BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Refining a model electron-density map *via* the *Phantom Derivative* method

Maria Cristina Burla,[a,b] Benedetta Carrozzini,[b] Giovanni Luca Cascarano,[b] Carmelo Giacovazzo[b]* and Giampiero Polidori[b]

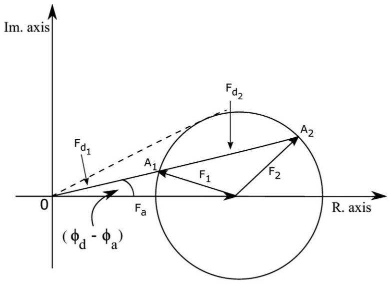
[a]Dipartimento di Fisica e Geologia, Università di Perugia, Via Pascoli, 06123 Perugia, Italy, and [b]Istituto di Cristallografia, CNR, Via G. Amendola 122/o, 70126 Bari, Italy. *Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 8 June 2015
Accepted 6 July 2015

Edited by Z. Dauter, Argonne National Laboratory, USA

Keywords: phasing; phase refinement; molecular replacement; *Phantom Derivative* method.

The *Phantom Derivative* (*PhD*) method [Giacovazzo (2015), *Acta Cryst.* A**71**, 483–512] has recently been described for *ab initio* and non-*ab initio* phasing. It is based on the random generation of structures with the same unit cell and the same space group as the target structure (called ancil structures), which are used to create derivatives devoid of experimental diffraction amplitudes. In this paper, the non-*ab initio* variant of the method was checked using phase sets obtained by molecular-replacement techniques as a starting point for phase extension and refinement. It has been shown that application of *PhD* is able to extend and refine phases in a way that is competitive with other electron-density modification techniques.



1. Introduction

The theoretical basis of a new phasing method, the *Phantom Derivative* (*PhD*) approach, has recently been published (Giacovazzo, 2015; referred to in the following as Paper I). According to this method, a large number of structures (called ancil structures, from the Latin *ancilla*) with the same unit-cell parameters and the same space group as the target structure, uncorrelated with each other and uncorrelated with the target, are randomly created. This may be achieved by the following simple algorithm.

The atomic positions of each ancil structure are randomly fixed by triples of random numbers in the interval (0, 1) without any control of interatomic distances and angles, and without any attempt to generate chemically consistent molecular fragments. In the case where ancils with the same scattering power as the target are preferred, the simplest algorithm is the following. Let $n_1, n_2, n_3, \ldots$ be the number of atoms in the asymmetric unit of the target structure for atomic species $1, 2, 3, \ldots$, respectively. Then, for each ancil structure, n_1 random atomic positions are associated with atomic species 1, n_2 with atomic species 2, n_3 with atomic species 3, *etc.*

It is also convenient for practical reasons to create ancils with the same average thermal factor as the target. This may be achieved by generating a Wilson plot using the target diffraction data and then assigning the average thermal factor thus found to all of the ancil atoms.

Let $\rho_a(j)$, $j = 1, \ldots, n$, be the electron densities of the ancil structures, and $|F_a(j)|$ and $\varphi_a(j)$ be the corresponding amplitudes and phases, both of which are perfectly known *a priori*. Also, let ρ be the unknown electron density of the target structure and let $|F|$ and φ be the corresponding amplitudes and phases. The $|F|$ values are known from a diffraction experiment, while the φ values are unknown.

http://dx.doi.org/10.1107/S1399004715013024

The *PhD* method aims to exploit the derivative electron densities

$$\rho_{\rm d}(j) = \rho + \rho_{\rm a}(j), \quad j = 1, \ldots, n, \qquad (1)$$

to phase the target. Since the ancil structures are not real structures (they are randomly generated) the derivative structures are also unreal, and therefore no experimental amplitude is available for them. This justifies the name the *Phantom Derivative* method.

The Fourier transform of (1) gives

$$F_{\rm d}(j) = F + F_{\rm a}(j). \qquad (2)$$

At beginning of the phasing process, the values of $|F_{\rm d}(j)|$, $\varphi_{\rm d}(j)$ and φ are unknown. It is therefore impossible at this stage to estimate the derivative amplitudes and phases by applying (2). The *PhD* method aims to use the experimental diffraction amplitudes of the target structure $|F|$, and the calculated diffraction amplitudes and phases of the ancil structures $|F_{\rm a}(j)|$ and $\varphi_{\rm a}(j)$, to progressively provide better estimates of derivative amplitudes and phases, from which better φ estimates should be progressively obtained. The method relies on a probabilistic approach, working in reciprocal space and on the following two algebraic conditions (see equation 1).

(i) For reflections for which $|F_{\rm a}(j)| \gg |F|$, $\varphi_{\rm a}(j)$ is a sufficiently good approximation of $\varphi_{\rm d}(j)$ and $|F_{\rm a}(j)|$ is a not too rough an approximation of $|F_{\rm d}(j)|$.

(ii) For reflections for which $|F| \gg |F_{\rm a}(j)|$, φ (even if unknown) is sufficiently close to $\varphi_{\rm d}(j)$ and $|F|$ is not too rough an approximation of $|F_{\rm d}(j)|$.

(1) clearly suggests the qualitative correctness of statements (i) and (ii). For example, if $|F| = 0$ then $\varphi_{\rm d}(j) = \varphi_{\rm a}(j)$ and $|F_{\rm d}(j)| = |F_{\rm a}(j)|$; if $|F_{\rm a}(j)| = 0$ then $|F_{\rm d}(j)| = |F|$ and $\varphi_{\rm d}(j) = \varphi$. A quantitative study of the errors (both in modulus and phase) involved in the approximations (i) and (ii) has been performed in paper I (see §§3 and 6), to which the reader is referred. We only mention here that errors in the amplitudes are more critical for the phasing process than phase errors, in agreement with the principle stated by direct methods, according to which phases are not lost in the diffraction experiments but are only hidden in the amplitudes.

The *PhD* approach has been designed to be a fully *ab initio* method because it only needs the experimental diffraction amplitudes of the target structure. However, the approach may also be used as a non-*ab initio* technique (see Paper I, §§8 and 11). This paper is dedicated to this second problem: we will check the potential of *PhD* by applying it to cases in which a model electron-density map is available from other phasing methods. This non-*ab initio* problem is more easy to attack: indeed, in this case the enantiomorph is perfectly determined by the model (compared with the ambiguity that exists for *ab initio* attempts), and φ estimates are available for all of the reflections (in contrast to the *ab initio* approach where only φ_{aj} phases, which are uncorrelated with φ, may be used). We will apply *PhD* to a large number of test structures with variable size and variable data resolution, the electron-density maps of which are provided by molecular-replacement (MR) techniques.

On the other hand, the *PhD* practice of using ancil structures that are completely uncorrelated with the target structure is maintained in our applications. We will verify whether *PhD*, by using random ancil structures, is really useful for driving the model phases closer to the target values. Thus, favourable results in this paper may prove the founding *PhD* conjecture: that random structures may be usefully employed for phasing a given target structure.

In §2, we will adapt the so-called sum function (see §8 of Paper I) to MR data and we will also optimize the various parameters that influence its efficiency. In §3, we will present and discuss the results obtained by applying *PhD* to the selected test structures.

2. The theoretical basis of *PhD* for non-*ab initio* approaches

Let us suppose that a model electron-density map $\rho_{\rm MR}$ is available at the end of an MR phasing procedure: $|F_{\rm MR}|$ and $\varphi_{\rm MR}$ are the corresponding calculated amplitudes and phases. We will suppose that the density is not immediately interpretable by automated model-building programs and that electron-density modification (EDM) techniques (Cowtan, 1994, 1999; Abrahams, 1997; Abrahams & Leslie, 1996; Refaat & Woolfson, 1993; Giacovazzo & Siliqi, 1997) are a necessary additional step to complete the crystal structure solution. It is not uncommon that even the most efficient EDM techniques are unable to provide phases of sufficient quality for a successful automated model-building process. Thus, new approaches such as *PhD* are welcome if they prove to be able to further reduce the average phase error corresponding to the best available electron-density map.

Let ρ_0 be the best electron-density map obtained by the application of an effective EDM procedure to the best $\rho_{\rm MR}$ map produced by the MR approach: it may be considered as a starting point for the application of *PhD*. $|F_0|$ and φ_0 represent the best amplitude and phase estimates obtained by Fourier inversion of ρ_0, respectively. The non-*ab initio* direct-space *PhD* approach suggests the creation of n ancil structures with the same unit cells and space group as the target structure and, correspondingly, the calculation of n derivative model electron densities,

$$\rho_{\rm d0}(j) = \rho_0 + \rho_{\rm a}(j), \quad j = 1, \ldots, n. \qquad (3)$$

If these n derivatives are summed into the sum function (see §8 of Paper I),

$$\rho_{\rm s} = \sum_{j=1}^{n} \rho_{\rm d0}(j) = n\rho_0 + \sum_{j=1}^{n} \rho_{\rm a}(j), \qquad (4)$$

$\rho_{\rm s}$ cannot provide any additional information on the target structure: indeed, the ancil structures are uncorrelated with each other and are uncorrelated with the target. Accordingly, the ancil contribution on the right-hand side of (4) is confined to the background of the sum map, and the map signal will coincide with an emphasized ρ_0 map. Expectations change if, in agreement with Paper I, each derivative electron density is submitted to EDM before being summed, and also if the corresponding sum function is itself submitted to EDM. Then, instead of (4), (5) should be used,

$$\rho_{\rm s} = \sum_{j=1}^{n} \rho_{\rm dmod}(j), \tag{5}$$

where $\rho_{\rm dmod}(j)$ is the jth EDM-modified derivative electron density. In turn, each $\rho_{\rm dmod}(j)$ may be interpreted as the sum of the original ancil densities $\rho_{\rm a}(j)$ with the jth modified electron density of ρ_0 [called $\rho_{0\rm mod}(j)$ on the right-hand side of equation 6],

$$\rho_{\rm s} = \sum_{j=1}^{n} \rho_{0\rm mod}(j) + \sum_{j=1}^{n} \rho_{\rm a}(j). \tag{6}$$

It may be expected that each $\rho_{0\rm mod}(j)$ will show structural features that were originally present in ρ_0 plus additional structural features generated by the EDM procedure. Summing the n $\rho_{0\rm mod}(j)$ maps will increase the contrast between a better target model and the background created by the ancil densities. A subsequent application of EDM to $\rho_{\rm s}$ may improve the current model even further.

The main problem in the application of (6) is that the derivative amplitudes are not experimentally known, and therefore the usual EDM procedures cannot be applied to refine derivative electron-density maps. In accordance with Paper I, we will use hybrid Fourier syntheses employing target amplitudes and derivative phases as EDM tools: the hope is that the target amplitudes will drive the derivatives phases closer to the target values.

In Paper I, a slightly different approach was also suggested. It is possible to write (6) as sum of difference maps,

$$\sum_{j=1}^{n} \rho_{0\rm mod}(j) = \sum_{j=1}^{n} [\rho_{\rm dmod}(j) - \rho_{\rm a}(j)], \tag{7}$$

which therefore should be submitted to EDM procedures. The two approaches are algebraically equivalent, but in practice they may provide slightly different results according to the behaviour of the EDM procedure. Indeed, in the case of (6), EDM is applied to a map that is expected to be positive everywhere; in the case of (7) the map may be negative in more extended regions of the unit cell.

(6) and (7) have a reciprocal-space counterpart which is summarized below. Taking the Fourier transform of (6) and (7) provides the amplitudes and phases of the new model, *i.e.*

$$F_{\rm ds} = \sum_{j=1}^{n} F_{\rm dmod}(j) = \sum_{j=1}^{n} F_{0\rm mod}(j) + \sum_{j=1}^{n} F_{\rm a}(j) \tag{8}$$

and

$$\sum_{j=1}^{n} F_{0\rm mod}(j) = \sum_{j=1}^{n} [F_{\rm dmod}(j) - F_{\rm a}(j)], \tag{9}$$

respectively, where the F values are obtained by Fourier inversion of the corresponding electron-density maps. For any given reflection (hkl) the $\varphi_{\rm a}(j)$ values, for $j = 1, \ldots, n$, are expected to be randomly distributed on the trigonometric circle so that the sum of the $F_{\rm a}(j)$ vectors, for $j = 1, \ldots, n$, is expected to be distributed about a null vector. On the contrary, $F_{0\rm mod}(j)$ vectors are light modifications of the target structural model and therefore will play a dominant role.

It is immediately clear that non-*ab initio PhD* is not an alternative to the present EDM methods and is itself an EDM method which has to be used in cooperation with other EDM techniques, possibly to overcome their present limitations. This belief will guide the applications described in this paper. The practical use of the non-*ab initio PhD* method requires the prior optimization of various parameters and choices between alternative variants. We refer to the following.

(i) The number of random ancil structures. In Paper I it was guessed that 100–300 ancil structures would be necessary for *ab initio* phasing. It should not be surprising that the simpler non-*ab initio PhD* approach should require a smaller number of ancils. We checked several n values and we found that in most cases 15 ancils are sufficient to significantly improve the initial structural model. Increasing this number improves the result but in a marginal way and with a larger computational cost. Thus, in all of our tests we will use $n = 15$.

(ii) The phasing process may be cyclical. As soon as the approach outlined above has generated a new target electron-density model (ρ_1), this density may be the starting point for a new application of *PhD* using other n additional and randomly created ancil structures. In this way, a second target model (ρ_2) may be obtained, and so on: in this cyclical approach each target model ρ_i is the starting point for cycle $i + 1$.

This procedure is not very rewarding, probably because some useless structural features that are obtained in the ith cycle are transmitted to cycle $i + 1$. The general effect is the following: the target phase estimates become stationary in spite of the larger computational cost. This approach has been abandoned.

(iii) Our preferred technique is the following. 15 ancil structures are randomly generated and are subdivided into three batches of five. *PhD* is then separately applied to each batch: at the end of the procedure $F_{\rm bat}(i)$, $i = 1, 2, 3$, are obtained, where $F_{\rm bat}(i)$ is the best target structure-factor estimate arising from the ith batch. The final target phase estimate for a given reflection **h** arises by applying the tangent formula to the three $F_{\rm bat}(i)$ estimates. As in (ii), a larger number of ancil structures may be used, but the advantages are marginal.

(iv) It may be worthwhile noting that n ancil structures may be used to generate both the n derivative densities (6) and the following ones,

$$\bar{\rho}_{\rm s} = \sum_{j=1}^{n} \bar{\rho}_{\rm dmod}(j) = \sum_{j=1}^{n} \rho_{0\rm mod}(j) - \sum_{j=1}^{n} \rho_{\rm a}(j). \tag{10}$$

$\bar{\rho}_{\rm s}$ is now negative in extended regions of the unit cell. The $\rho_{\rm a}(j)$ densities are again randomly distributed and therefore become, as in (6), part of the background of the sum map. We verified that the use of (10) leads to results that are very similar to those obtained *via* (6) and we therefore renounced the use of (10).

(v) We separately checked the efficiency of (6) and (7) on all of our test structures. The corresponding results were very

Table 1
The PDB code (CODE) of the target structure, the data resolution (RES), the number of residues in the asymmetric unit of the target (NresT), the mean phase error obtained by MR ($\langle|\Delta\varphi_{\rm MR}|\rangle$), the mean phase error obtained by applying *DM* to the set of observed reflections phased by MR ($\langle|\Delta\varphi_{\rm DM}|\rangle$) and the mean phase error obtained by applying *PhD* to the reflections phased by *DM* ($\langle|\Delta\varphi_{\rm PhD}|\rangle$) for each test structure.

CODE	RES (Å)	NresT	$\langle\|\Delta\varphi_{\rm MR}\|\rangle$ (°)	$\langle\|\Delta\varphi_{\rm DM}\|\rangle$ (°)	$\langle\|\Delta\varphi_{\rm PhD}\|\rangle$ (°)
1dy5	0.87	248	74	76	21
1bxo	0.90	323	74	71	21
2fc3	1.54	124	54	49	35
1tgx	1.55	180	58	54	41
2a46	1.65	217	69	57	37
1lys	1.72	258	53	51	45
1cgo	1.79	127	74	66	55
2otb	1.79	432	58	54	42
1kqw	1.80	134	59	54	43
2sar	1.85	192	52	47	37
1lat	1.90	145	70	68	66
1e8a	1.95	175	69	57	48
2f53	1.99	811	59	52	45
2ayv	2.00	148	53	48	43
2pby	2.07	1155	79	75	72
2f8m	2.09	472	64	57	51
1yxa	2.10	740	74	69	66
2f84	2.10	323	55	49	44
1cgn	2.15	125	73	65	57
1xyg	2.19	1380	64	58	54
2a4k	2.30	439	60	55	47
2b5o	2.50	584	50	49	45
1ycn	2.51	619	56	50	44
2iff	2.58	555	62	61	61

similar: a slightly smaller phase error may be obtained by combining both of the estimates, but the increased computing time did not encourage us to follow this practice. Thus, in all of our tests we applied (6) only.

(vi) We also checked the usefulness of the reciprocal *PhD* variant involving the use of (8) or (9). The results were greatly inferior to those obtained *via* (6) and (7): this result confirms the historical superiority of direct-space EDM procedures with respect to reciprocal-space variants.

(vii) The general *PhD* theory allows great freedom in the choice of the scattering power of the ancil structures. In the case of *ab initio* phasing, special considerations suggest the use of ancils with a scattering power equal to that of the target. In our non-*ab initio* case the choice is more free: we checked the efficiency of *PhD* by using ancils with scattering powers that were smaller than, equal to or larger than that of the target. The results are not very different, provided that the ratio of the scattering powers (ancil:target) is between 0.3 and 1.5. We choose ancils with a scattering power that was half of that of the target for our tests.

(viii) In analogy with Patterson deconvolution methods, we also applied the *PhD* procedure by replacing the sum function by the minimum function (which takes the minimum value calculated for the derivative electron densities pixel by pixel). The results were slightly worse than those obtained *via* the sum function, which remains our standard choice.

Thus far we have mentioned EDM procedures without explicitly specifying the technique and the programs that we use. Phase extension and refinement are performed by two EDM procedures: *DM* (Cowtan, 1994, 1999; available from *CCP*4; Winn *et al.*, 2011) and *DSR* (Giacovazzo & Siliqi, 1997; Caliandro *et al.*, 2014). Both apply real-space constraints to the electron-density map to meet the expected protein features. *DM* is very popular in protein crystallography and is highly effective at resolutions that are worse than atomic, but is inefficient when the data resolution is better than 1.4 Å (for specific test cases, see Carrozzini *et al.*, 2013). It automatically stops when a suitable figure of merit overcomes a given limit.

DSR is more suitable for high-resolution data (*e.g.* up to about 1.8 Å) and will only be used in our tests for the two test structures with data at atomic resolution.

3. Applications

In order to check the usefulness of the non-*ab initio* *PhD* variant, we used 24 test structures, a subset of the 45 structures used by Carrozzini *et al.* (2013) for checking an MR pipeline. The selected set only involves the cases for which the MR step provided phase errors larger than 50°. Their Protein Data Bank (PDB) codes are quoted in Table 1, together with their data resolution (RES) and the number of residues in the asymmetric unit (NresT). $\langle|\Delta\varphi_{\rm MR}|\rangle$ is the average phase error at the end of the MR step.

We extended and refined the phases by the default use of *DM*: the $\langle|\Delta\varphi_{\rm DM}|\rangle$ column in Table 1 shows the *DM* results in terms of average phase error. This column gives a state-of-the-art standard with which to compare the efficiency of *PhD*. The *PhD* procedure described in §2 was then applied to the phases refined by *DM* to check whether they may subsequently improve. The corresponding average phase errors are shown in the $\langle|\Delta\varphi_{\rm PhD}|\rangle$ column. We notice the following.

(i) *DM* used in default mode worked quite usefully in most of the cases: the phases were efficiently extended and refined, and in most of the cases $\langle|\Delta\varphi_{\rm DM}|\rangle$ is significantly better than $\langle|\Delta\varphi_{\rm MR}|\rangle$. *DM* is, however, highly inefficient when applied to PDB entries 1dy5 and 1bxo, the two test structures with atomic resolution data.

(ii) *PhD* starting from the *DM* phases significantly improved their quality. In all cases $\langle|\Delta\varphi_{\rm PhD}|\rangle \le \langle|\Delta\varphi_{\rm DM}|\rangle$, and in most cases the difference is large: the average phase error is in some cases reduced by 20°. Correspondingly (for brevity we do not give the figures), the map correlation between the target electron-density map corresponding to the published structure and the final density map obtained by *PhD* is significantly larger than that with the map obtained by *DM*. This makes the automatic interpretation of the density maps by automated model-building programs (AMB) more easy.

The obvious conclusion is as follows: even if *PhD* may conveniently start from MR phases, it may be more useful when applied to phases refined by *DM*. It introduces information that is not available for *DM* into the refinement step.

We now observe that the efficiency of any EDM technique depends on the quality of the initial phases. Generally speaking, if the phases are too far away from the true values then EDM techniques are not able to reduce the phase error (indeed, EDM techniques are not *ab initio* techniques). If the

Table 2
The PDB code (CODE), the mean phase error obtained by applying *VLD* + *FL* to the set of reflections phased by MR ($\langle|\Delta\varphi_{\rm V+F}|\rangle$) and the mean phase error obtained by the supplementary application of *PhD* to the phases refined by *VLD* + *FL* ($\langle|\Delta\varphi_{\rm V+F+P}|\rangle$) for each test structure.

CODE	$\langle\|\Delta\varphi_{\rm V+F}\|\rangle$ (°)	$\langle\|\Delta\varphi_{\rm V+F+P}\|\rangle$ (°)
1dy5	22	21
1bxo	20	19
2fc3	35	33
1tgx	41	38
2a46	31	26
1lys	44	43
1cgo	58	49
2otb	42	39
1kqw	41	39
2sar	37	34
1lat	66	66
1e8a	49	45
2f53	46	46
2ayv	43	43
2pby	72	72
2f8m	51	49
1yxa	67	66
2f84	45	45
1cgn	57	52
1xyg	54	54
2a4k	45	44
2b5o	43	44
1ycn	44	44
2iff	63	64

initial set of phases is of very high quality, EDM techniques are inefficient and superfluous (in these conditions the phase error may only decrease if molecular models rather than electron-density maps are used).

It may thus be of interest to check whether *PhD* may subsequently improve the quality of more deeply refined sets of phases, for example phases already treated with *VLD* (Burla *et al.*, 2011) and subsequently refined by *free lunch* (referred to as *FL* in the following; Caliandro *et al.*, 2005, 2007) techniques. We will refer to this method as *VLD* + *FL*.

For the benefit of the reader, we recall that *VLD* is based on the difference Fourier synthesis

$$E_q \simeq (mR - R_p)\exp(i\varphi_p),$$

where m is a suitable weight and R and R_p are the normalized structure-factor moduli corresponding to the observed and calculated model amplitudes, respectively. E_q, obtained by Fourier inversion of the difference electron-density map, is then added to E_p to estimate the phase of a new model electron density *via* the tangent formula

$$\tan\varphi = \frac{R_p\sin\varphi_p + w_q R_q \sin\varphi_q}{R_p\cos\varphi_p + w_q R_q \cos\varphi_q},$$

where w_q is a suitable weight. This new density is then resubmitted to additional *DM* cycles.

The application of *FL* requires the extrapolation of the amplitudes and phases of a large number of non-measured reflections both beyond and behind the experimental resolution. The extrapolation limit is determined by the observed data resolution (it makes no sense to extrapolate reflections up to 1 Å resolution when the observed resolution is 3 Å). Indeed, the number of extrapolated reflections actively used in the electron-density maps is never greater than twice the number of observed reflections.

As in our previous papers, the *VLD* + *FL* approach implies that the best map produced by *VLD* is the initial map for the *FL* step. It is also useful to notice that *VLD* + *FL* is a short abbreviation of the sequence *DM*–*VLD*–*DM*–*FL*, since *DM* is an important step of the technique (see Carrozzini *et al.*, 2013).

In additional calculations, as described below, instead of applying *DM* to MR phases we apply *VLD* + *FL*, and we then use the final phases as a starting point for *PhD*. The *VLD* + *FL* average phase errors obtained by Carrozzini and coworkers for our test structures are shown in Table 2 (in the $\langle|\Delta\varphi_{\rm V+F}|\rangle$ column). A simple comparison of the $\langle|\Delta\varphi_{\rm V+F}|\rangle$ column in Table 2 with the $\langle|\Delta\varphi_{\rm DM}|\rangle$ column in Table 1 suggests the greater efficiency of *VLD* + *FL* with respect to the simple application of *DM*: indeed, in most of the cases $\langle|\Delta\varphi_{\rm V+F}|\rangle$ is significantly better than $\langle|\Delta\varphi_{\rm DM}|\rangle$.

To check whether *PhD* may be able to subsequently improve a set of phases previously refined by *VLD* + *FL*, we integrated the two techniques: the resulting approach will be referred as *VLD* + *FL* + *PhD*. The *PhD* step includes a final application of *FL* at its end.

At the end of the *PhD* procedure we have two different (although correlated) phase estimates for each reflection: the value obtained at the end of the *PhD* procedure and the initial value, which is also expected to be of high quality. The two estimates are then combined *via* a tangent technique: the corresponding average phase error for each test structure is reported in Table 2 in the $\langle|\Delta\varphi_{\rm V+F+P}|\rangle$ column. We notice the following.

(i) In some cases $\langle|\Delta\varphi_{\rm V+F}|\rangle$ is too small for further significant improvements (PDB entries 1dy5 and 1bxo).

(ii) In most cases $\langle|\Delta\varphi_{\rm V+F+P}|\rangle < \langle|\Delta\varphi_{\rm V+F}|\rangle$ by a few degrees, but in some cases the difference is considerable. The closeness between the two methods is also owing to the common use of tools such as *VLD* and *FL*.

(iii) Some structures (PDB entries 1lat, 2pby, 1yxa, 1xyg and 2iff) are resistant to any effort. Neither *VLD* nor *PhD* are able to obtain average phase errors that are better than those obtained by *DM*. The reason for this is not completely clear, but some deeper insight is given below.

(iv) The *PhD* average phase errors obtained using the *VLD*-refined phases as initial phases are significantly smaller than those obtained when *PhD* starts from *DM*-refined phases. Accordingly, the quality of the starting phases strongly influences the quality of the final *PhD* phases.

The above results suggest that the most effective procedure for refining MR phases is to combine *DM* with the *VLD*, *FL* and *PhD* approaches. However, we have to check whether the phase improvement obtained by this method is significant in terms of whether it would allow a structure to be solved when it remains unsolved by applying *DM* techniques only. Let $\rm CORR_{DM}$, $\rm CORR_{PhD}$ and $\rm CORR_{V+F+P}$ be the correlation values between the electron-density map corresponding to the published structure and the maps obtained by the application

Table 3
$\mathrm{CORR}_{\mathrm{DM}}$, $\mathrm{CORR}_{\mathrm{PhD}}$, $\mathrm{CORR}_{\mathrm{V+F+P}}$ and the corresponding free *R* values for each test structure.

CODE	$\mathrm{CORR}_{\mathrm{DM}}$	$\mathrm{Rf}_{\mathrm{DM}}$	$\mathrm{CORR}_{\mathrm{PhD}}$	$\mathrm{Rf}_{\mathrm{PhD}}$	$\mathrm{CORR}_{\mathrm{V+F+P}}$	$\mathrm{Rf}_{\mathrm{V+F+P}}$
1dy5	0.67	0.45	0.95	0.24	0.95	0.28
1bxo	0.64	0.29	0.95	0.25	0.96	0.25
2fc3	0.80	0.36	0.90	0.30	0.91	0.29
1tgx	0.76	0.43	0.85	0.36	0.86	0.33
2a46	0.69	0.38	0.88	0.33	0.94	0.29
1lys	0.78	0.33	0.82	0.27	0.83	0.26
1cgo	0.54	0.49	0.69	0.36	0.74	0.31
2otb	0.78	0.35	0.87	0.32	0.88	0.34
1kqw	0.74	0.32	0.83	0.36	0.85	0.39
2sar	0.79	0.30	0.86	0.28	0.88	0.28
1lat	0.51	0.52	0.53	0.56	0.52	0.59
1e8a	0.65	0.34	0.78	0.31	0.81	0.29
2f53	0.80	0.34	0.84	0.35	0.82	0.35
2ayv	0.83	0.34	0.87	0.34	0.87	0.32
2pby	0.49	0.48	0.50	0.53	0.50	0.52
2f8m	0.70	0.33	0.75	0.32	0.76	0.33
1yxa	0.55	0.50	0.58	0.50	0.59	0.52
2f84	0.77	0.34	0.81	0.32	0.81	0.31
1cgn	0.60	0.56	0.68	0.46	0.73	0.27
1xyg	0.69	0.30	0.72	0.29	0.71	0.31
2a4k	0.71	0.37	0.78	0.33	0.80	0.30
2b5o	0.78	0.33	0.82	0.32	0.82	0.33
1ycn	0.77	0.32	0.82	0.32	0.81	0.30
2iff	0.54	0.47	0.52	0.47	0.48	0.47

Table 4
$\langle|\Delta\varphi_{\mathrm{MR}}|\rangle$ (as quoted in Table 1) and the average phase errors $\langle|\Delta\varphi_{\mathrm{R}}|\rangle$, $\langle|\Delta\varphi_{\mathrm{R+V+F}}|\rangle$ and $\langle|\Delta\varphi_{\mathrm{R+V+F+P}}|\rangle$ for each test structure.

CODE	$\langle|\Delta\varphi_{\mathrm{MR}}|\rangle$ (°)	$\langle|\Delta\varphi_{\mathrm{R}}|\rangle$ (°)	$\langle|\Delta\varphi_{\mathrm{R+V+F}}|\rangle$ (°)	$\langle|\Delta\varphi_{\mathrm{R+V+F+P}}|\rangle$ (°)
1dy5	74	21	21	22
1bxo	74	34	34	19
2fc3	54	34	33	31
1tgx	58	36	35	34
2a46	69	40	37	23
1lys	53	29	29	32
1cgo	74	57	53	35
2otb	58	37	36	32
1kqw	59	35	34	31
2sar	52	44	41	33
1lat	70	61	59	50
1e8a	69	40	39	36
2f53	59	30	29	36
2ayv	53	33	32	33
2pby	79	45	43	35
2f8m	64	44	42	37
1yxa	74	45	43	38
2f84	55	37	36	36
1cgn	73	49	46	30
1xyg	64	42	40	38
2a4k	60	37	35	31
2b5o	50	38	37	36
1ycn	56	33	32	33
2iff	62	65	64	79

of the three procedures defined by the subscripts. Also let $\mathrm{Rf}_{\mathrm{DM}}$, $\mathrm{Rf}_{\mathrm{PhD}}$ and $\mathrm{Rf}_{\mathrm{V+F+P}}$ be the free *R*-factor values (Brünger, 1992) obtained by applying the automatic model-building program *Buccaneer* (Cowtan, 2006) to the final sets of phases obtained by the same three procedures. We assume that Rf values smaller than 0.40 may correspond to solved structures, while Rf values significantly larger than 0.40 indicate unsolved structures. The results are shown in Table 3, where we give the corresponding CORR and Rf values for each test structure.

The reader will immediately observe that the map correlations increase when *DM* is replaced by *DM* + *PhD* and when this is replaced by *DM* + *VLD* + *FL* + *PhD*: the new procedures therefore improve the quality of the final electron-density maps.

It may also be noticed that eight test structures may remain unsolved when only *DM* is applied (*i.e.* PDB entries 1dy5, 1tgx, 1cgx, 1lat, 2pby, 1yxa, 1cgn and 2iff). For 1tgx the value of $\mathrm{Rf}_{\mathrm{DM}}$ is however not significantly larger than 0.40: indeed, the phase error corresponding to the *Buccaneer* structural model is 41° and the coverage index is 0.82. The remaining seven structures have a final phase error of greater than 55°.

When *PhD* is combined with *DM* only five test structures show $\mathrm{Rf}_{\mathrm{PhD}}$ values larger than 0.40 (*i.e.* PDB entries 1lat, 2pby, 1yxa, 1cgn and 2iff). For 1cgn $\mathrm{Rf}_{\mathrm{PhD}}$ is not much greater than 0.40: indeed, the phase error corresponding to the *Buccaneer* structural model is 47° and the coverage index is 0.88.

The number of structures with $\mathrm{Rf}_{\mathrm{V+F+P}}$ greater than 0.40 decreases to four (*i.e.* PDB entries 1lat, 2pby, 1yxa and 2iff): $\mathrm{Rf}_{\mathrm{V+F+P}}$ for 1cgn is now 0.27 and the corresponding *Buccaneer* coverage index is 0.95.

Thus far, we have described procedures which may also be applied to sets of phases generated by any other *ab initio* or non-*ab initio* approach: the procedures only need a model electron-density map and the observed diffraction amplitudes. Indeed, to start the phase-refinement processes described above the model electron-density map may be Fourier inverted to obtain model phase values which will constitute the starting point for the *VLD* + *FL* + *PhD* procedure. In fact, a molecular model is not necessary.

MR, however, ends with a molecular model that is suitably oriented and translated: the question is whether the availability of a molecular model may be used to subsequently reinforce the *VLD* + *FL* + *PhD* refinement procedure.

In a recent paper (Carrozzini *et al.*, 2015) aiming at solving *via* crystallographic techniques a set of structures originally solved (DiMaio *et al.*, 2011) by combining MR with a suite using physically realistic all-atom potential functions (*Rosetta*; see Das & Baker, 2009), an extraordinarily intensive use of *REFMAC* (Murshudov *et al.*, 2011) is described. Such intensive application is one of the phasing tools which allowed Carrozzini and coworkers to succeed (without *Rosetta*) in phasing most of the difficult test structures proposed by Di Maio and coworkers. The main problem in such intensive use is to decide at which *REFMAC* cycle refinement should be stopped, otherwise overrefinement will usually lead *REFMAC* to diverge.

For the benefit of the reader, we report the basic criterion adopted by Carrozzini *et al.* (2015). The crystallographic residual R_{cryst} is calculated every 15 cycles. At cycle $15(n + 1)$ R_{cryst} is compared with that calculated at cycle $15n$. If the value at cycle $15(n + 1)$ is larger, and if the average absolute difference between the phase values at cycle $15n$ and the phase values at cycle $15(n - 1)$ is less than 3°, then the program stops and the phases obtained at cycle $15n$ are restored. The above criterion often allows a large number of *REFMAC* cycles, even greater than 150, but owing to the robustness of the *REFMAC* algorithms this overwork often leads to significant phase improvement.

A new procedure then may be invoked that is specifically designed for MR cases. The first step of the refinement procedure involves the intensive application of *REFMAC* to MR phases. Let $\langle|\Delta\varphi_{\rm R}|\rangle$ be the average phase error corresponding to this step. The phases from the last cycle of *REFMAC* are then submitted to *VLD* + *FL*, thus producing new phase estimates with a mean error denoted $\langle|\Delta\varphi_{\rm R+V+F}|\rangle$. Alternatively, the *REFMAC* phases are submitted to *VLD* + *FL* + *PhD*, which ends with new phase estimates: let $\langle|\Delta\varphi_{\rm R+V+F+P}|\rangle$ be their average phase error (again the *DM* step is contained in the *VLD* notation). For each test structure in Table 4, we report for the benefit of the reader the value $\langle|\Delta\varphi_{\rm MR}|\rangle$ already quoted in Table 1 and the average phase errors $\langle|\Delta\varphi_{\rm R}|\rangle$, $\langle|\Delta\varphi_{\rm R+V+F}|\rangle$ and $|\Delta\varphi_{\rm R+V+F+P}|\rangle$. We notice the following.

(i) In no case does our algorithm allow *REFMAC* to diverge: for almost all of the test structures the phase improvement is surprisingly high. The number of cycles varies from a minimum of 30 to a maximum of 180, which is reached for PDB entry 2pby: in this last case the large number of cycles is necessary to decrease the average phase error from 79 to 45°. However, the reader should consider that in 11 cases the algorithm stops *REFMAC* at a cycle number equal to or greater than 90.

(ii) All of the structures benefit by the intensive use of *REFMAC*, including four of the five resistant test cases (PDB entries 1lat, 2pby, 1yxa and 1yxg). Only 2iff does not improve, for the following reason. The scattering power of the 2iff MR model is a small percentage (about 23%) of that of the target: therefore, the use of *REFMAC* may improve the geometry of the monomer but cannot generate other monomers.

(iii) The application of *DM* to *REFMAC* phases lowers the average phase error by a few degrees. A significant improvement with respect to the *DM* result is obtained by applying *VLD* + *FL* + *PhD* to *REFMAC* phases. The average phase error for all of the test structures is now always smaller than 40° (except for 2iff).

It may now be appropriate to apply *Buccaneer* to the structures refined by *REFMAC* + *VLD* + *FL* + *PhD* to verify its efficiency with the model maps provided by this procedure. For the sake of brevity, we do not report its full outcome in Table 3: indeed, the final average phase errors shown in Table 4 are in general very small, and in these cases *Buccaneer* easily succeeds. We only mention the key values obtained for the four test structures (PDB entries 1lat, 2pby, 1yxa and 2iff) that remained unsolved when the procedure *VLD* + *FL* + *PhD* was applied. Only two structures remain now unsolved, 1lat and 2iff, with Rf values of 0.51 and 0.50, respectively. However, 1lat now has a better model, with a coverage index of 0.48 compared with 0.12 obtained *via VLD* + *FL* + *PhD* and a mean phase error of 61° compared with 76°. 2pby and 1yxa are now fully solved: the Rf values are 0.27 and 0.30, respectively, with corresponding coverage indices of 0.96 and 0.94.

4. Conclusions

The *PhD* method has been tested in its non-*ab initio* variant. We selected a set of test structures for which the average phase error obtained *via* MR is larger than 50° and we checked whether *PhD* is able to reduce the average phase error and therefore to improve the quality of the model. We compared the efficiency of *PhD* with other EDM techniques such as *DM* and *VLD*. Our calculations clearly show the capacity of *PhD* to extend and refine phases in a manner that is competitive with any other current approach. Indeed, the best results were obtained by integrating *PhD* with the *DM*, *VLD* and *FL* techniques.

A procedure specifically designed for MR has been tested: it involves intensive use of *REFMAC*, followed by *DM*, *VLD*, *FL* and *PhD* steps. The procedure improves the previous results even further, and may be considered the most effective technique to apply when well oriented and translated molecular models are obtained by MR.

Finally, a few conclusive words about the role of this paper. Its experimental results prove the founding conjecture of Paper I: that structures that are completely uncorrelated among themselves, and are uncorrelated with the target, are able to improve the current target phase estimates. Owing to such a result, this conjecture may be now transformed into a formal statement, which establishes the following principle: information additional to that contained in the target diffraction amplitudes may be found in structures that are completely uncorrelated with the target.

This is the most radical way to increase the information contained in the diffraction data: milestones on this pathway are MR (information contained in a molecular fragment similar to the target molecule), SIR/MIR (additional information provided by the derivative diffraction data), SAD/MAD (a further case of isomorphism) and *free lunch* (additional information obtained by the extrapolation of non-measured amplitudes). It may therefore be foreseen that *PhD* may find useful applications when combined with any *ab initio* or non-*ab initio* phasing technique, including EDM techniques other than *DM*: this paper just opens the way.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 1143–1151.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Comunale, G., Giacovazzo, C. & Mazzone, A. (2014). *Acta Cryst.* D**70**, 1994–2006.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Moustiakimov, M. & Siliqi, D. (2005). *Acta Cryst.* A**61**, 343–349.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 931–937.

Carrozzini, B., Cascarano, G. L., Comunale, G., Giacovazzo, C. & Mazzone, A. (2013). *Acta Cryst.* D**69**, 1038–1044.

Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Mazzone, A. (2015). *Acta Cryst.* D**71**, 1856–1863.

Cowtan, K. D. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Cowtan, K. (1999). *Acta Cryst.* D**55**, 1555–1567.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Das, R. & Baker, D. (2009). *Acta Cryst.* D**65**, 169–175.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London)*, **473**, 540–543.

Giacovazzo, C. (2015). *Acta Cryst.* A**71**, 483–512.

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.