Benchmark problems for phase retrieval

Veit Elser and Ti-Yen Lan

***Abstract*—The hardest instances of phase retrieval arise in crystallography, where the signal is periodic and comprised of atomic distributions arranged uniformly in the unit cell of the crystal. We have constructed a graded set of benchmark problems for evaluating algorithms that perform this type of phase retrieval. A simple iterative algorithm was used to establish baseline runtimes that empirically grow exponentially in the sparsity of the signal autocorrelation. We also review the algorithms used by the leading software packages for crystallographic phase retrieval.**

***Index Terms*—phase retrieval, periodic signals, reconstruction algorithms**

I. Introduction

In the last ten years there has been a revival of interest in the problem of phase retrieval [PR]. By far the largest application of phase retrieval algorithms is crystallography. In 2016 about 50,000 crystal structures were deposited in the Cambridge Structural Database [CSD], each made possible by a “phasing algorithm” that reconstructs a periodic signal — the electron density in a crystal — from its Fourier magnitudes. Structures in the database are of modest size, and the diffraction data is of sufficient quality that individual atoms — about 100 per unit cell on average — can be resolved as a sparse signal. Signal sparsity in these applications of phase retrieval offsets the data deficiency in magnitude measurement; it is also a common theme in the current wave of phase retrieval research. What is puzzling, and the motivation for the work described here, is the absence of demonstrations of the new methods in the domain where it is most used.

Phase retrieval also finds application outside of crystallography, in the reconstruction of aperiodic signals. The production of diffraction data from individual (non-crystallized) biomolecules, made possible by X-ray free electron lasers, is expected to see significant development over the coming decade [SPI]. But this new source of data is unlikely to supplant crystallography because crystals, when available, are still by far the easiest means for amplifying the inherently weak signal of a single biomolecule. Phase retrieval for aperiodic signals is significantly easier, and corresponds to knowing all the atoms in the unit cell are located in a given sub-volume. The main challenge faced by practitioners of “single particle imaging” is not this easier form of phase retrieval, but dealing with a weak, shot noise limited signal in the presence of background.

The reluctance of contemporary phase retrieval researchers to study periodic signals may simply be the result of not appreciating the significance of periodicity on the complexity of the problem. Not having to deal with complicating factors incidental to phase retrieval, *e.g.* space groups, may also be a contributing factor. In any case, this state of affairs is easily addressed by making available instances of phase retrieval[1] that (*i*) are seen as challenging by crystallographers and (*ii*) are presented with an eye toward accessibility for non-crystallographers. Below we describe the construction of a set of benchmark problems with these characteristics. In addition to providing a basis for comparing different algorithms, the graded difficulty of the instances will provide evidence of the complexity behavior of individual algorithms. We did not think it was necessary to construct benchmark problems for aperiodic signals because periodic signals are the harder case and the leading application.

II. Description of the datasets

Here we describe our synthetic datasets and what it means to solve an instance. Details on the construction of the benchmark problems are given in the next section and can be skipped by readers just wishing to solve the benchmarks.

Data sets are identified by an integer N, the number of atoms in the signal, and a suffix characterizing difficulty: E (easy), M (medium), H (hard). All data have the same format: an $M \times M$ table of integers (photon counts) representing the measurements of Fourier intensities $|\hat{\rho}|^2$ of a signal ρ sampled on a periodic $M \times M$ grid. All instances have $M = 128$. This size was chosen to discourage methods that represent the signal in terms of a dense matrix on which a rank-1 (or low rank) constraint is imposed. In protein crystallography the corresponding size is even larger. Regarding the dimensionality of the signal, we believe this has no effect on phase retrieval complexity in the periodic case; two dimensions was chosen only for ease of rendering the signal.

Figure 1 shows a rendering and excerpt of the data file for the easiest instance, **`data100E`**. The table of photon counts contains several zero entries because intensities are normally measured out to frequencies where the Fourier transform is small in magnitude (and few photons are detected). The $(0, 0)$ intensity is never measured and appears as a 0 in the data file. All other intensities have been symmetrized, $|\hat{\rho}(p, q)|^2 = |\hat{\rho}(-p, -q)|^2$, because the electron density signal of X-ray crystallography is real-valued. The data files comprise just the 128×64 half-table of symmetrized photon counts.

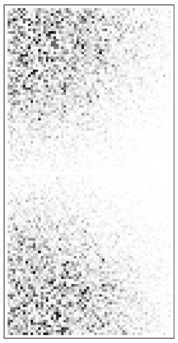
Solving an instance entails the following. Square roots of the data are taken and define the Fourier magnitudes $|\hat{\rho}(p, q)|$. The phasing algorithm being demonstrated reconstructs $\hat{\rho}(0, 0) > 0$ and the phases $\phi(p, q)$ of the periodic signal

$$\rho(x, y) = \frac{1}{\sqrt{M^2}} \sum_{p=0}^{M-1} \sum_{q=0}^{M-1} e^{i \frac{2\pi}{M}(px+qy)} \, |\hat{\rho}(p, q)| e^{i\phi(p,q)}. \quad (1)$$

V. Elser and T.-Y. Lan are with the Department of Physics, Cornell University, Ithaca, NY, 14853-2501 USA e-mail: ve10@cornell.edu.

This work received support from DOE grant de-sc0005827 and a Taiwan Government Scholarship to Study Abroad.

[1] **https://github.com/veitelser/phase-retrieval-benchmarks**



0	41	…	11	14
93	586	…	2	53
126	226	…	12	5
78	961	…	11	44
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
171	47	…	18	10
78	125	…	10	3
126	617	…	0	17
93	303	…	0	9

Fig. 1. Rendering (left) and excerpt (right) of benchmark instance **data100E**. The $(0, 0)$ photon count at the upper left corner is not measured and set to zero.

The solution phases must satisfy $\phi(p, q) = -\phi(-p, -q)$ in order that $\rho(x, y)$ is real.

Solutions are required to be consistent with a prior constraint on the support S of the signal. The cardinality of S is $8N$, that is, on average each of the N atoms is supported on 8 pixels. In practice, S is the set of pixels on which $\rho(x, y)$ has its $8N$ largest values because the signal is not only real but non-negative. Consistency with the support constraint is established by checking a power inequality. From the data, as well as the reconstructed $(0, 0)$ intensity, the total Fourier power is given by

$$I_{\mathrm{F}} = \sum_{p=0}^{M-1} \sum_{q=0}^{M-1} |\hat{\rho}(p, q)|^2. \tag{2}$$

In a successful reconstruction the power in the support

$$I_{\mathrm{S}} = \sum_{(x,y)\in S} |\rho(x, y)|^2, \tag{3}$$

matches the Fourier power. Because of noise the power in the support falls short of the power in the data. However, all the benchmark instances have solutions that satisfy

$$\frac{I_{\mathrm{S}}}{I_{\mathrm{F}}} > 0.95. \tag{4}$$

An instance is declared to be solved when this criterion is met. All noise in the benchmarks arises from the discreteness (Poisson sampling) of the Fourier intensity measurements (details in Section III).

We emphasize that phase retrieval need not be performed on $M \times M$ grids. Computations may be performed on finer or coarser grids, or without a grid sampling of the signal at all. However, at the end of the computation the algorithm is required to output the phase angles and $\hat{\rho}(0, 0)$ needed to check criterion (4).

III. Construction details

Figure 2 shows a successfully reconstructed signal for benchmark instance **data300E** next to the signal that was constructed by hand to produce the data (a translation and inversion through the origin was applied to the latter to aid comparison). Of the 128^2 pixels, only 8×300 have significant signal (appear gray). These are bandlimited signals and less noisy than they appear. After zero-padding their Fourier transforms on a 512×512 grid and back transforming, contour plots of the resulting signals in Figure 3 reveal "atoms" of two types with relatively precise positions.

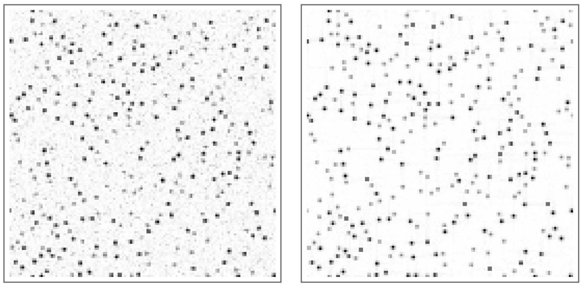


Fig. 2. A phase retrieval solution (left) and signal as constructed by hand (right) for benchmark **data300E**. These images are rendered on a grid of the same size (128×128) as the grid that holds the data.

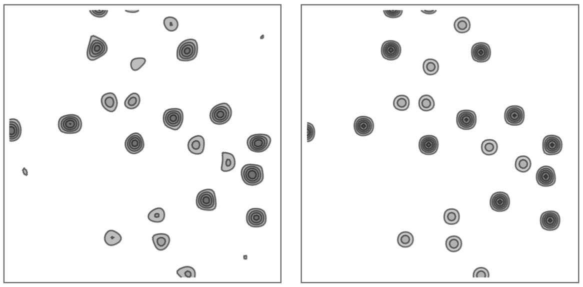


Fig. 3. Corresponding contour renderings of the upper left corners of the signals in Figure 2.

To construct data for an instance with N atoms, first a set of N pixels on a 512×512 gird was sequentially sampled, uniformly but with the constraint that the distance between pixels is at least 12. After downsampling the signal by a factor of 4, this gives a minimum separation of 3 pixels between atom centers or about 3 Å units in data where the pixel resolution is 1 Å. The signal value was set to 1 on half of the N selected pixels and 2 on the other half. This mimics two species with atomic number ratio 2. Minor variants of this part of the construction, described below, give the E, M and H grades of instances.

The Fourier intensities (squared magnitudes) of the downsampled signal were multiplied by a Gaussian filter that diminishes the lowest-frequency unmeasured intensities by a factor of 25 relative to intensities at the center of the transform. These filtered intensities were then rescaled and the result was used as the mean for Poisson samples of simulated photon counts. A single photon count was generated at each frequency. The final step was to symmetrize the data by summing the counts at (p, q) and $(-p, -q)$.

The same intensity rescaling factor was used in all the benchmarks. This has the effect that individual atoms have the same characteristics (amplitude, width) across all the benchmarks. It also implies the total photon count in the data sets is proportional to N, the number of atoms. This normalization convention can be defended on information theoretic grounds: to reconstruct the types and positions (in a fixed field) of N atoms, the quantity of information should be proportional to N. Since each detected photon delivers the same quantity of information, this proportionality is maintained when the total photon count is also proportional to N.

Phase retrieval complexity is more directly linked to the sparsity of the signal-autocorrelation than the sparsity of the signal itself. When the autocorrelation is very sparse, it reveals unique inter-atomic vectors from which the signal can be reconstructed by direct search with little if any branching. Since the number of atom pairs grows as N^2, dividing this by the number of possible inter-atomic vectors estimates the average multiplicity μ of atom-atom vectors in the signal. We expect phase retrieval to be easy for any N as long as μ is small.

For the benchmark problems the number of possible inter-atomic vectors is a constant of order M^2, the number of grid points. This in turn was set by the number of measured Fourier samples in the data. Because real data and our synthetic data is well characterized by Gaussian decay of intensity with frequency, we chose to define the effective number of Fourier samples in any number of dimensions by the formula

$$V=\sum_{\mathbf{q}\in\Lambda^*}e^{-b|\mathbf{q}|^2},\tag{5}$$

where the sum is over the (infinite) lattice dual to the crystal lattice Λ and b is a parameter. In real (3D) crystals the intensity decay is reported as the Wilson B-factor [G], and $b=B/6$. For large V the sum can be approximated by an integral and we obtain in three dimensions

$$V=\left(\frac{6\pi}{B}\right)^{3/2}\mathrm{vol}(\Lambda),\tag{6}$$

where the last factor is the volume of the crystal unit cell. The mean multiplicity of interatomic vectors is now defined as

$$\mu=\frac{N^2}{V}.\tag{7}$$

Upon performing the sum (5) for the Gaussian filter of the benchmarks, we obtain the formula

$$\mu=\left(\frac{N}{64.17}\right)^2\tag{8}$$

for the benchmark instances. The numbers of atoms N of the instances was chosen to sample μ by roughly equal intervals, from $\mu=2.4$ to $\mu=39$.

The benchmark signals all have trivial space group ($P1$), limiting comparisons with real-data phase retrieval. To expand the comparison group we propose that in a space group with point group order Z, both N and V in (7) should be divided by Z. This has the effect of replacing μ by μ/Z. To the best of our knowledge and with this generalized definition, $\mu\approx 13$ is the hardest reported case of phase retrieval with real data from comparable structures (lacking heavy atoms; see Section V).

When the atom positions are uniformly and independently sampled, the Fourier amplitudes (before filtering) have a complex-normal distribution and their intensities are exponentially distributed (Wilson statistics [G]). This is a reasonable statistical model for the benchmarks, since the minimum distance constraints are weak for the densities of atoms considered. In real data there often are intensities that are unusually strong by this model, and their existence can be exploited by clever algorithms. Conversely, phase retrieval appears to be more challenging when the data lacks such outliers. The extreme case was recently studied for two-valued 1D signals [E1], where it is possible to construct signals whose intensities are all equal. Since the intensity-distribution characteristics are clearly important, we implemented the following modification in the construction of the atom positions.

We used the normalized second-moment of the intensities

$$i_2=\frac{\langle|\hat{\rho}|^4\rangle}{\langle|\hat{\rho}|^2\rangle^2}\tag{9}$$

to quantify the outlier content of the intensity distribution. The averages are over all the "measured" intensities, leaving out $\hat{\rho}(0,0)$. This statistic is increased when the high intensity tail is enhanced and decreases when the intensity distribution is made more uniform. Without any intervention, when atom positions are uniformly sampled (rejecting positions that violate the minimum distance constraint), we obtain $i_2\approx 4$. To make such instances easier, we select an atom at random and propose a new random position (still satisfying the distance constraint), accepting the proposal whenever the value of i_2 is increased. These increases are small, and many such moves had to be made to arrive at the value $i_2=4.5$ that define the E instances. The harder (H) instances were produced by the same procedure but where proposals are accepted whenever i_2 is decreased, continuing until $i_2=3.5$. Relatively few atom-position re-samplings were needed to arrive at the value $i_2=4.0$ that defines our medium difficulty (M) instances.

IV. Baseline results

To the best of our knowledge, the only known algorithms that reliably solve the benchmark problems are heuristic in nature. A common feature of these algorithms is that they act iteratively on the signal. To set a baseline for the benchmarks we have selected a simple exemplar: **Algorithm 1**. This section describes the algorithm, addresses some common misconceptions about this type of algorithm, and tries to establish some standards for reporting results.

Almost all crystallographic phase retrieval algorithms repeatedly use a "Fourier synthesis" step, where a signal is constructed from the known Fourier magnitudes and some set of phases. The simplest such operation is the Fourier magnitude projection, $\rho\to\rho_2=P_2(\rho)$, where ρ_2 inherits its Fourier phases from ρ and combines these with the Fourier magnitudes of the data, when available. When magnitude data is not available, say at frequency $\mathbf{q}$, the Fourier transform at $\mathbf{q}$ is simply copied.

Most algorithms also repeatedly do "direct space refinement", where prior information is imposed on the signal. One of the simplest operations of this kind is the support-size projection $\rho\to\rho_1=P_1(\rho)$, where ρ_1 is unchanged on the $|S|$ highest valued pixels and set to zero on the rest.

Algorithm 1 Simple Phase Retrieval

```
input $|\hat{\rho}|, |S|, \beta$                                  Fourier magnitudes, support size, RRR parameter
  $\rho \leftarrow \mathrm{rand}()$                                random initial signal
  $i \leftarrow 0$                                                 zero the iteration counter
  repeat
      $(\rho_1, S) \leftarrow P_1(\rho\,; |S|)$                    support-size projection
      $\rho_2 \leftarrow P_2(2\rho_1 - \rho\,; |\hat{\rho}|)$      Fourier magnitude projection
      $\rho \leftarrow \rho + \beta(\rho_2 - \rho_1)$              increment by the projection discrepancy
      $i \leftarrow i + 1$                                         increment counter
  until $\mathrm{pow}(\rho_2, S) > 0.95$                           termination criterion
output $\rho_2, i$                                                 phased input magnitudes (solution), iteration count
```

This is significantly weaker in constraining the signal than the analogous operation applied to a known support region. When the support region S is sufficiently compact so it avoids aliasing (in the crystal unit cell), the phase retrieval problem reverts to the easier aperiodic case.

Pseudocode for **Algorithm 1** is given above. In addition to the operations P_2 and P_1 already described, the algorithm calls on a function to initialize the signal and another function that computes the power ratio (4) for terminating iterations. Although the evolution of the signal is acutely sensitive to initial conditions (see below), the number of iterations required to find solutions, when averaged over initial conditions, is not. Our implementation used a pseudo-random number generator to avoid pathological cases, such as the all-zero signal. **Algorithm 1** has one parameter, β, that lies between 0 and 2. Our baseline results are for $\beta = 0.5$.

The logical origin of **Algorithm 1** is Fienup's "hybrid input-output" algorithm [F], the key elements of which are projections to constraint sets in Euclidean space. In crystallographic phase retrieval these are the operations P_2 and P_1. **Algorithm 1** is a slightly tidier version of Fienup's called "relaxed-reflect-reflect" (RRR) [BCL], [E2]. Written compactly it corresponds to the iterated map

$$\rho \mapsto \rho + \beta \left(P_2(2P_1(\rho) - \rho) - P_1(\rho) \right). \tag{10}$$

When reviewing the literature, some authors often lump this algorithm (and closely related ones) with another algorithm, called Fienup's "error reduction" algorithm [F], for which the map is

$$\rho \mapsto P_2(P_1(\rho)). \tag{11}$$

Aside from acting on the same space and being built from the same pair of projections, these two schemes have almost nothing in common. The error reduction algorithm, also called "alternating projections", is never used for hard (crystallographic) phase retrieval. In just a few iterations it converges on one of a multitude of uninteresting fixed points: signals with correct Fourier magnitudes that are proximal to, but not coincident with, a signal of the correct support size. By contrast, when the first scheme (RRR) has a fixed point it is because the correctly supported signal $\rho_1 = P_1(\rho)$ is in the range of P_2 — it also has the correct Fourier magnitudes. The presence of noise renders the fixed point inexact, but that is easily remedied by testing the power inequality in each iteration.

It is also not correct to characterize algorithms based on Fienup's hybrid input-output idea, in particular RRR, as "alternating" or "cyclic" in the usual sense [PR]. One can understand this by noticing that a fixed point ρ of RRR is in general not a solution. Instead it is the signals ρ_2 and ρ_1 generated by the projections and equal at a fixed point that are solutions (see **Algorithm 1** pseudocode). The truer sense in which this algorithm alternates is the "alternating direction method of multipliers" (ADMM) principle [E2].

Contemporary accounts often are dismissive of projection-based algorithms because convergence is not guaranteed and there have been reports of "stagnation" in the behavior of the iterates. While this criticism certainly applies to alternating projections, the direct opposite is empirically the case for RRR. The latter map has the characteristics of a strongly mixing dynamical system in mechanics, where ergodic behavior is the rule rather then the exception. It is for this reason that initialization is unimportant. As in mechanics the evidence of ergodicity is very strong even while prospects of a proof are dim. Every attempt by RRR (**Algorithm 1**) to solve a benchmark problem produced a solution: the success rate was 100%.

There is no better illustration of the statistical inevitability of solution discovery than the time series of the power ratio (4). A typical time series for instance **`data100H`** is shown in Figure 4. The solution — marked by the sudden jump — is not constructed incrementally but appears as an isolated event, when the chaotic dynamics arrives by chance at a fixed point's basin of attraction. RRR and related algorithms are also used in convex optimization, where the behavior is very different and in fact convergent. However, when these algorithms are applied to hard phase retrieval only the very short capture-phase of the solution process would appear to fall under the purview of convex analysis.

The power ratio time series also illustrates the limits of using just the support size to constrain the signal. Figure 5 shows how the plot in Fig. 4 changes as the support size increases. The hardest benchmark instances, with $N = 400$ atoms, are just short of the point where criterion (4) fails as a valid certificate. It is only for this reason that the benchmarks do not go beyond $N = 400$ ($\mu = 39$). Algorithms that seek signals with additional prior characteristics — *e.g.* peaks — could in principle succeed beyond this limit on the number of atoms. Indeed, algorithm developers are encouraged to exploit any of the prior signal information given in Section

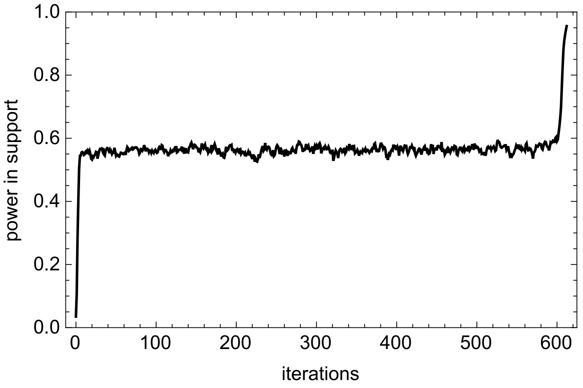


Fig. 4. Time series of the power ratio (4) over the course of solving **`data100H`** with **Algorithm 1**. Not only is the final capture by the solution fixed point very brief, so is the transient from the random initial signal to the family of signals explored in the search. Incorrectly phased signals in the long epoch of search have about 55% power in a dynamic support constrained only by size.

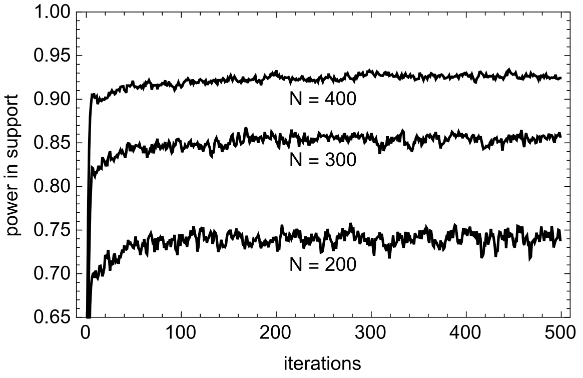


Fig. 5. Power ratio time series generated by **Algorithm 1** for instances with progressively more atoms showing just the steady state behavior before the solution-discovery jump seen in Figure 4. Beyond 400 atoms the power-in-support fraction fails as a solution certificate.

III. Criterion (4) should only be seen as a convenient solution certificate that holds for $N \leq 400$.

Benchmark results are most useful when behavior can be assessed with respect to hardness parameters. Though actual runtimes are important, trends in behavior are best reported in terms that do not depend on implementation details. In the case of **Algorithm 1**, the average number of iterations in repeated trials serves this purpose[2]. Success rates, when greater than zero, are more useful when converted to an expected runtime per solution. Had **Algorithm 1** been run with a bound on the number of iterations, an expected runtime could have been computed from the total number of solutions found and the total number of iterations performed.

Iteration counts per solution for **Algorithm 1**, averaged over 20 trials per instance, are given in Table I. Figure 6 shows the behavior with respect to the multiplicity-of-inter-atomic-vectors parameter μ defined by (8). At each N, almost without exception, the E instance is easier than the M instance, which in turn is significantly easier than the H instance. The behavior with μ is consistent with a simple exponential. Linear fits to the logarithms of the iteration counts give the following factors by which the mean count grows when μ is increased by 1: 1.56 (E), 1.72 (M), 1.90 (H).

TABLE I
ALGORITHM 1 MEAN ITERATION COUNTS (LOG_{10})

N	E	M	H
100	1.87	2.15	3.01
140	2.37	3.00	3.93
175	3.23	3.55	5.20
200	3.42	4.57	5.48
225	3.47	5.12	6.92
245	4.33	5.77	7.03
265	5.81	6.02	7.60
285	6.06	5.98	7.62
300	5.55	6.97	–
315	6.46	7.29	–
330	6.58	8.41	–
345	7.83	–	–
360	6.86	–	–
375	8.00	–	–

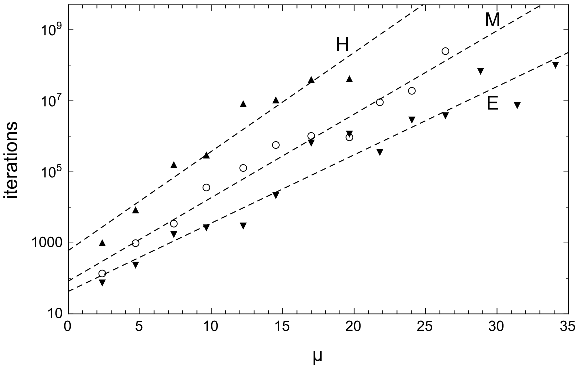


Fig. 6. Exponential growth of the mean iteration count for **Algorithm 1** as a function of μ for the three difficulty grades of instances.

We expect our baseline results to be an easy target. The best outcome of phase retrieval algorithm development would of course be a fundamentally different algorithm, promising subexponential cost in the parameter μ. In the absence of this and precisely because of the exponential growth, even incremental improvements will bring substantial dividends. Although we have not performed the experiment, the extrapolation of our results indicate that **Algorithm 1** would require roughly three cpu-years[3] to solve **`data330H`**, compared with the single hour needed for **`data330E`**.

V. STATE OF THE ART

Probably the last time a phase retrieval milestone — on real data — was hailed was the Shake-and-Bake (SnB) solution of triclinic lysozyme in 1998 [DM1]. The SnB algorithm was

[2] The runtime per iteration in our implementation, about 1 msec, is essentially constant across all instances.

[3] By the strong mixing hypothesis **Algorithm 1** parallelizes trivially by independent runs differing only in the initial signal. A 1000-cpu cluster should find the solution in about one day.

the product of a long history of developments that drew inspiration from various disciplines, including signal processing, probability theory and iterative methods for solving non-linear equations. In this section we review the principles behind SnB [SnB] as well as those used in three leading crystallographic packages: SHELXD [SX], SIR2004 [SIR], SUPERFLIP [SF].

Because the earliest phase retrieval algorithms were developed in the pre-FFT era, they imposed prior information on the signal not directly in real-space, but indirectly during Fourier synthesis. The simplest such strategy, known as David Sayre's "tangent formula" [TF], is based on the observation that in a signal ρ comprised of equal atom-like distributions (*e.g.* Gaussians), the Fourier phases of ρ^2 and ρ are the same. This opens up the possibility that iterating $\rho \mapsto P_2(\rho^2)$ might by itself produce as fixed points a signal that (*i*) has been synthesized from the known Fourier magnitudes and (*ii*) corresponds to an atomic distribution — at least for crystals of sufficiently identical atoms. It was possible to efficiently implement this map working just with the Fourier coefficients by expressing the transform of the square as the convolution of the transforms and approximating the convolution by terms where both Fourier factors have a large magnitude. SHELXD and SIR2004 have the option to alternate the tangent formula iteration with a direct space refinement operation.

The tangent formula modification of the Fourier synthesis operation (P_2) is just one way the alternating scheme (11) is made viable again, in effect eliminating a host of uninteresting fixed points. The identical-atom model of the signal on which the method is based is also the premise behind another modification of P_2. This is the SnB objective function on the phases that is first minimized before phases are combined with magnitudes [SnB]. The simplest form of the objective function is based on the observation that the distribution of the product $\hat{\rho}(\mathbf{q}_1)\hat{\rho}(\mathbf{q}_2)\hat{\rho}(\mathbf{q}_3)$ is invariant, for $\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 = 0$, under translation of all the atoms and therefore exhibits a nontrivial dependence on the sum of the corresponding phases, $\phi_1 + \phi_2 + \phi_3$. The distribution of this "triplet" phase depends on the data via the known magnitude of the cubic product. The resulting conditional distribution for the triplet phase can be calculated explicitly for the case of equal atoms uniformly distributed in the crystal unit cell and serves as a model for triplet distributions in a typical crystal [G]. SnB tries to bring the cosines of the triplet phases in line with their expectation in the random model by minimizing a sum-of-squares objective function.

It is important that modified Fourier synthesis, by either the tangent formula or the SnB objective function, is combined with a robust direct space refinement operation like the support size projection P_1. This is because the phase interventions, or modifications of P_2, are based on approximate models and should only serve to bias the search for phases. When the bias built into P_2 has sufficient strength, one improves the probability that the signal $P_2(P_1(\rho))$ is also fixed by P_1 and is therefore a solution. After all, the bias in both methods (tangent formula, SnB objective function) is derived from a direct space model. Phase retrieval often succeeds simply by alternating a modified P_2 and a direct space refinement P_1 of some type. In SnB, SHELXD and SIR2004 the P_1 operation can impose prior information on the signal beyond just the size of its support. This may include knowledge of the minimum atom-atom distance, the presence of a known number of heavy atoms, or even the expected histogram of the signal values.

The phase retrieval algorithm in SUPERFLIP also alternates between Fourier synthesis and direct space refinement, but unlike the other packages it avoids false fixed points through a modification of the direct space operation [DM10]. When written in terms of the **Algorithm 1** projections, SUPERFLIP iterates a close approximation of the map

$$\rho \mapsto P_2(2P_1(\rho) - \rho). \tag{12}$$

The algorithm's name is derived from the argument of P_2, wherein the sign of the signal is reversed wherever it is judged not to be in the support (unchanged otherwise). In a solution the "charge flipping" step has no effect and the signal is a fixed point of the map. On the other hand, one cannot theoretically rule out the possibility of exotic non-solution fixed points, where charge flipping only changes the Fourier magnitudes — that are then restored and the charge re-flipped by P_2. The P_1 used by SUPERFLIP [DM10] differs from the one in **Algorithm 1** by being parameterized through a small positive lower bound on the signal value in the support rather than a bound on the support size. Although **Algorithm 1** is not based on an alternation of two operations, it is intriguing and perhaps not a complete coincidence that Fourier synthesis in this algorithm is also preceded by charge flipping.

Direct comparison of the accomplishments of the crystallographic packages with the benchmark problems is complicated by a number of factors. First and foremost is the fact that data sets of sufficient quality, on which sparsity can be imposed, become increasingly rare as the number of atoms in the unit cell grows. In crystals of large protein molecules there is too much variability in the electron density from one unit cell to another (solvent disorder) that individual protein atoms cannot be resolved. The signal encoded by the Fourier magnitudes in this case is that of the average of the atomic distributions, and its support is so large that it has become too weak to be used as a constraint. Almost all large protein structures are solved with the help of additional data derived from atom-specific inelastic scattering. A large share of the credit for structures with N above 1000 that did not rely on such additional data and yielded to "direct methods" goes to the crystal growers who managed to significantly reduce disorder in their crystals.

Another factor that complicates comparisons is the presence of heavy atoms. Large proteins often contain a minority admixture of heavy atoms, and it is well known that this makes phase retrieval easier, even when this information is not used. The benchmark problems have equal numbers of atoms of two types and therefore correspond to harder instances in this respect.

Table II lists what we judge to be the hardest instances of successful real-data phase retrieval for crystals with (*i*) resolved atoms and (*ii*) the fewest number of heavy atoms. The highest value of μ in this list is near the middle of the benchmark problems.

TABLE II
LARGE NEARLY EQUAL-ATOM STRUCTURES SOLVED BY DIRECT METHODS

structure	PDB or CSD	group, Z	res. (Å)	N/Z, heavy	μ/Z	iter. or time	software	ref.
HEW lysozyme	–	$P1$, 1	0.85	1001, 10S	–	3,400	SnB	[DM1]
alpha-1 peptide	1BYZ	$P1$, 1	0.90	408, 1Cl	1.73	4,500	SnB	[DM2]
acutohaemolysin	1MC2	$C2$, 4	0.85	975, 18S	4.45	–	SnB	[DM3]
scorpion toxin II	1AHO	$P2_12_12_1$, 4	0.96	500, 10S	4.46	413,000	SnB	[DM4]
actinomycin D	1A7Y	$P1$, 1	0.94	270	1.40	–	SHELXD	[DM5]
feglymycin	1W7Q	$P6_5$, 6	1.10	828	6.39	–	SHELXD	[DM6]
HEW lysozyme	4LZT	$P1$, 1	0.95	1001, 10S	11.06	–	SHELXD	[DM7]
human cyclophilin G	2WFI	$P2_12_12_1$, 4	0.75	1486, 2Mg 15S	11.17	60 min	SHELXD	[DM8]
hirustasin	1BX7	$P4_32_12$, 8	1.20	366, 12S	4.29	546 min	SIR2004	[DM9]
pheromone ER-1	2ERL	$C2$, 4	1.00	303, 8S	4.72	19 min	SIR2004	[DM9]
Kunitz domain C5	2KNT	$P2_1$, 2	1.20	460, 1P 6S	6.85	22 min	SIR2004	[DM9]
bovine ribonuclease	1DY5	$P2_1$, 2	0.87	1894, 31S	12.53	131 min	SIR2004	[DM9]
$2C_{72}N_4O_6$	PAWVEO	$P1$, 1	0.80	164	0.58	100	SUPERFLIP	[DM10]
$2C_{77.5}N_4O_{12.5}$	GOFMOD	$P1$, 1	0.80	188	0.66	250	SUPERFLIP	[DM10]
apamin	–	$P2_1$, 2	0.95	385	4.36	20,000	SUPERFLIP	[DM11]

VI. SUMMARY

Phase retrieval can be decisive in the success of crystal structure discovery. Algorithms that apply to periodic signals are heuristic and their success and runtime behavior is poorly documented. It is not normal practice for crystallographers to test algorithms with synthetic data; failures usually are attributed to real data that has been compromised. And when data quality is good, Nature does not always cooperate to create instances with graded hardness for the study of algorithm behavior.

Phase retrieval theory moved into a new era of systematic study when it was taken up by applied mathematicians about ten years ago. However, the algorithms generated by this development have had no impact on phase retrieval for crystals. Periodicity of the signal in crystallography is not a minor property or treatable as a limiting case. Algorithms that retrieve phases efficiently for aperiodic signals do not automatically generalize to periodic signals.

The circumstances just described are both addressed by our benchmark problems. Crystallographers should test their algorithms with synthetic data and applied mathematicians need access to simply formulated, realistic problems in order to develop relevant algorithms. Careful design and compromise went into the construction of the benchmark problems. The simple format of the data sets make them interpretable to expert and non-expert alike. The solution criterion is simple and unambiguous. The signals are realistic atomic distributions, with harder instances differing only in the number of atoms. The easiest instances yield to even naive algorithms, while the hardest defeat the state-of-the-art. And the subdivision of the hardness scale (autocorrelation sparsity parameter μ) is fine enough that algorithms can be compared even when runtimes grow exponentially.

REFERENCES

[PR] Y. Shechtman, et al., Phase retrieval with application to optical imaging: a contemporary overview, IEEE Signal Processing Magazine **32**, 87-109 (2015).

[CSD] C. R. Groom and F. H. Allen, The Cambridge Structural Database in retrospect and prospect, Angew. Chem. Int. Ed. **53**, 662-671 (2014).

[SPI] A. Aquila, et al., The linac coherent light source single particle imaging road map, Structural Dynamics **2**, 041701 (2015).

[G] C. Giacovazzo, Direct Phasing in Crystallography (Oxford University Press: Oxford, UK, 1998).

[E1] V. Elser, The complexity of bit retrieval, submitted (2016).

[F] J. R. Fienup, Phase retrieval algorithms: a comparison, Applied Optics **21**, 2758-2769 (1982).

[BCL] H. H. Bauschke, P. L. Combettes and D. R. Luke, Finding best approximation pairs relative to two closed convex sets in Hilbert spaces, J. Approx. Theory **79**, 418-443 (1994).

[E2] V. Elser, Matrix product constraints by projection methods, Journal of Global Optimization **68**, 329-355 (2017).

[SnB] R. Miller, et al., On the application of the minimal principle to solve unknown structures, Science **259**, 1430-1433 (1993).

[SX] G. M. Sheldrick, A short history of SHELX, Acta Crystallographica Section A: Foundations of Crystallography **64**, 112-122 (2008).

[SIR] M. C. Burla, et al., SIR2004: an improved tool for crystal structure determination and refinement, Journal of Applied Crystallography **38**, 381-388 (2005).

[SF] L. Palatinus and G. Chapuis, Superflip: a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions, Journal of Applied Crystallography **40**, 786-790 (2007).

[TF] D. Sayre, The squaring method: a new method for phase determination, Acta Crystallographica **5**, 60-65 (1952).

[DM1] A. M. Deacon, et al., The Shake-and-Bake structure determination of triclinic lysozyme, Proceedings of the National Academy of Sciences **95**, 9284-9289 (1998).

[DM2] G. G. Privé, et al., Packed protein bilayers in the 0.90 Å resolution structure of a designed alpha helical bundle, Protein Science **8**, 1400-1409 (1999).

[DM3] Q. Liu, et al., The crystal structure of a novel, inactive, lysine 49 PLA2 from Agkistrodon acutus venom: an ultrahigh resolution, ab initio structure determination, Journal of Biological Chemistry **278**, 41400-41408 (2003).

[DM4] G. D. Smith, et al., Ab initio structure determination and refinement of a scorpion protein toxin, Acta Crystallographica Section D: Biological Crystallography **53**, 551-557 (1997).

[DM5] Schäfer, Martina, et al., Crystal structures of actinomycin D and actinomycin Z3, Angew. Chem. Int. Ed. **37**, 2381-2384 (1998).

[DM6] G. Bunkóczi, L. Vértesy, and G. M. Sheldrick, The antiviral antibiotic feglymycin: first direct methods solution of a 1000+ equal-atom structure, Angew. Chem. Int. Ed. **44**, 1340-1342 (2005).

[DM7] G. M. Sheldrick, et al., International Tables for Crystallography, Vol. F, Ch. 16.1, 413-432 (2012).

[DM8] C. M. Stegmann, et al., The thermodynamic influence of trapped water molecules on a protein-ligand interaction, Angew. Chem. Int. Ed. **48**, 5207-5210 (2009).

[DM9] M. C. Burla, et al., The revenge of the Patterson methods. I. Protein ab initio phasing, Journal of Applied Crystallography **39**, 527-535 (2006).

[DM10] G. Oszlányi, Gábor, and Andras Sütő, Ab initio structure solution by charge flipping, Acta Crystallographica Section A: Foundations of Crystallography **60**, 134-141 (2004).

[DM11] C. Dumas and A. van der Lee, Macromolecular structure solution by charge flipping, Acta Crystallographica Section D: Biological Crystallography **64**, 864-873 (2008).