# Unsupervised determination of protein crystal structures

**Ivan S. Ufimtsev**[a,1] **and Michael Levitt**[a,1]

[a]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305

Contributed by Michael Levitt, March 25, 2019 (sent for review January 10, 2019; reviewed by Pavel V. Afonine and Alexandre Urzhumtsev)

**We present a method for automatic solution of protein crystal structures. The method proceeds with a single initial model obtained, for instance, by molecular replacement (MR). If a good-quality search model is not available, as often is the case with MR of distant homologs, our method first can automatically screen a large pool of poorly placed models and single out promising candidates for further processing if there are any. We demonstrate its utility by solving a set of synthetic cases in the 2.9- to 3.45-Å resolution.**

massively parallel | unsupervised method | protein crystal structure | Bcl-xL

At present, most protein crystal structures are determined in the absence of experimental phase information [the phase problem (1–3)], principally by molecular replacement (MR) methods (4–8). This is especially relevant for the recently introduced serial femtosecond crystallography (SFX) method made possible by the availability of free electron laser (XFEL) light sources (9–11), wherein de novo phasing is difficult due to inherent inaccuracies of the scattering data (12–14). Success of MR methods strongly depends on the quality of the search model, which often is expressed as Cα-rmsd distance from the solution. This distance correlates with the sequence identity between the protein and the homolog used to build the model (15). Typically 25–30% or better sequence identity is needed to build a promising search model (7, 16). At lower sequence identity the quality of the model deteriorates quickly, and MR typically is unable to reliably place the model, instead producing many (tens, hundreds) similarly poor solutions as evaluated by log-likelihood and translation z scores (7, 17). Choosing the initial model for building and refinement in this case is a nontrivial problem and subject to a degree of chance. If one model fails, one has to start with a different model, and the process can take many months, and sometimes the structure is not solved.

Modern packages for automatic model building and refinement (18–25) use smart algorithms for density map interpretation, yet still rely to a large extent on human input, especially at the beginning of the solution process when phases of sufficient quality are not available. Here we build on these previous approaches and present a method for solving protein crystal structures from low-quality initial models that generally converges to the solution with little or no human supervision. Contrary to the modern paradigm in crystallography software development, we do not try to develop an algorithm that can interpret electron densities on par with human crystallographers. Instead, we use a statistical approach wherein thousands of automatically built models of reasonable (but far from the best) quality are combined together and lead to the solution. This is possible since fast and reasonably accurate algorithms for automatic model building are readily available (19). As demonstrated below, our method has a large "radius of convergence" expressed as the Cα-rmsd of the initial model from the converged solution. Finally, our method can readily take advantage of computational clusters to quickly screen pools of initial models and find good candidates for further processing.

Here we test the method on several small proteins serving as synthetic examples (Fig. 1) showing that we can generally solve structures when the initial model is within 3 Å Cα-rmsd. In the best-case scenario the method will produce a high-quality solution (better than what a human could attain) or an improved model that can be corrected manually and then run through another cycle of unsupervised solution process. In Ufimtsev et al. (26), the method is used to solve the crystal structure of the human lethal giant larvae (LGL2) protein that resisted years of human efforts due to very low 10% sequence identity to the closest solved homolog (27).

## Results

**Description of the Method.** We begin with an initial model obtained by MR or some other method, the sequence of the molecule we are building, and the experimental structure factor amplitudes with standard deviations (SDs) ($F_{obs,}$ $\sigma_{obs}$). The model can be a polyalanine chain (recommended at early macrocycles) or can have partial or full sequence and consist of one or many chains. This model is referred to as parent model $M_0$. At this point we enter the macrocycle loop (Fig. 2).

At the end of each macrocycle we anticipate the model to be deformed toward the solution (if one chooses to refine $M_m$ with $F_{obs}$ and $\varphi_{ave}$ restraints to produce $M_{m+1}$, i.e., the "refinement" mode) or fully rebuilt (if $M_{m+1}$ is built in $\rho_{obs} = \{F_{obs}, \varphi_{ave}\}$ density, i.e., the "full" mode). The refinement mode is good in the beginning of the procedure when maps are poorly interpretable and the chances of breaking the parent model in the auto-building step are high. The full mode is good at late stages, when the parent model has phases of good quality able to produce interpretable maps to guarantee rebuilding does not break the model. One also can combine refinement and full modes by

### Significance

**Solving crystal structures of large biological macromolecules in the absence of experimental phase information, especially at low resolution, is a tedious problem prone to mistakes and overfitting. Due to errors stemming from poorly interpretable parts of the electron density map, human experience and intuition are imperative for building a correct atomistic model. For this reason, tools for automatic structure determination used nowadays require constant human supervision. Here we present a method that can overcome the difficulties; it greatly reduces or even eliminates human involvement in the solution process by working with ensembles of possible solutions. This approach can find solution of a higher quality than can humans and can solve difficult cases not amenable to other methods.**

Author contributions: I.S.U. and M.L. designed research; I.S.U. and M.L. performed research; I.S.U. contributed new reagents/analytic tools; I.S.U. and M.L. analyzed data; and I.S.U. wrote the paper.

Reviewers: P.V.A., Lawrence Berkeley National Laboratory; and A.U., Institut de Génétique et de Biologie Moléculaire et Cellulaire.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: ufimtsev@stanford.edu or michael.levitt@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1821512116/-/DCSupplemental.

Published online May 14, 2019.

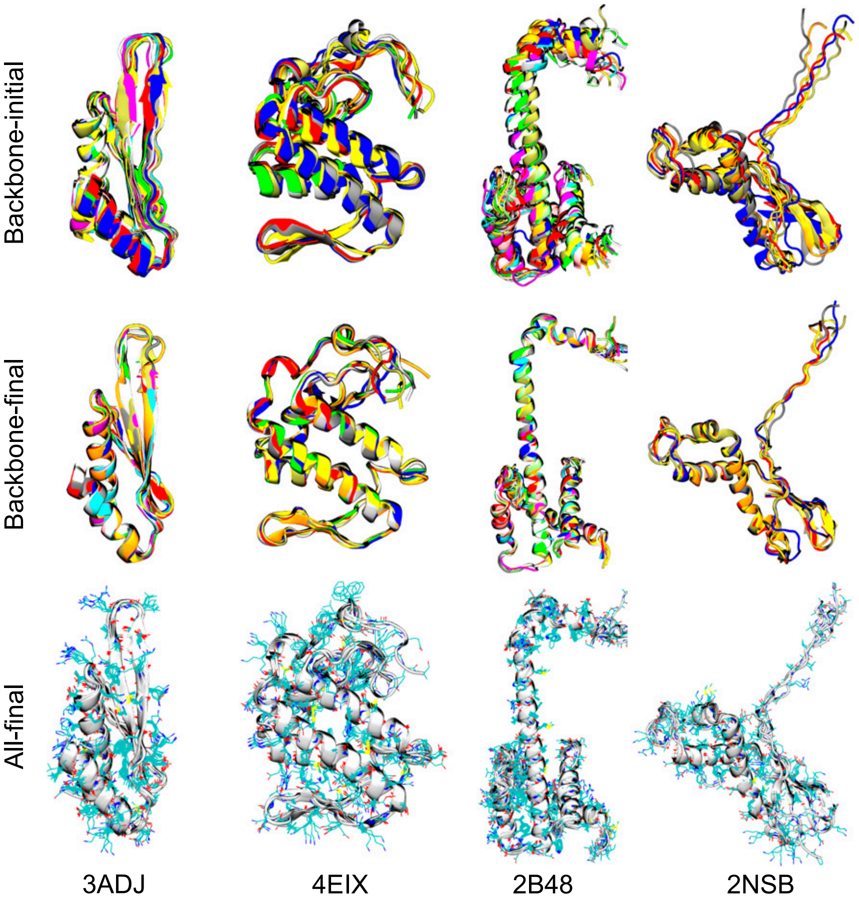

**Fig. 1.** Protein models in the 2.9- to 3.45-Å resolution range used in the synthetic tests. (*Top* row) Multiple initial decoys were generated by random deformation of the deposited structures along lowest-frequency normal modes (33). All of the side chains were removed and renamed UNK. (*Middle* row) Solved structures. (*Bottom* row) Same structures as in the *Middle* row displayed with side chains. Only solved models and the corresponding initial decoys are displayed for clarity.

refining and rebuilding the parent model at every macrocycle and passing the rebuilt model to the next macrocycle if the model is of higher quality than the refined model. When run in refinement mode, one macrocycle is akin to one iteration of a standard refinement program like Refmac or phenix.refine, yet it is more robust with respect to the choice of the optimization direction and the ability to escape local minima. It is also ∼100 times more computationally expensive and typically is executed in parallel due to the embarrassing parallelism of the macrocycle loop (Fig. 2).

At every macroiteration $m$, the algorithm proceeds through a pipeline composed of standard (density modification → auto build → refinement) steps, which is repeated 50 times (the inner microcycle in Fig. 2). The density modification tool is our in-house developed code. The density modification code generates an electron density map based on the parent model $M_m$, sequence (to estimate the solvent content), and [$F_{obs}$, $\sigma_{obs}$] subject to a set of standard restraints: (*i*) density histogram and bulk solvent restraints (solvent flattening), (*ii*) [$F_{obs}$, $\sigma_{obs}$], and (*iii*) $M_m$'s low-resolution phase restraints up to some resolution threshold $s_{max}$ that is adjusted dynamically at the end of each macrocycle. All these restraints are enforced through a set of real ↔ real and real ↔ reciprocal space projections: (*i*) density histogram projection, (*ii*) $2mF_{obs}$-$DF_c$ projection computed by the program Sigmaa (28), and (*iii*) phase projection computed by Fourier transform of the density map followed by the inverse Fourier transform of the computed amplitudes and target phases. This runs for a fixed number of 30 iterations. The procedure starts with a map combining random amplitudes $r\ exp(-B_0\ s^2)$ with $M_m$'s phases, where $r$ is a random number uniformly distributed in the [0,1] range, $B_0$ is the overall $B$ factor, and $s$ is the length of the reciprocal space vector.

Repeating the procedure 50 times produces 50 different density maps, which are quite diverse at early macroiterations and overlap strongly at late iterations. For each of the maps $\rho_C$ we compute its correlation with the $M_m$ density map $\rho_M$ and average all of the 50 correlation coefficients. If the average is greater than 0.6, $s_{max}$ is decreased by 10% (fewer phases constrained to $\varphi_M$ in the next macrocycle, i.e., more relaxed phase constraints) and is increased by 10% otherwise. Parameter $s_{max}$ is initialized at the beginning of the solution process to the resolution of the $N_{SF}$/2th structure factor after sorting all of the $N_{SF}$ structure factors by resolution. Next, for each map $\rho_C$ we build a trial model by Buccaneer 1.6.1 (19) and refine it by Refmac 5.8.0135 (29) with default settings against $F_{obs}$ subject to secondary structure restraints generated by Prosmart (30). The restraints do not include so-called h-bond terms to avoid any unnecessary bias. Likewise, disulfide bond restraints are not applied. The 50 newly built trial structures are ranked by their R-free values (31), and the 20 best structures are used to derive phase restraints used to refine (in refine mode) or rebuild (in full mode) $M_m$ to produce the next parent model $M_{m+1}$. We combine the 20 models together by averaging their figure-of-merit–weighted (FOM-weighted) density maps as

$$F^k{}_{ave}exp\left(i\varphi^k{}_{ave}\right)=\left\langle m_n{}^kF_n{}^k{}_c exp\left(i\varphi_n{}^k{}_c\right)\right\rangle_{n=1..20}, \qquad [1]$$

where $F_n{}^k{}_c exp(i\varphi_n{}^k{}_c)$ is the computed $k$th structure factor of trial structure $n$ and $m_n{}^k$ is its figure of merit as computed by Sigmaa.

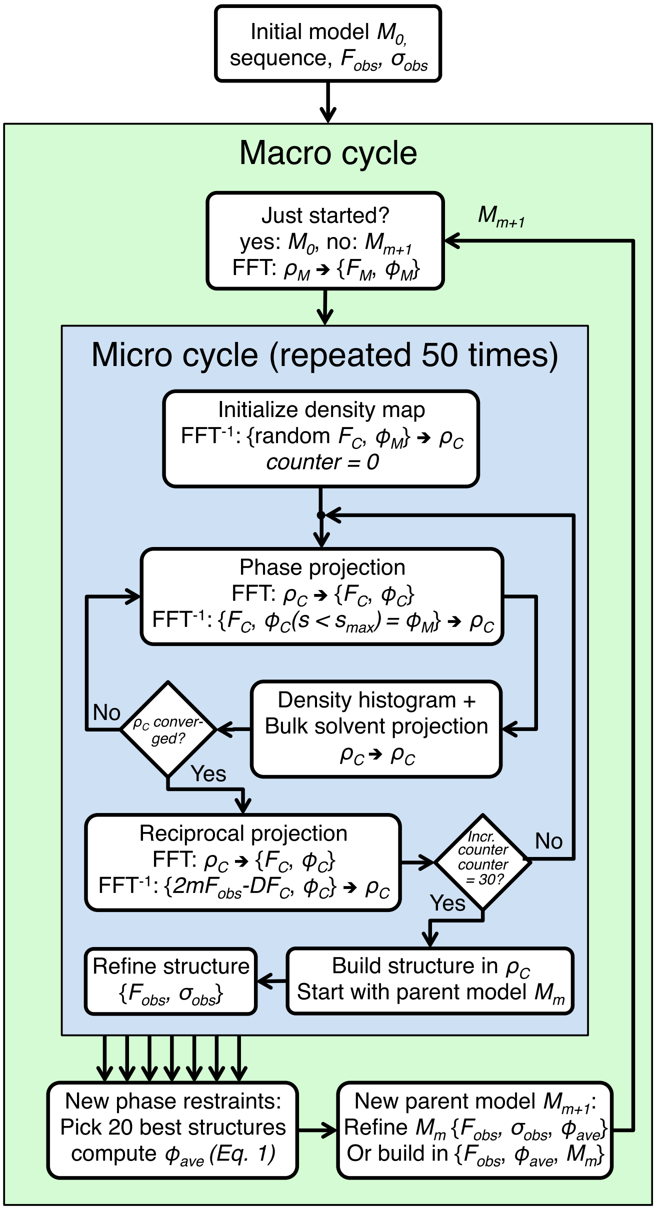
Initial model $M_0$, sequence, $F_{obs}$, $\sigma_{obs}$
Macro cycle
Just started?
yes: $M_0$, no: $M_{m+1}$
FFT: $\rho_M \rightarrow \{F_M, \phi_M\}$
$M_{m+1}$
Micro cycle (repeated 50 times)
Initialize density map
FFT$^{-1}$: {random $F_C$, $\phi_M$} $\rightarrow \rho_C$
counter = 0
Phase projection
FFT: $\rho_C \rightarrow \{F_C, \phi_C\}$
FFT$^{-1}$: $\{F_C, \phi_C(s < s_{max}) = \phi_M\} \rightarrow \rho_C$
No
$\rho_C$ converged?
Yes
Density histogram + Bulk solvent projection
$\rho_C \rightarrow \rho_C$
Reciprocal projection
FFT: $\rho_C \rightarrow \{F_C, \phi_C\}$
FFT$^{-1}$: $\{2mF_{obs}\text{-}DF_C, \phi_C\} \rightarrow \rho_C$
Incr. counter counter = 30?
No
Yes
Refine structure $\{F_{obs}, \sigma_{obs}\}$
Build structure in $\rho_C$
Start with parent model $M_m$
New phase restraints:
Pick 20 best structures
compute $\phi_{ave}$ (Eq. 1)
New parent model $M_{m+1}$:
Refine $M_m \{F_{obs}, \sigma_{obs}, \phi_{ave}\}$
Or build in $\{F_{obs}, \phi_{ave}, M_m\}$

**Fig. 2.** Detailed flowchart of our unsupervised determination of protein crystal structures. The parent model $M_m$ is updated at each macrocycle based on information derived from the 20 best auto-built trial models (of 50). The trial models are built by Buccaneer (19) and refined by Refmac (29), and their quality is assessed by their free R factors (31). The models are built from density maps generated by our in-house–developed density modification code. The density modification solver is seeded randomly, thereby generating every time a different map. Here, $M_m$ is the structure of the parent model at macrocycle $m$, $F_{obs}$, $\sigma_{obs}$ are experimental structure factor amplitudes and SDs, $\rho_M$ is the density map of model $M$, $\{F_M, \varphi_M\}$ are computed structure factors of model $M$, $\rho_C$ is the density map used to build the trial models, $\{F_C, \varphi_C\}$ are computed structure factors of $\rho_C$, and $\varphi_{ave}$ are the averaged phases. FFT is a fast Fourier transform of the density map and FFT$^{-1}$ is the inverse transform. The resolution of the data is measured by 1/$s$ with $s_{max}$ determining the degree of phase projection in each microcycle.

We also experimented with an alternative averaging scheme with all ${F_n}^k{}_c$ set equal to 1.0 and found that it performed similarly well. A more accurate approach to combine the phases and filter out outliers would be based on cluster analysis (32); however, it is not employed in the current version of the code. In refine mode, $M_{m+1}$ is generated by refining $M_m$ against the complex value structure factors $\{F_{obs},\varphi_{ave}\}$. In full mode, $M_{m+1}$ is built in the $\{F_{obs},\varphi_{ave}\}$ density map starting with $M_m$'s alpha carbons. Finally, this step completes one macrocycle, and $M_{m+1}$ is passed to the next macrocycle if needed.

**Validation with Synthetic Data.** To test our method we selected four small proteins solved at different fairly low resolution: 3ADJ at 3.0 Å, 4EIX at 2.9 Å, 2NSB at 3.2 Å, and 2B48 at 3.45 Å. For each protein we generated several hundred near-native decoys by computing the protein's normal modes and randomly exciting the 10 lowest-frequency normal modes (33). After removing structures with clashes, for each protein we obtained a set of several hundred decoys with Cα-rmsd from the corresponding deposited structure in the 0.1- to 5.0-Å range. Next, we removed all side chains, ligands, and waters and renamed all residues as UNK.

All of the decoys are represented in Fig. 3 as gray shaded circles. The $x$ and $y$ coordinates were obtained from multidimensional scaling (MDS) analysis (34) of the decoy all-to-all pairwise rmsd matrix. Here, MDS seeks for $n$ points on a 2D plain, with pairwise Euclidian distances approximating the $n$-by-$n$ rmsd matrix in the least-squares way, and provides a 2D representation of the rmsd data. Coordinates $\{x_i,y_i\}$ of all decoys were shifted by the same amount to place the deposited structure $\{x_0,y_0\}$ at the origin of the plot, and then each $\{x_i,y_i\}$ pair was scaled so that its Euclidean distance from the origin, $(x_i^2 + y_i^2)^{1/2}$, would be exactly equal to the decoy's rmsd from the deposited structure. The deposited structure is represented by the magenta circle in Fig. 3, and the concentric circles, therefore, define regions of constant rmsd.

Next, for each protein we handpicked 20–30 decoys uniformly distributed in the 2.3- to 3.5-Å rmsd zone (green circles in Fig. 3). This level of deviation is much larger than the theoretical convergence radius (35). We then processed each decoy for 120 macrocycles in refine mode (i.e., the parent model was only refined and not rebuilt at each macrocycle). If a refined model was inside the 2-Å rmsd zone outlined by the thick circle in Fig. 3, we define it as a converged model and the corresponding initial decoy is shown as a large green circle; otherwise it is shown as a small green circle. The refined models are shown as large red circles for those that converged and small red circles for those that did not converge.

As anticipated, the converged models (large green circles in Fig. 3 for 43 decoys in total) tend to localize closer to the origin than unconverged models (small green circles for 47 decoys), indicating a strong dependency of the quality of our solution on the degree of deformation of the initial decoy. The statistics are summarized in Fig. 4, where the relative fractions of converged and unconverged decoys in various rmsd zones are shown by green and red bars, correspondingly. The relative fraction of solved decoys (green bars in Fig. 4) decreases almost linearly with the magnitude of the initial deformation: At 2.3-Å rmsd all structures are solved, while at 3.5-Å rmsd only 10% of structures are solved. Approximately one-half of all decoys that start at 2.9-Å rmsd deformation are solved, allowing this value to be considered as the radius of convergence of our method.

One can see in panel 2B48 in Fig. 3 that the converged solutions (red circles inside the 2-Å zone) tend to cluster in the 1.5-Å rmsd zone. To find the origin of such clustering, we rerefined the structure, starting with the best trial structure in terms of R-free and chain integrity that was generated in one of the later macrocycles. Compared with the original deposited structure, our solution has a larger number of protein atoms (1,230 vs. 1,145) and lower R factors (work/free) (0.243/0.250 vs. 0.261/0.305). The major structural difference is between residues R100 and T118 as shown in Fig. 5.

Because a similar systematic shift is observed in panel 4EIX in Fig. 3, we rerefined the structure starting from one of the best trial structures. However, unlike the 2B48 case we did not

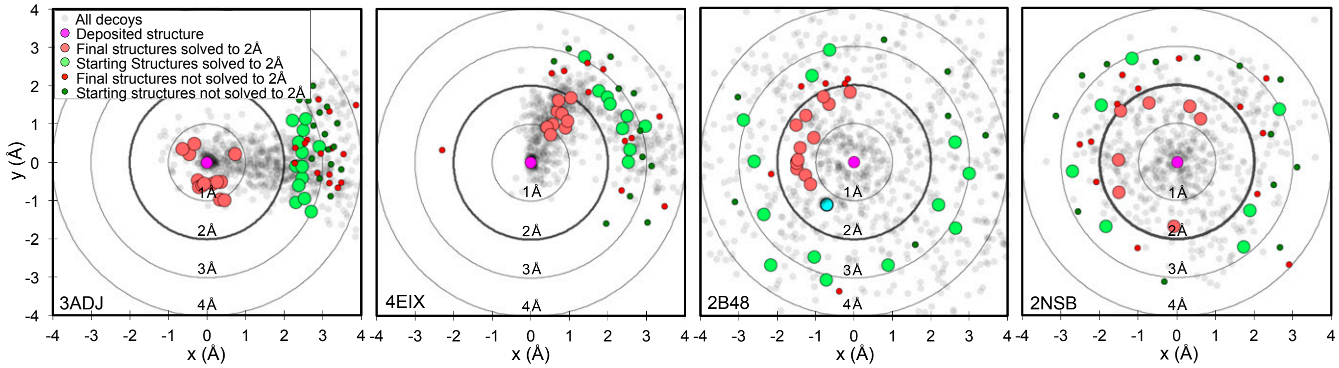

**Fig. 3.** Decoy statistics for the four proteins. For every structure we generated several hundred decoys at different levels of deformation (gray shaded circles). The green decoys were selected for refinement and produced converged (solved to within 2-Å rmsd from the deposited structure, large red circles inside the 2-Å circle) and unconverged (small red circles outside the circle) structures, correspondingly. The plots were built by multidimensional scaling representation of the decoys' pairwise rmsd matrix. The cyan circle for 2B48 shows the rerefined structure.

find any relevant structural differences from the deposited structure. Our method produced the systematic bias because 4EIX has an unstructured C-terminus loop that is supported by two disulfide bonds at C115 and C121 residues. Ignoring these bond restraints gave rise to unbalanced model bias and produced structures with displaced loops and poorer density maps.

To visualize a possible solution trajectory, we chose one 3ADJ decoy that was refined from an initial 2.7-Å rmsd down to 0.65-Å rmsd in a single unsupervised run. In Fig. 6 we plot the highest (red) and lowest (blue) R-free factor of the 50 trial structures generated at every macrocycle. In addition, the black line in Fig. 6 represents rmsd from the deposited structure of the parent model plotted on another scale. The structure was solved in 80 macrocycles, with the 0.65-Å residual rmsd being likely due to the 3.0-Å resolution of the data. One important thing to note is that at final macrocycles Buccaneer and Refmac consistently produce structures that are better than the deposited structure (dashed line in Fig. 6), which was also refined by Refmac (36). In addition, the high-quality trial structures generated at late-stage macrocycles form an ensemble of possible solutions of the phase problem and thus provide insights into the structural heterogeneity of different parts of the protein and the lower bound of the atomic coordinate errors (37). This is strikingly different from the amount of information contained in the single-structure solutions typically built by human crystallographers, where the structural heterogeneity is modeled by temperature factors, which sometimes does not provide accurate interpretation of the data (38).

## Discussion

Unlike standard refinement protocols using maximum-likelihood estimator target functions to fit the experimental data (39, 40), our method deforms a model in the direction that improves interpretability of density maps produced by combining experimental amplitudes with phases derived from the ensembles of trial models. We define interpretability as the ability of a computer program, in our case Buccaneer, to automatically build a good model whose quality is quantified by the model's R-free value. The particular method or program used to build and refine the models should not matter as long as it is applied consistently everywhere during the

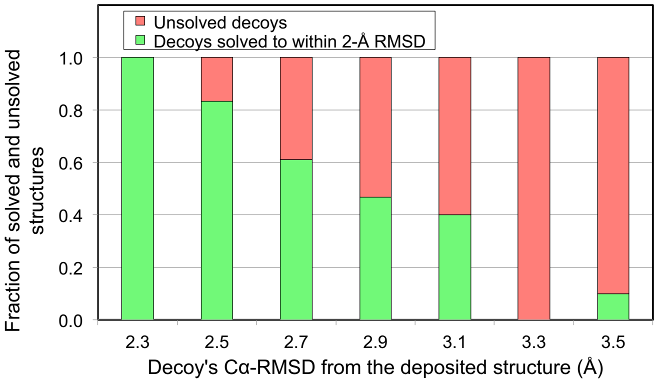

**Fig. 4.** The fraction of solved (green) and unsolved (red) decoys depends monotonically on the magnitude of their initial deformation with ~50% of cases solved at 2.9-Å Cα-rmsd deformation.

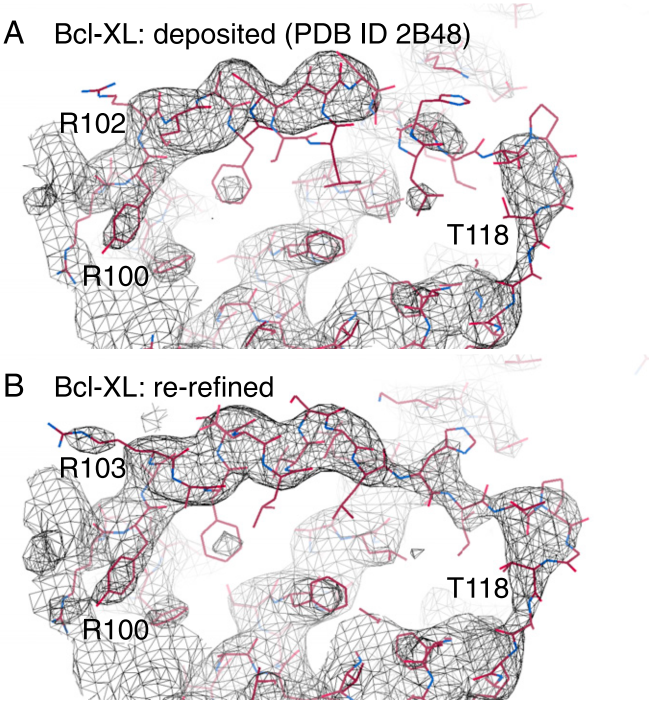

**Fig. 5.** Deposited (*A*) and rerefined (*B*) structure of Bcl-XL at 3.45-Å resolution. The $2mF_{obs}$-$DF_c$ electron density map is contoured at the $2\sigma$ level. The major structural difference is between R100 and T118 residues. Note the density peaks on R100 and R103 side chains in *B*, which are missing in *A*. The rerefined structure is depicted by the cyan circle in Fig. 3.

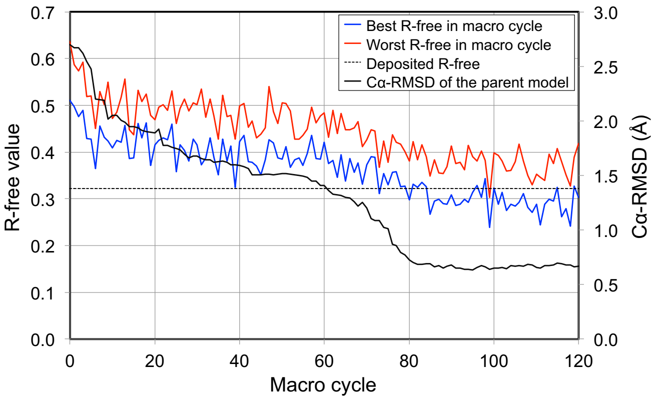

**Fig. 6.** Solution trajectory of one of the 3ADJ decoys. The red and blue lines, correspondingly, denote the highest and lowest R-free factor of the 50 trial structures generated at every macrocycle. The rmsd from the deposited structure of the parent model is shown as a solid black line. The dashed black line is the R-free value of the deposited structure.

solution process. In fact, we observed that it was better to trade accuracy for speed and build as many as possible trial models rather than rely on a smaller number of higher-quality models. This is partially due to the fact that electron density maps in principle cannot be fully interpreted at early stages of structure solution.

We observed that unlike the R-factor signal which degrades quickly with rmsd, the interpretability signal is more robust with respect to model deformations and allows us to explore and navigate through "flat" R-free surfaces. For instance, in the 30- to 60-macrocycle range in Fig. 6, the correlation coefficient between the best R-free value in a cycle (blue line in Fig. 6) and rmsd of the parent model (black line) is zero (−0.01), although rmsd still exhibits steady progress toward the solution.

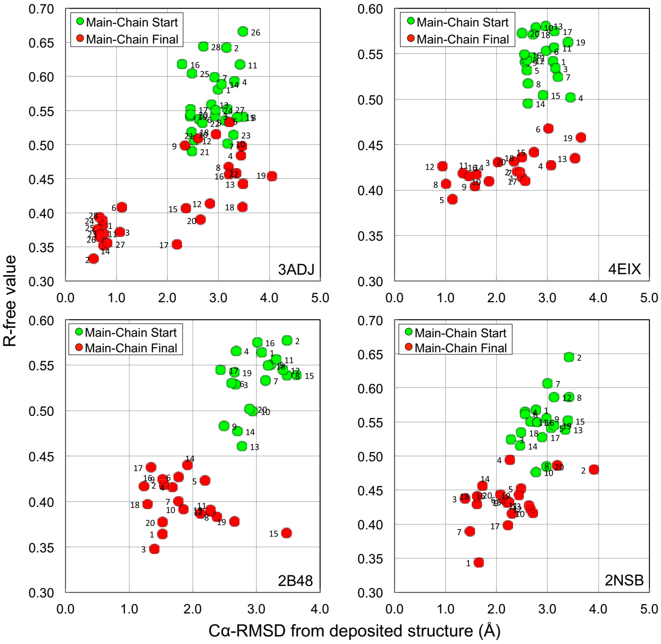

**Fig. 7.** Initial and final R-free values and Cα-rmsd of all of the parent structures (polyalanine backbones) used in the benchmark. Most models are consistently improved. There are a few false-positive structures with low R-free and large rmsd. None of these led to acceptable all-atom models.

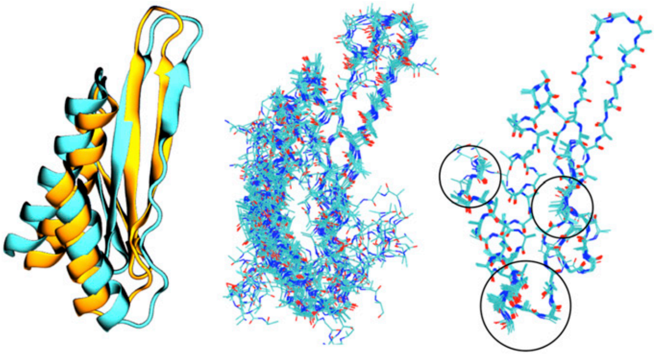

**Fig. 8.** (*Left*) Deposited 3ADJ structure (orange) and the initial decoy (cyan). (*Center*) Trial structures built at the first macrocycle by Buccaneer based on the density maps computed by our density modification code. One beta strand was interpreted quite accurately but the other strand was less clear. Attempts to build the beta turn at the right place are visible. Alpha helices are interpreted as cylinders due to the large displacement of the helices in the initial decoy. (*Right*) Trial structures built in the last macrocycle. Regions of structural heterogeneity are circled. All trial structures were superimposed to minimize their mutual rmsd and all side chains were removed for clarity.

In addition, Fig. 7 shows that essentially all decoys are improved in terms of rmsd and R-free metrics. Even the models that did not converge to the deposited structures to within 2-Å rmsd demonstrated systematic improvements. To our surprise, we discovered a few polyalanine chains (for instance, 2B48 decoy 15) that fitted the experimental data quite accurately yet deviated substantially from the corresponding deposited structures. Building full sequence models from such backbones never succeeded.

Fig. 8 shows the 20 overlaid best trial structures generated at the first and the last macrocycle iteration in the 3ADJ run shown in Fig. 6. More interpretable parts of the electron density map can be traced well enough to be visible in this ensemble representation and indicate the high-resolution part of the structure. Less interpretable parts of the density are represented by more random atom distribution that forms low-resolution structures, such as cylinders in place of alpha helices. Furthermore, regions of structural heterogeneity in the solved structure are visible in Fig. 8, *Right*.

## Methods

In the density modification protocol, the spacing of the real space grid was set to one-quarter of the dataset resolution, and all Fourier transforms were performed by the Nvidia CUDA FFT library. The calculations were carried out inside the unit cell with all space group symmetry operations handled in the real space explicitly. The Protein Data Bank (PDB) ID 5DTE structure was used to compute the reference density histogram for the density histogram projection. The binary protein mask includes all grid points located within 1.3-Å distance from any atom of the parent model. If the size of this distance-based protein mask is smaller than that estimated from the protein content of the unit cell *cV*, we compute the Gaussian-weighted density map fluctuations

$$\sigma_n{}^2 = \sum\nolimits_i \left(w_{i,n}\rho_i{}^2\right) \Big/ \sum\nolimits_i w_{i,n} - \left[\sum\nolimits_i \left(w_{i,n}\rho_i\right) \Big/ \sum\nolimits_i w_{i,n}\right]^2,$$

$$w_{i,n} = exp\left(-ad_{i,n}{}^2\right),$$

where *c* is the estimated protein content, *V* is the unit cell volume, $d_{i,n}$ is the Euclidian distance between points *i* and *n*, and the summation is performed over the entire unit cell. Then we add as many points with the largest $\sigma$ to the distance-based protein mask as needed to make its size will be equal to *cV*.

**ACKNOWLEDGMENTS.** M.L. is the Robert W. and Vivian K. Cahill Professor of Cancer Research. This research was supported by NIH Grant 5R35GM122543. This work used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (Grant ACI-1429830).

1. Hauptman HA (1991) The phase problem of X-ray crystallography. *Rep Prog Phys* 54: 1427.
2. Sayre D (2002) X-ray crystallography: The past and present of the phase problem. *Struct Chem* 13:81–96.
3. Taylor G (2003) The phase problem. *Acta Crystallogr D Biol Crystallogr* 59:1881–1890.
4. Brunger A (1990) Extension of molecular replacement: A new search strategy based on Patterson correlation refinement. *Acta Crystallogr A* 46:46–57.
5. Rossmann MG (1990) The molecular replacement method. *Acta Crystallogr A* 46: 73–82.
6. Read RJ (2001) Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr* 57:1373–1382.
7. Arnold E, Himmel DM, Rossman MG (2012) Crystallography of biological macromolecules. *International Tables for Crystallography* (Wiley & Sons, Chichester, United Kingdom), Vol F, pp 333–366.
8. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Cryst* 40:658–674.
9. Boutet S, et al. (2012) High-resolution protein structure determination by serial femtosecond crystallography. *Science* 337:362–364.
10. Chapman HN (2015) Serial femtosecond crystallography. *Synchrotron Radiat News* 28: 20–24.
11. Schlichting I (2015) Serial femtosecond crystallography: The first five years. *IUCrJ* 2: 246–255.
12. Sauter NK, et al. (2014) Improved crystal orientation and physical properties from single-shot XFEL stills. *Acta Crystallogr D Biol Crystallogr* 70:3299–3309.
13. Uervirojnangkoorn M, et al. (2015) Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *eLife* 4:e05421.
14. Nass K, et al. (2016) Protein structure determination by single-wavelength anomalous diffraction phasing of X-ray free-electron laser data. *IUCrJ* 3:180–191.
15. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826.
16. DiMaio F, et al. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* 473:540–543.
17. Evans P, McCoy A (2008) An introduction to molecular replacement. *Acta Crystallogr D Biol Crystallogr* 64:1–10.
18. Terwilliger TC (2003) Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallogr D Biol Crystallogr* 59:38–44.
19. Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* 62:1002–1011.
20. Cowtan K (2008) Fitting molecular fragments into electron density. *Acta Crystallogr D Biol Crystallogr* 64:83–89.
21. Terwilliger TC, et al. (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* 64:61–69.
22. Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3:1171–1179.
23. Adams PD, et al. (2010) PHENIX: A comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66:213–221.
24. Winn MD, et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242.
25. Afonine PV, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68:352–367.
26. Ufimtsev IS, Almagor L, Weis WI, Levitt M (2019) Solving the structure of Lgl2, a difficult blind test of unsupervised structure determination. *Proc Natl Acad Sci USA* 116:10819–10823.
27. Hattendorf DA, Andreeva A, Gangar A, Brennwald PJ, Weis WI (2007) Structure of the yeast polarity protein Sro7 reveals a SNARE regulatory mechanism. *Nature* 446: 567–571.
28. Read RJ (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr A* 42:140–149.
29. Murshudov GN, et al. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67:355–367.
30. Nicholls RA, Fischer M, McNicholas S, Murshudov GN (2014) Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr D Biol Crystallogr* 70:2487–2499.
31. Brünger AT (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472–475.
32. Buehler A, Urzhumtseva L, Lunin VY, Urzhumtsev A (2009) Cluster analysis for phasing with molecular replacement: A feasibility study. *Acta Crystallogr D Biol Crystallogr* 65: 644–650.
33. Chopra G, Summa CM, Levitt M (2008) Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci USA* 105:20239–20244.
34. Carroll JD, Arabie P (1980) Multidimensional scaling. *Annu Rev Psychol* 31:607–649.
35. Jack A, Levitt M (1978) Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallogr A* 34:931–935.
36. Yang SW, et al. (2010) Structure of Arabidopsis HYPONASTIC LEAVES1 and its molecular implications for miRNA processing. *Structure* 18:594–605.
37. Terwilliger TC, et al. (2007) Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. *Acta Crystallogr D Biol Crystallogr* 63:597–610.
38. Kuzmanic A, Pannu NS, Zagrovic B (2014) X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat Commun* 5: 3220.
39. Pannu NS, Read RJ (1996) Improved structure refinement through maximum likelihood. *Acta Crystallogr A* 52:659–668.
40. Adams PD, Pannu NS, Read RJ, Brünger AT (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc Natl Acad Sci USA* 94:5018–5023.