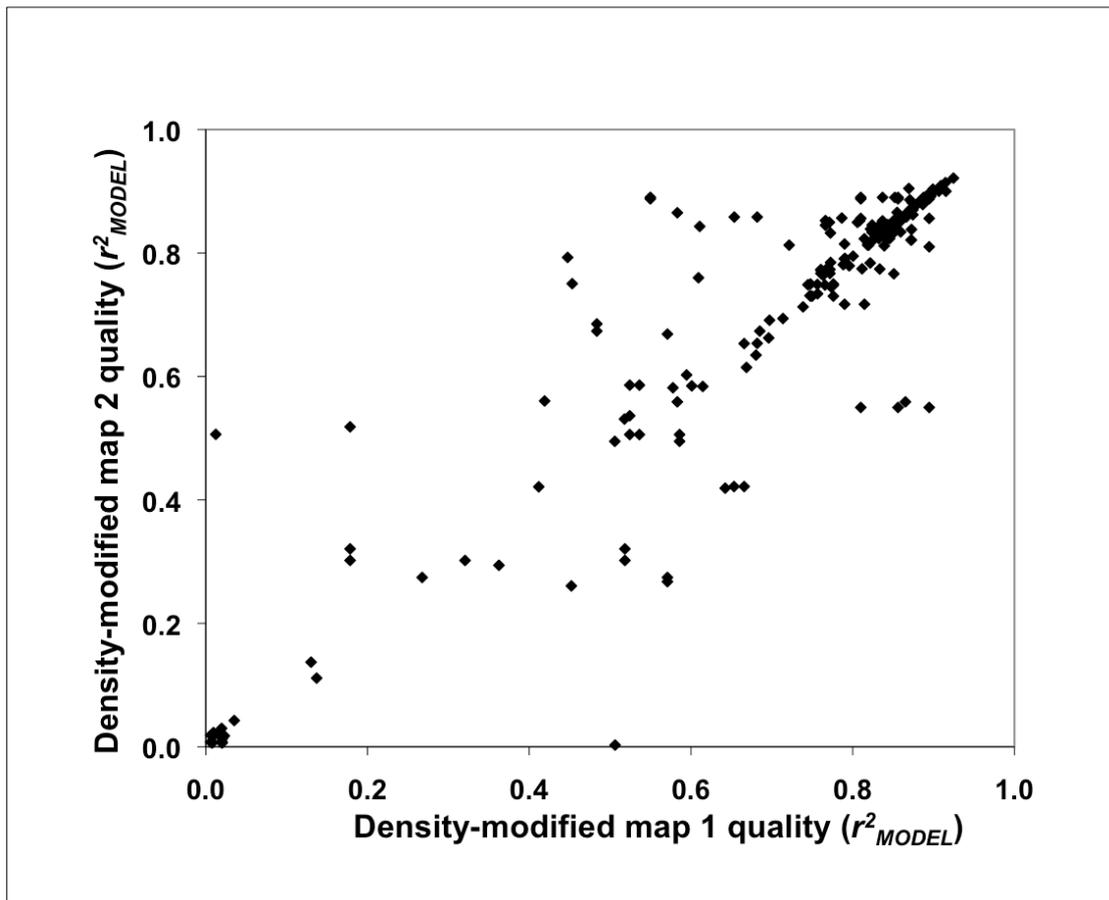


Fig. 5 A

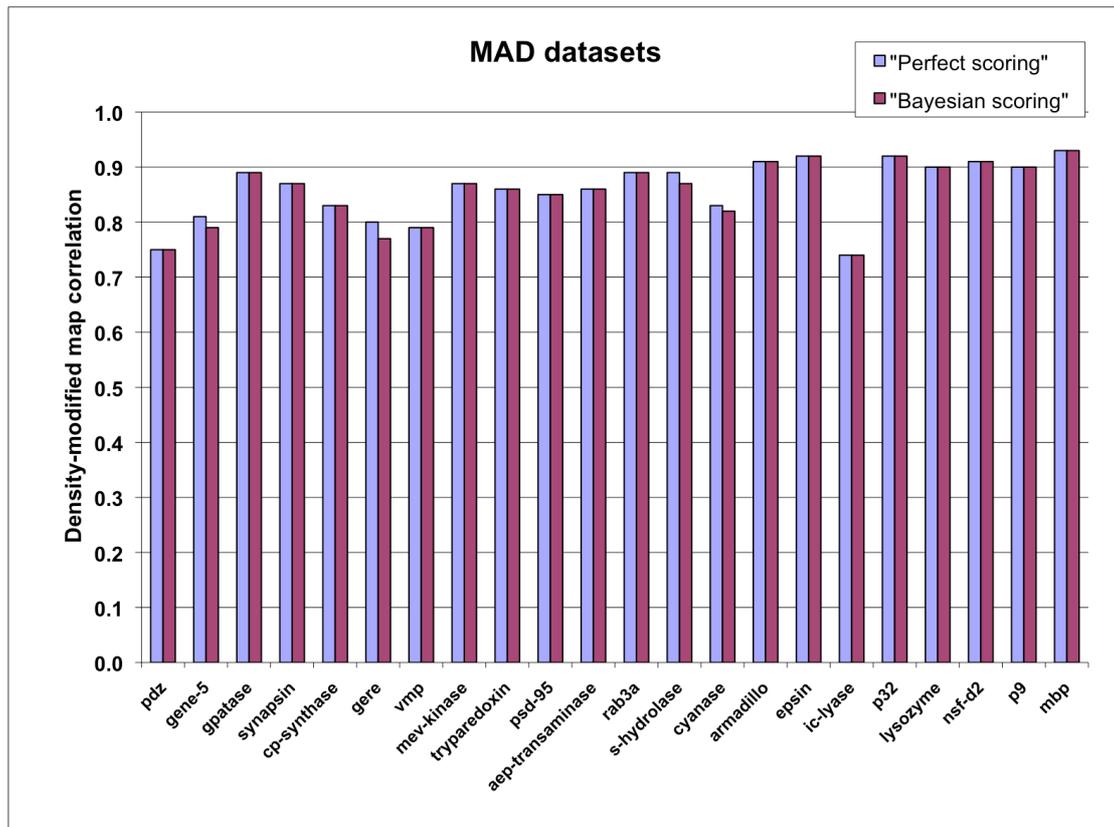


Fig. 5B

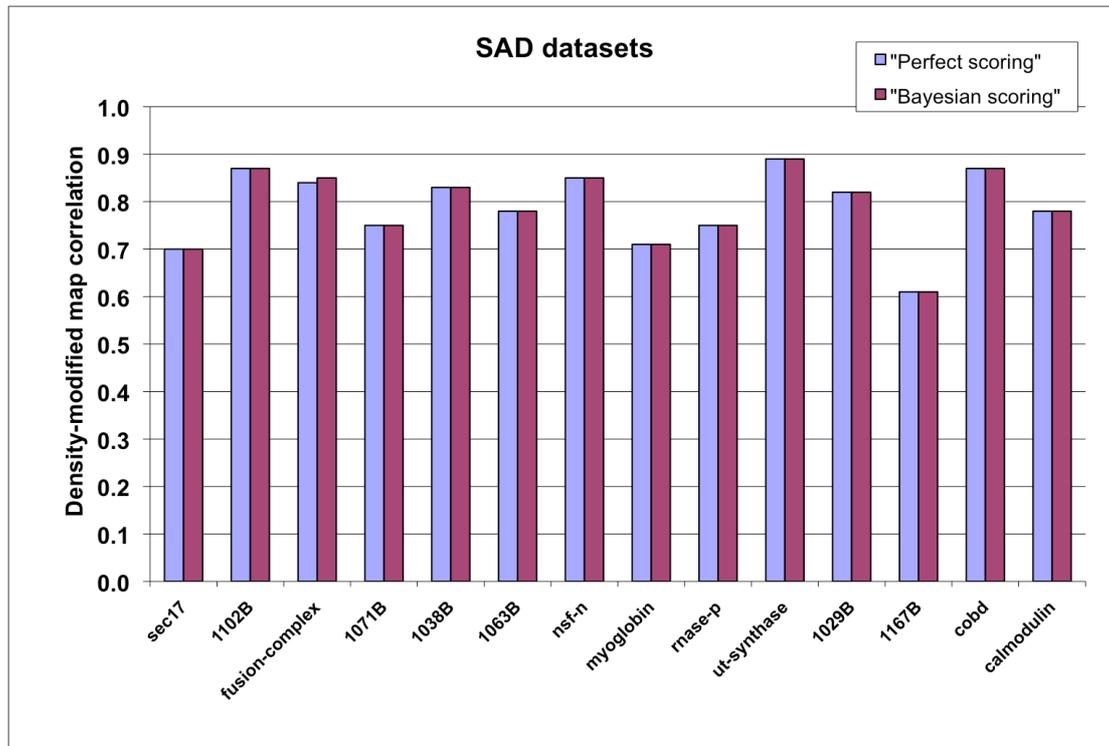


Fig. 5 C

