AlphaFold as a Prior: Experimental Structure Determination Conditioned on a Pretrained Neural Network

Alisia Fadini[1†], Minhuan Li[2†], Airlie J. McCoy[1], Thomas C. Terwilliger[4], Randy J. Read[1*], Doeke Hekstra[2,3*], Mohammed AlQuraishi[5*]

[1]Cambridge Institute for Medical Research, University of Cambridge.
[2]John A. Paulson School of Engineering & Applied Sciences, Harvard University.
[3]Department of Molecular and Cellular Biology, Harvard University.
[4]New Mexico Consortium.
[5]Department of Systems Biology, Columbia University.

*Corresponding author(s). E-mail(s): rjr27@cam.ac.uk; doeke_hekstra@harvard.edu; m.alquraishi@columbia.edu;
Contributing authors: af840@cam.ac.uk; minhuanli@g.harvard.edu; ajm201@cam.ac.uk; tterwilliger@newmexicoconsortium.org;
†These authors contributed equally to this work.

Abstract

Advances in machine learning have transformed structural biology, enabling swift and accurate prediction of protein structure from sequence. However, challenges persist in capturing sidechain packing, condition-dependent conformational dynamics, and biomolecular interactions, primarily due to scarcity of high-quality training data. Emerging techniques, including cryo-electron tomography (cryo-ET) and high-throughput crystallography, promise vast new sources of structural data, but translating raw experimental observations into mechanistically interpretable atomic models remains a key bottleneck. Here, we aim to address these challenges by improving the efficiency of structural analysis through combining experimental measurements with a landmark protein structure prediction method – AlphaFold2. We present an augmentation of AlphaFold2, ROCKET, that refines its predictions using cryo-EM, cryo-ET, and X-ray crystallography data, and demonstrate that this approach captures biologically important structural variation that AlphaFold2 does not. By performing structure optimization in the space of coevolutionary embeddings, rather than Cartesian coordinates, ROCKET automates difficult modeling tasks, such as flips of functional loops and domain rearrangements, that are beyond the scope of current state-of-the-art methods and, in some instances, even manual human modeling. The ability to efficiently sample these barrier-crossing rearrangements unlocks a new horizon for scalable and automated model building. Crucially, ROCKET does not require retraining of AlphaFold2 and is readily adaptable to multimers, ligand-cofolding, and other data modalities. Conversely, our differentiable crystallographic and cryo-EM target functions are capable of augmenting other structure prediction methods. ROCKET thus provides an extensible framework for the integration of experimental observables with biomolecular machine learning.

1 Introduction

Machine learning (ML) has revolutionized structural biology by enabling highly accurate protein structure prediction. Breakthrough models such as AlphaFold2 (AF2) [1], RoseTTAFold [2], and their descendants harness large-scale sequence data and coevolutionary signals to produce predictions with atomic-level precision and with accuracy often rivaling experimentally resolved structures [3, 4].

Despite this remarkable accomplishment, computational approaches still face significant limitations, particularly in capturing important properties such as sidechain packing, condition-dependent dynamics relevant to protein function and disease, and large-scale molecular assembly [5–9].

The success of current ML models relies heavily on a vast collection of experimentally resolved structures – an achievement made possible by decades of scientific effort by the structural biology community. However, deep learning-based structure prediction is constrained by the scarcity of high-quality ground-truth data that capture dynamic conformational states, biomolecular interactions, and structural variations observed in low-resolution experimental datasets. Recent advances in high-throughput experimental techniques are enabling the generation of large-scale datasets that could drive the supervised training of next-generation prediction methods [10, 11]. For example, modern crystallography beamlines and single-particle cryo-electron microscopy (cryo-EM) enable the resolution of numerous structures under diverse perturbations, creating new opportunities for drug screening and structural dynamics studies. X-ray crystallography can report on conformations induced by ligand binding with sub-Ångstrom-level accuracy [12–14] and can track structural changes along a reaction coordinate in real-time [15, 16], while advances in cryo-electron tomography (cryo-ET) enable *in situ* observations of macromolecular complexes [17, 18]. These developments hold the potential to accelerate our understanding of conformational flexibility and biomolecular complex assembly. Realizing this potential, however, requires scalable methods that can effectively process noisy experimental data.

A key hurdle in the effective interpretation of structural data is the necessary reconstruction of atomic models from raw observations. To a certain extent, ML methods have helped streamline this process by providing high-quality starting points for atomic model building [6, 19, 20]. Standard refinement software [21–23] operates on these starting points in Cartesian coordinate space, optimizing structures using gradients from experimental likelihoods, and is complemented by pattern matching-based model building to overcome local barriers [24–26]. This combination is effective at high resolution but struggles when the structural rearrangements presented by the data are large, such as switches in secondary structure, flips in flexible loops, or shifts in relative domain orientations. Model building becomes particularly challenging, even for humans, below 4–5 Å resolution, where critical structural details such as sidechains become indistinct. This difficulty is especially pronounced in the emerging field of cryo-ET, which currently tends to produce data at very low resolution (between 3 Å and 40 Å for structures deposited in 2024 [27]). Leveraging new priors that guide model building could therefore both improve accuracy in the low-resolution regime and reduce the need for labor-intensive manual intervention when the initial structures deviate from the experimental data.

We hypothesized that the implicit structural knowledge embedded in pre-trained ML structure prediction methods could guide atomic model building from experimental data more efficiently than traditional geometric restraints [28]. Historically, the integration of model and experiment was facilitated by the explicitly physics-based nature of structure prediction methodology, which readily permitted the incorporation of experiment-derived potential functions [29, 30]. The advent of "black-box" ML models like AF2, however, has made this type of integration less straightforward. Recent efforts aimed at adapting ML-based structure prediction to incorporate experimental constraints have done so either through fine-tuning of existing methods [31, 32] or full retraining, as exemplified by the ModelAngelo method for cryo-EM data [33]. While promising, these strategies are computationally expensive, data-hungry, and specific to the data modality they were trained on. Another, arguably more flexible approach, operates on the outputs of ML methods without modifying these methods directly. One such example is PredictAndBuild [6, 34], which does not require additional training but instead iterates between AF2 predictions and targeted rebuilding and refinement based on experimental data. However, its decoupled prediction and rebuilding steps can work against each other and lead to suboptimal convergence, limiting its effectiveness in complex cases.

Ideally, existing ML methods could be leveraged – without retraining – to accelerate and automate atomic model building. Indeed, some contemporaneous works have begun exploring this direction, but their applicability to challenging modeling tasks remains unclear [35, 36]. Here, we sought to combine the high accuracy of pre-trained AF2 structure prediction with guidance from experimental data in a way that can be flexibly adapted to different data modalities. We achieve this through ROCKET (Refining OpenFold with Crystallographic/Cryo-EM liKElihood Targets), a framework that integrates OpenFold [37], a trainable re-implementation of AF2, with differentiable crystallographic and cryo-EM likelihood targets [38]. Unlike approaches that require retraining or fine-tuning,

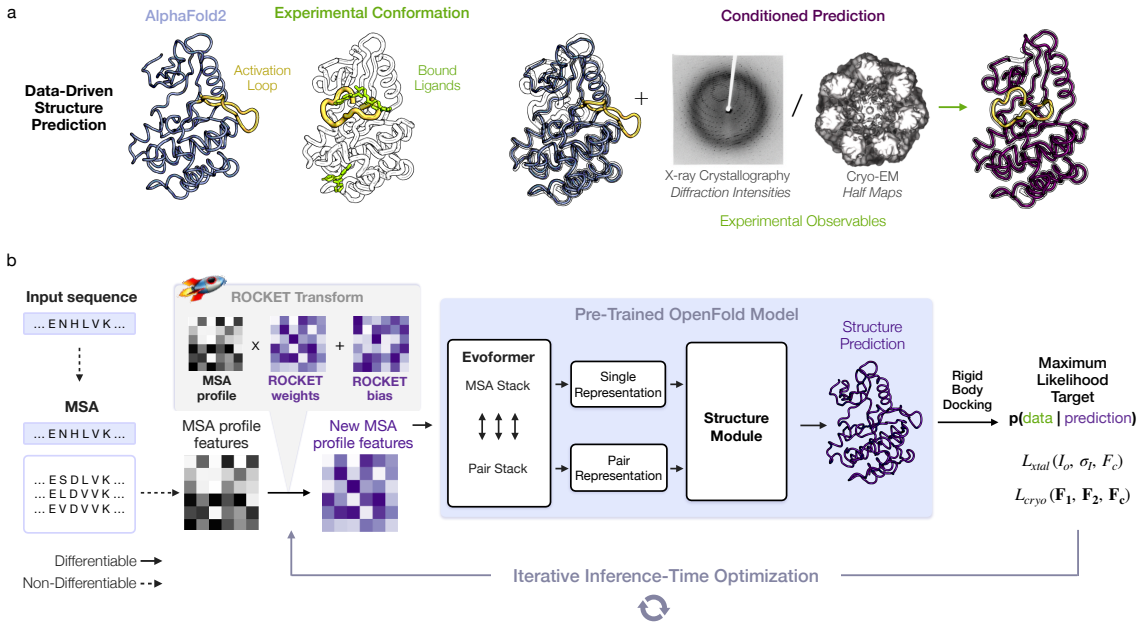


Fig. 1 Data-Driven Structure Prediction Refinement with ROCKET. (a) Experimental techniques such as X-ray crystallography and cryo-EM provide high-fidelity observations of conformational states, capturing, for example, ligand-induced changes (right; activation loop of human c-Abl kinase trapped in the active state by a small-molecule activator – PDB ID 3PYY) that may not be modeled by ML-based predictions (left; AF2 prediction of c-Abl in the untrapped state). (b) ROCKET extends OpenFold by integrating crystallographic and cryo-EM likelihood targets within its differentiable prediction pipeline. It accomplishes this by learning, at inference-time, multiplicative and additive adjustments to MSA cluster profiles that maximize their agreement with experimental data as computed by an experimental likelihood function. For the crystallographic target, the function (L_{xtal}) depends on observed crystallographic diffraction intensities and their measurement errors (I_o, σ_I) and on the structure factor amplitudes computed from the predicted model (F_c). For the cryo-EM target, the function (L_{cryo}) depends on the complex Fourier terms from experimental half-maps ($\mathbf{F_1}$ and $\mathbf{F_2}$) and the complex structure factors computed from the predicted model ($\mathbf{F_c}$).

ROCKET refines structures at inference time, steering predictions toward experimentally supported conformations (Fig. 1a). Inspired by recent strategies for expanding AF2 to conformational sampling [39–43], we directly optimize the embedded multiple sequence alignment cluster profile [40, 41] in AF2, transforming structure refinement to a data-guided search within AF2's continuous space for coevolutionary information.

Cryo-EM and crystallography are the predominant techniques for producing atomic-level insights into conformational heterogeneity, macromolecular interactions, and functional structural rearrangements. Despite crystallography's promise for high-resolution, high-throughput screening of different experimental conditions [44–46] and drug candidates [47], machine learning applications in protein crystallography [38, 48] have lagged behind cryo-EM [33, 49–51], largely due to inherent challenges presented by crystallographic symmetry and phase loss. ROCKET enhances the data-guiding capabilities of both techniques by efficiently integrating experimental data directly into OpenFold's inference process. We find that performing optimization in the latent embedding, instead of in traditional Cartesian coordinate space, allows for barrier-crossing structural rearrangements – such as domain shifts and drug-induced conformational changes – that conventional refinement methods struggle to access. Our evaluations on crystallographic test cases demonstrate that ROCKET can reliably capture backbone and sidechain conformations with accuracy comparable to that of manually curated models. Moreover, subsampling multiple sequence alignments to generate alternative profile embeddings improves initial predictions, helping to overcome gradient-descent convergence barriers in challenging cases.

ROCKET is particularly valuable for the unsolved challenge of model building at very low resolution [52], which is especially relevant in cryo-ET applications. Existing software struggles below 4–5 Å: ModelAngelo was not trained to work with maps below 4 Å [33] and PredictAndBuild will not automatically rebuild below 3.5 Å [6]. We find that ROCKET's inference-time refinement allows it to explore a wide conformational space and that it remains robust for atomic modeling in noisy, low-resolution maps. Our results show that ROCKET not only improves automated model building

but also outperforms expert manual refinement in particularly challenging cases. Since ROCKET relies on an input sequence for AF2 inference, it is complementary to other ML approaches that are designed to identify novel proteins in cryo-EM maps [33, 53, 54]. These capabilities render ROCKET a scalable and generalizable approach for integrating experimental data with ML-based biomolecular modeling.

2 Results

2.1 Method Overview

AF2-based structure prediction begins with the construction of a multiple sequence alignment (MSA), typically through a search of sequence databases for proteins homologous to the query of interest. The resulting MSA comprises aligned sequences of identified homologs, summarizing a protein's evolutionary history including its patterns of residue-residue coevolution. AF2 transforms raw MSAs into an input representation suitable for neural network computation, known as the MSA cluster profile. Through many studies and ablations, it has been shown that the depth and diversity of the MSA, and the statistical patterns found therein, determine the geometry and quality of a structure predicted by AF2 [1]. Given this central role, we reasoned that direct optimization of the continuous space of MSA cluster profiles would provide the greatest lever for influencing AF2 predictions; a hypothesis supported by recent observations made by other groups [40, 41] as well as our own experiments (Fig. S1).

To operationalize this principle, ROCKET augments OpenFold with a new module that optimizes MSA cluster profiles to maximize agreement between predictions and experimental observables (Fig. 1b; see Methods for algorithmic details). Through an experimental target function that quantifies this agreement, ROCKET performs gradient descent in the space of MSA cluster profiles. During each descent step, ROCKET computes a forward pass through OpenFold, evaluates the target function and its derivatives, then updates the MSA cluster profile to increase the function's agreement with model predictions. ROCKET currently provides two target functions, for crystallographic and cryo-EM data, and is extensible to other data modalities.

Since OpenFold-predicted structures are generated in an arbitrary reference frame, ROCKET first performs molecular replacement [55] or cryo-EM docking [56] to align the predicted model with the experimental data before starting the iterative refinement process. The resulting roto-translation is applied at every subsequent iteration to align the model. While only approximately optimal, previous observations indicate that initial AF2 predictions are usually sufficiently accurate for robust placement in the data [19, 20, 34, 57]. After iterative refinement through OpenFold is complete, a final local structure refinement is performed using phenix.refine [21] to optimize local geometry and atomic displacement parameters. Owing to memory limitations, ROCKET currently only operates on one protein chain at a time and requires special handling for cases involving multiple chains (see "Multi-Chain Dataset Handling" in Methods).

2.2 Evaluation Dataset and Approach

We took a two-pronged approach to assessing ROCKET's effectiveness in guiding structure prediction with experimental data. In the first, we focused on minor refinements and high-resolution datasets that are already well-served by existing, conventional methods, and where ROCKET's primary utility would be in streamlining experimental model building through integration with ML-based structure prediction. Here, our focus was on matching the best methods on relatively easy problems. In the second and more ambitious prong, we tackled cases where existing automated methods break down, either due to the magnitude and complexity of the required rearrangements or the low resolution of the data; cases in which manual human intervention is necessary and may not even be sufficient.

For the first prong, we curated a diverse set of 27 high-resolution, single-chain X-ray crystallographic datasets and their corresponding deposited structures (Fig. S2). We prioritized crystallography because, as noted earlier, its integration with ML-based methods remains underdeveloped [38]. All 27 datasets were solved at resolutions better than 3 Å (Table 1), a regime where conventional methods, such as phenix.refine, perform well and hybrid methods, such as PredictAndBuild, perform extremely well [34].

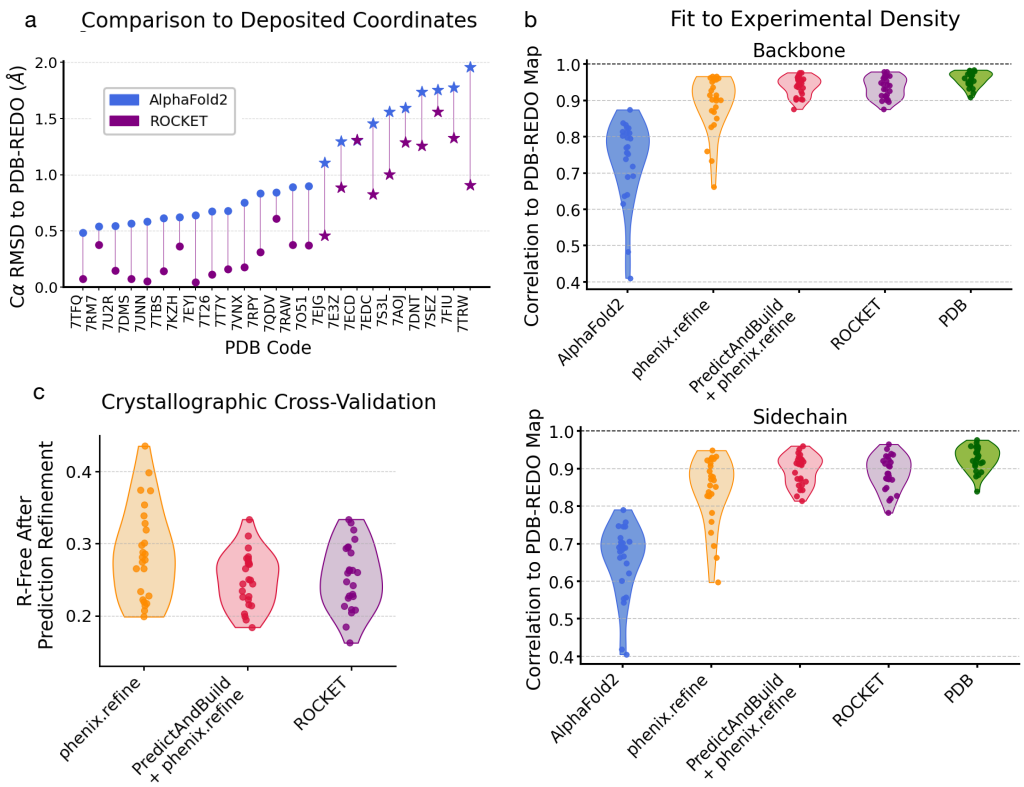


Fig. 2 ROCKET Refines a Diverse Set of Predictions Using High-Resolution X-ray Crystallography Datasets. (a) Cα RMSDs of conventional AF2 and ROCKET-refined predictions with respect to "ground truth" PDB-REDO models. AF2 predictions deviating by more than 1 Å from PDB-REDO models are highlighted with a star and generally correspond to instances where ROCKET performs the largest structural corrections. (b) Real-space Pearson correlation coefficient (RSCC) values of electron density maps derived from model structures vs. experimental amplitudes and unbiased phases from PDB-REDO. PDB-REDO maps are expected to have a favorable phase bias for deposited PDB models, artificially increasing their reported RSCCs. (c) Crystallographic R-free values for models placed by molecular replacement and refined using different methods.

For the second prong, we examined large-scale structural changes in ligand-induced loop rearrangements in three human proteins: c-Abl kinase (PDB ID 3PYY), protein tyrosine phosphatase 1B (PTP-1B, PDB ID 1NWL), and the serpin plasminogen activator inhibitor-1 (PAI-1, PDB ID 1JL5). We also examined a cryo-EM structure of *E. coli* GroEL (PDB ID 9C0B) where the subunits adopt a markedly different conformation from the original AF2 prediction. Finally, to assess performance in the low-resolution regime, we selected a 3.82 Å crystallographic dataset of the human multidomain protease inhibitor (HAI-1, PDB ID 5H7V) and a 9.60 Å sub-tomogram average of *E. coli* GroEL (PDB ID 8P4P).

2.3 ROCKET Reliably Refines High-Resolution Structures

At high resolution, we expect deposited backbone and sidechain coordinates to serve as reliable ground truths for evaluation. To validate ROCKET's performance on the 27 crystallographic structures in our first prong, we compared its refined models to those from the PDB-REDO database [58], which systematically re-refines and validates human-deposited coordinates. Fig. 2a shows Cα root-mean-squared deviations (RMSDs) with respect to PDB-REDO models for original AF2 predictions and ROCKET-refined ones. ROCKET improves all AF2 predictions except one, bringing them closer to the experimentally determined structures. Focusing on the ten most difficult cases, where AF2 predictions deviate by more than 1 Å from the PDB-REDO models, ROCKET achieves substantial structural corrections (Fig. 2a, stars; mean RMSD drop of 0.47 Å), demonstrating robustness in challenging scenarios.

Although Cα RMSD is a convenient measure of overall model quality, it is prone to underestimating discrepancies between atomic models and data, especially in flexible loops or sidechains that exhibit low occupancy or disorder. For a more sensitive assessment, we employed the real-space Pearson correlation coefficient (RSCC), which directly compares model-derived electron density maps to maps

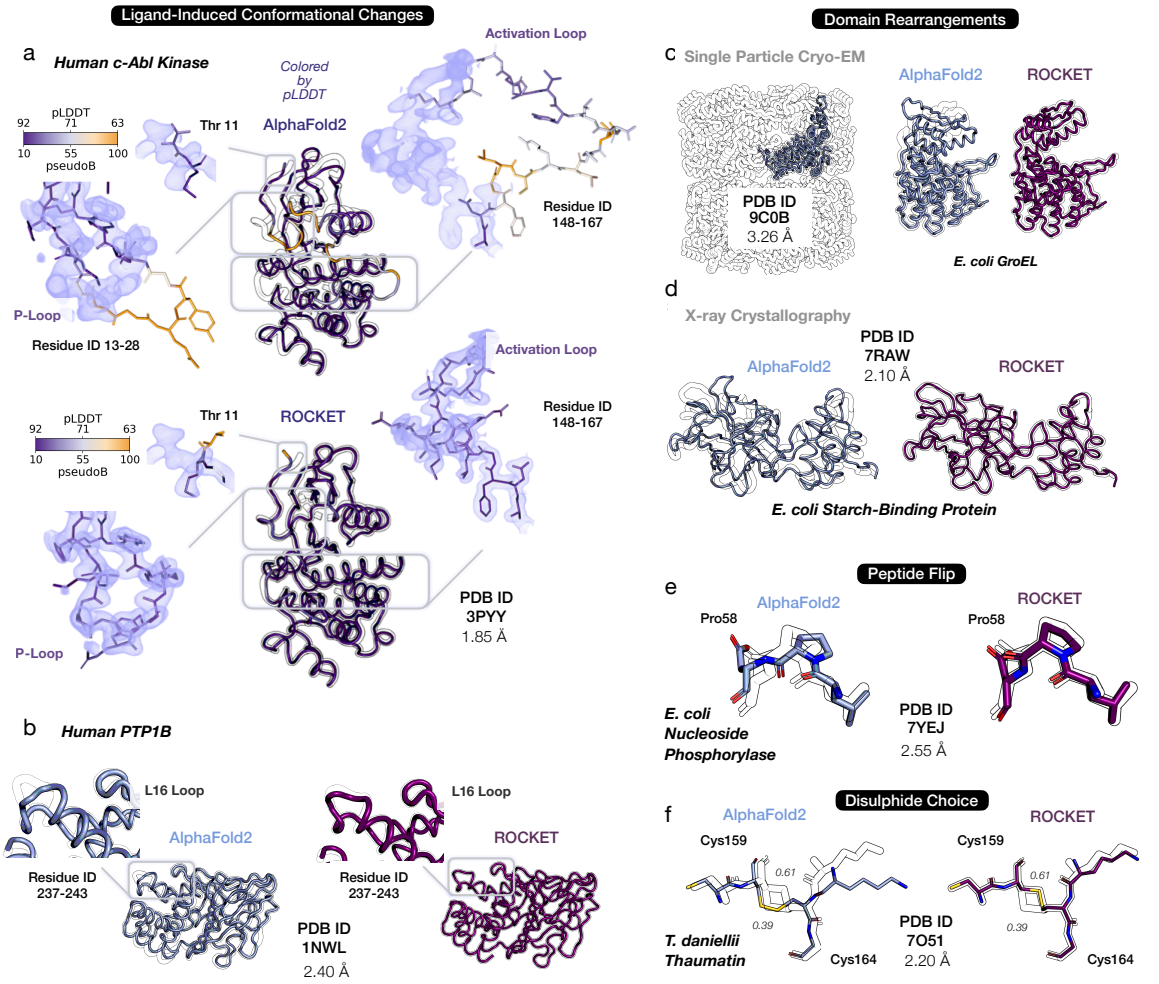


Fig. 3 ROCKET Overcomes Structural Barriers and Samples Diverse Conformations. Shown are representative cases that highlight ROCKET's ability to sample structural changes inaccessible to conventional refinement methods. In each panel, predictions by AF2 (top or left; light blue) and ROCKET (bottom or right; dark purple) are superimposed on experimental structures (black outline). Insets highlight local regions that undergo major changes. (a) Ligand-induced loop rearrangement in human c-Abl kinase (PDB ID 3PYY). ROCKET produces large conformational shifts of the activation and P-loops upon ligand binding. Per-residue confidence (pLDDT; blue-orange color scale) from AF2 is tracked by ROCKET and found to positively correlate with the final prediction's agreement with experimental data. (b) Conformational changes in protein tyrosine phosphatase PTP-1B (PDB ID 1NWL). A peptide flip in the L16 loop, induced by ligand binding, is captured by ROCKET. Ligand molecules (not shown) are not considered by ROCKET, as its refinement depends purely on experimental density without explicit modeling of protein-ligand interactions. (c-d) Domain rearrangements for a cryo-EM dataset of *E. coli* GroEL (PDB ID 9C0B) and a crystallographic dataset of *E. coli* starch-binding protein (PDB ID 7RAW). (e) A peptide flip in the refinement of *E. coli* nucleoside phoshorylase (PDB ID 7YEJ). (f) A switch in a disulphide bond in the structure of thaumatin from *T. daniellii* (PDB ID 7AOJ). Two alternate conformations are present in the deposited structure, with refined occupancies of 0.31 and 0.69. ROCKET builds the conformation with highest occupancy.

from experimental amplitudes (here, combined with phases from PDB-REDO models). As shown in Fig. 2b, ROCKET substantially improves RSCC values for both backbones and sidechains across all test cases. Our fully automated workflow achieves accuracies comparable to human-deposited models and performs on par with PredictAndBuild. Notably, optimizing MSA cluster profiles not only corrects secondary structures but also improves sidechain fit to the data (Fig. S4). Further validation using crystallographic R-free factors [59] (Fig. 2c) confirms ROCKET's added value: all but two models see improvements relative to phenix.refine alone, with 10 models showing R-free reduction of more than 3%. ROCKET's performance on R-free is comparable to the more complex combination of PredictAndBuild + phenix.refine, and the two can complement one another (Fig. S5).

We also performed an important control experiment: AF2 includes built-in confidence metrics, which several studies have leveraged to explore conformational space, particularly in the context of protein design [11, 40, 60, 61]. We hypothesized that by optimizing MSA cluster profiles with respect to AF2's own confidence metrics, it may be possible to improve structure prediction without the

need for experimental data. To test this hypothesis, we used ROCKET to optimize pLDDT, AF2's primary confidence metric. This approach fails to improve the correspondence of AF2 predictions with experiment (Fig. S6), indicating that experimental data provide new and orthogonal information, beyond AF2's implicit scoring function, that is necessary for efficient sampling of functionally relevant conformations.

2.4 ROCKET Overcomes Structural Barriers and Samples Diverse Conformations

We next asked whether ROCKET can perform challenging large-scale structural rearrangements – such as functional loop movements, peptide flips, and domain shifts. Such rearrangements are often beyond the scope of current state-of-the-art methods but may be easy to sample in MSA space. Fig. 3 shows ROCKET-produced loop rearrangements in c-Abl kinase (Fig. 3a) and PTP-1B (Fig. 3b), as well as subunit rebuilding of *E. coli* GroEL (Fig. 3c). In all three cases, ROCKET successfully samples barrier-crossing transitions that are critical for modeling functional states. Beyond these large-scale motions, ROCKET also accurately refines domain rearrangements and smaller structural modifications, such as peptide flips in high-resolution structures (Fig. 3d-f). In contrast, PredictAndBuild, despite its own iterative use of AF2, largely fails to access these conformational changes (Fig. S7).

Intriguingly, we find that, across our datasets, ROCKET's AF2-derived model confidence (measured as per-residue pLDDT) is correlated with agreement with experimental data, as reflected in the positive correlation between final model confidence and RSCC (Fig. S8). We visually highlight an example of this in the refinement of the activation- and P-loops of c-Abl kinase (Fig. 3a). Since MSAs reflect the variety of conformations a protein can adopt [39], we suggest that ROCKET's data-guided optimization uses, and then resolves, this structural ambiguity, allowing it to reach functional conformations that conventional refinement protocols cannot access. This clarifies, but does not contradict, our finding from the preceding section that pLDDT-driven optimization cannot refine predicted structures. Specifically, it indicates that pLDDT can identify highly preferred conformations but cannot, by itself, distinguish the experimentally observed state (and provide a useful gradient) without further experimental information.

2.5 ROCKET Integrates MSA Subsampling to Escape Local Minima in Large Transitions

The effectiveness of a gradient descent approach is inherently limited by its convergence radius. When an initial AF2 prediction deviates significantly from the experimentally observed conformation, the gradient signal derived from the experimental likelihood target may become weak or misleading, impeding refinement. Inspired by recent findings that alternative folds can be uncovered through subsampling of the input MSA [39], ROCKET subsamples input MSAs to obtain diverse MSA cluster profiles – and thereby fresh initial models – that may lie closer to the experimental conformation. For each subsampled MSA, a single-shot prediction is made (as well as one from the full MSA) and scored using the experimental likelihood. The MSA resulting in the prediction with the highest likelihood (and thus most likely to accurately capture the experimentally resolved backbone arrangement) is subsequently selected for full refinement.

We illustrate this technique using the 1.8 Å crystallographic dataset of the human serpin PAI-1, captured in its energetically hyperstable latent state (Fig. 4a; left). The initial AF2 prediction, based on a full sequence alignment, adopts the protein's metastable active conformation and deviates substantially from the experimentally determined model (Fig. 4a; right; 10.77 Å overall RMSD and 40.4 Å loop RMSD). A notable structural difference involves the reactive center loop (RCL, residues 330–354): it undergoes a large-scale structural transition from a solvent-exposed position in the active state to being fully inserted into the central β-sheet in the latent conformation. This rearrangement poses a substantial challenge for gradient descent optimization when refinement is started from the full MSA (Fig. 4b; 3.2 Å overall RMSD and 36.8 Å loop RMSD). On the other hand, subsampled MSAs result in starting conformations that more closely resemble the latent state (Fig. 4c; 4.2 Å overall RMSD and 13.6 Å loop RMSD for best starting structure). We note here that the experimental likelihood score does not monotonically correlate with pLDDT (Fig. 4c; lower panel), indicating that MSA selection based solely on AF2's confidence metric offers no assurance of identifying the relevant

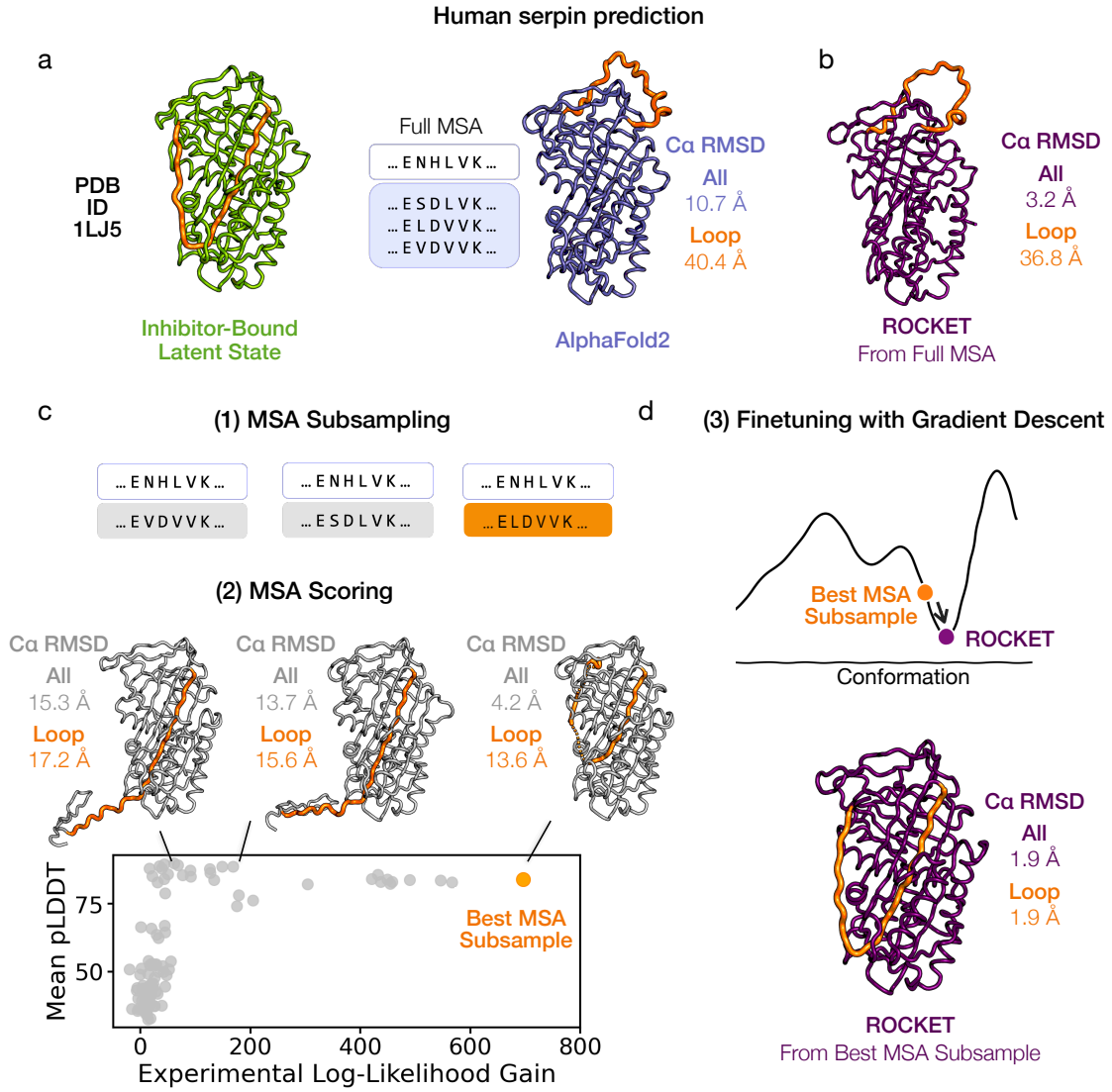


Fig. 4 Synergistic Integration of MSA Subsampling with Gradient-Based Optimization. Gradient-based refinement faces a limit when initial predictions deviate substantially from the target conformation. (a) The structure of a human serpin, plasminogen activator inhibitor-1 (PAI-1), in its latent state (PDB ID 1LJ5, 1.8 Å) illustrates this issue (left). A conventional AF2 prediction, generated from the full MSA, yields the metastable active conformation that diverges from the latent state captured by the experimental structure (right). (b) The latent state conformation is not accessible by gradient descent alone when ROCKET refinement is started from the full MSA (c) Subsampled MSAs (1) result in predictions that more closely resemble the experimental conformation (2). Ranking these predictions (bottom panel) by their experimental likelihoods identifies a better starting model for gradient-based refinement (samples are ordered along the x-axis by increasing gains in experimental likelihood), unlike pLDDT-based scoring, which does not (highest samples along the y-axis do not resemble the experimental conformation). (d) Gradient-based refinement of the best starting prediction from (c) results in a structure that closely resembles the latent conformation.

functional state. ROCKET subsequently uses the highest likelihood MSA to perform gradient-based optimization, producing a structure that closely resembles the latent state (Fig. 4d; 1.9 Å overall and loop RMSD).

2.6 ROCKET Extracts Essential Information from Low-Resolution Data

We next assess ROCKET's applicability in the low-resolution regime with two datasets that lie below the typical resolution limits for automated model building: a 3.82 Å human HAI-1 crystallographic dataset and a 9.60 Å sub-tomogram average of *E. coli* GroEL.

In well-resolved regions of HAI-1, ROCKET successfully converges to a backbone conformation that more closely matches the experimental density than the AF2 prediction (Fig. 5a-b; regions 1, 2, and 5). Conversely, in areas where the experimental map is poorly resolved – such as region 3, which is not modeled in the deposited structure – ROCKET refrains from forcing arbitrary changes and instead

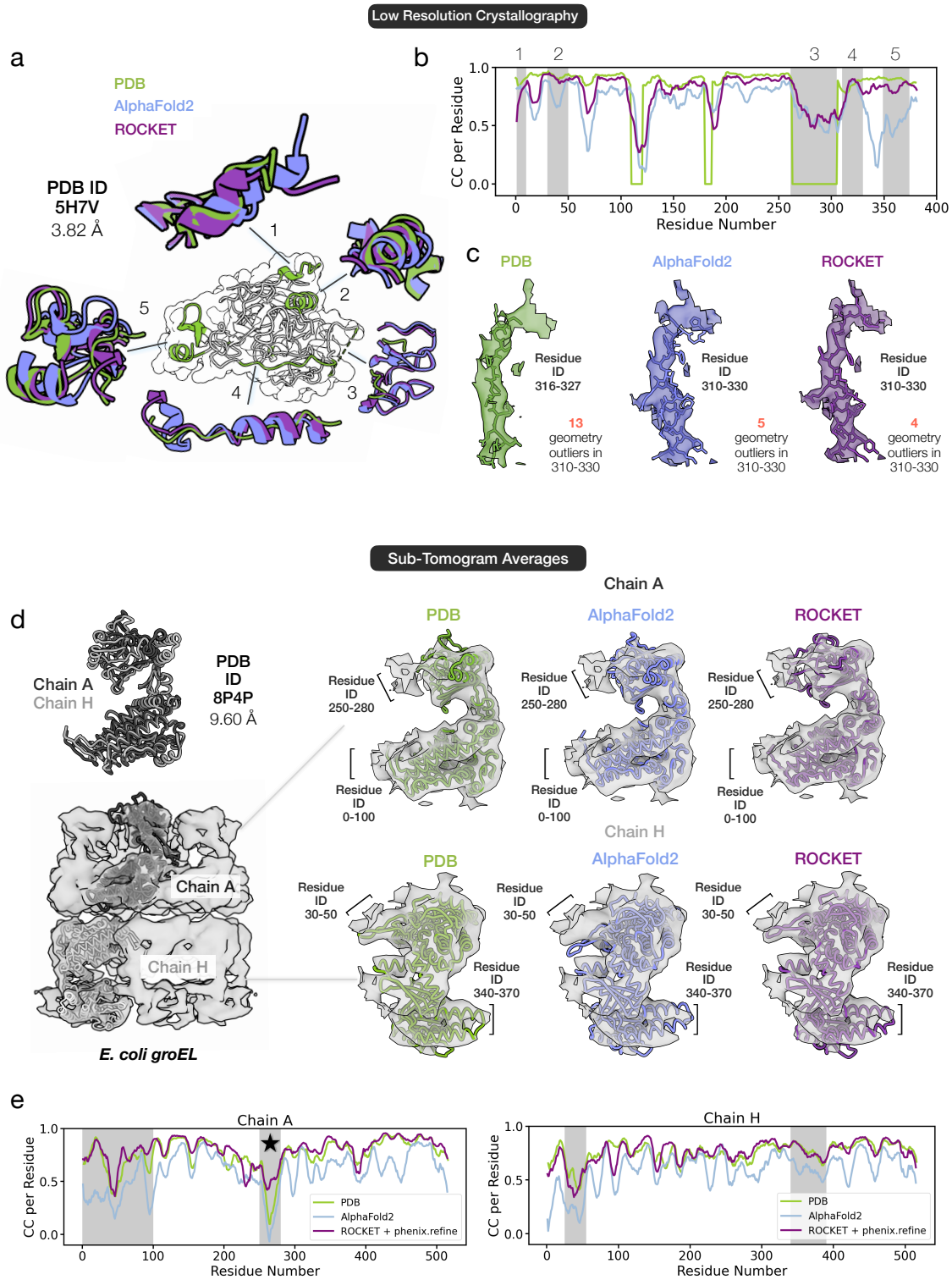


Fig. 5 Extracting Information from Low-Resolution Data. (a-c) Refinement of a 3.82 Å crystallographic dataset of the human protease inhibitor HAI-1 (PDB ID 5H7V). ROCKET accurately recovers backbone structures in regions where the data support the deposited model (regions 1, 2, and 5) and preserves the AF2 prediction in poorly defined regions (region 3). In region 4, where manual model building is hindered by noisy density and seemingly incorrect sequence register in the deposited model, ROCKET improves the initial AF2 prediction without introducing new geometric outliers. (d-e) Refinement of a 9.60 Å sub-tomogram average of *E. coli* GroEL (PDB ID 8P4P). ROCKET successfully predicts two distinct subunit conformations observed in the GroEL heptameric rings. Maps computed using the refined models correlate with the experimental map at a level comparable to human-built models. For the chain A conformer, ROCKET explores a broader conformational space, with its final model achieving a higher average RSCC of 0.5 for residues 250–280 (marked by a star) than the deposited model's RSCC of 0.2.

preserves the original AF2 prediction. Particularly noteworthy is region 4 (residues 310–330), where the density is highly noisy, making even manual model building difficult. In this segment, where the deposited model appears to have an incorrect sequence register, ROCKET improves AF2's initial prediction to better align with the density map without introducing new geometric outliers (Fig. 5c).

In the sub-tomogram average of GroEL (PDB ID 8P4P), the 9.60 Å map reveals two distinct conformations for the repeating subunit in the top and bottom heptameric rings (Fig. 5d). Despite this very low resolution, ROCKET accurately predicts both conformations (Fig. 5d), yielding refined models with map correlations comparable to manually built structures (Fig. 5e). Moreover, for chain A, ROCKET explores a broad conformational space, deviating substantially from the AF2 prediction. Notably, in the top domain (residues 250–280; marked by a star in Fig. 5e), ROCKET achieves an average RSCC of 0.5, substantially higher than the 0.2 achieved by the deposited model, supporting a closer match to experimental data.

These examples demonstrate that ROCKET extracts meaningful information at resolutions inaccessible to other methods. Moreover, in challenging regions such as the 310–330 helix of HAI-1 or the 250–280 residue range of GroEL chain A, ROCKET produces solutions that improve agreement with the experimental map while retaining structural plausibility, surpassing the accuracy of human-deposited models.

3 Discussion

In developing ROCKET, we have demonstrated that experimentally guided refinement of the embedded MSA profiles of AF2 enables efficient exploration of conformational space, capturing structural changes that conventional refinement in Cartesian space struggles to reach. Our results suggest that these embeddings provide access to paths along which the barrier for structural rearrangements is significantly reduced or eliminated, indicating that information about such rearrangements are encoded in evolutionary statistics.

We anticipate that data-guided, inference-time optimization will prove broadly valuable across diverse atomic model building scenarios. Beyond cryo-EM and crystallographic datasets, implemented here, our approach can be used for other, sparser data modalities, provided a likelihood-based target between data and prediction can be formulated. Extending ROCKET to handle multi-chain complexes or protein-ligand co-folding is straightforward within any AlphaFold-like framework, requiring in principle only a switch in the inference model used.

More ambitious directions may involve integrating generative models to account for conformational ensembles. Given our success with MSA profile biasing, we propose that learning a mapping from experimental observables to a profile bias matrix could effectively condition single-shot structure prediction towards experimentally probed conformations. Such a mapping would translate experimental data – such as cryo-EM density maps, X-ray diffraction intensities, or NMR-derived constraints – into a profile bias matrix that captures residue-level probabilities or pairwise constraints derived from experiment, effectively guiding the model towards relevant conformations without requiring exhaustive searches. By adopting amortized inference, which learns a reusable mapping to enable fast predictions for new inputs, we could further streamline the process by replacing the stepwise search (MSA subsampling and likelihood scoring, shown in Fig. 4) with a more efficient learned transformation.

While additional work is needed to fully realize these extensions, our study marks critical progress by showing that structure optimization in coevolutionary embeddings can overcome limitations of conventional refinement. Additionally, by introducing differentiable likelihood targets for cryo-EM and crystallography, which include a robust treatment of measurement error, we provide a framework well-suited for training future predictive models. More broadly, our Bayesian approach to combining information learned by ML models with information obtained by direct observation establishes a foundation for continuous interaction between machine learning and experimental data. This interplay is critical for the interconnected goals of scaling experimental throughput and training machine learning models with enhanced functionality.

Our approach is automated, requires no retraining, and for difficult cases produces models of quality comparable or even superior to those created by expert human modelers. However, certain limitations remain. The current approximation of the atomic displacement parameters is derived empirically

from model confidence, and could be improved by incorporating considerations of density fit at the residue level. Additionally, because OpenFold is not explicitly aware of crystal contacts, ROCKET may struggle to converge to certain lattice-constrained conformations. We anticipate that this limitation can be addressed by extending ROCKET using multi-chain models [62]. We also noticed that ROCKET can fail to flip small loops (3–4 residues in length) that contain bulky sidechains. We show examples of cases that are difficult for ROCKET in Fig. S9. Perhaps most importantly, iterative backpropagation through OpenFold is memory-intensive and limits the maximum size of the protein or domain that can be refined at once – about 500 residues on a 40 GB A100 GPU (Fig. S10). We believe this can be extended with further code optimization. The implementation of a learned mapping for single-shot MSA profile bias, mentioned above, could also help address this limit.

Naturally, assessing biological accuracy solely through the lens of how well the atomic model fits experimental maps has inherent limitations, as both cryo-EM and crystallography can introduce artifacts [63]. Ultimately, multi-modal information is essential for building a full picture of physiological protein states and functions. For this purpose, our approach is readily adaptable to alternative loss functions that combine multiple sources of experimental and computational data, supporting integrative modeling strategies for biological structure determination.

4 Methods

ROCKET Algorithm and Processing Pipeline

ROCKET's inputs and pre-processing steps are summarized in Algorithm 1, while the ROCKET inference-time optimization algorithm is summarized in Algorithm 2.

Inputs and Pre-processing Pipeline

For crystallographic datasets, a protein sequence and a reflection MTZ file (or CIF file) containing observed intensities and their uncertainties are required, while for cryo-EM datasets two half-maps are required. As outlined in Algorithm 1, to obtain an aligned reference model ($\mathbf{x}^{\text{ref}}$), we use Phasertng [64] for molecular replacement with crystallographic datasets and a likelihood-based docking tool, EM_placement [65], with cryo-EM datasets. These determine the correct pose from an initial, unconditional AF2 model. These tools also estimate experimental data parameters for refinement, such as E_e and D_{obs}, which represent observed normalized amplitudes and an accounting for measurement error, respectively [66], and are described further below.

Algorithm 1 ROCKET Preprocessing

```
function PREPROCESSING(ProteinSeq, ReflectionMTZ, HalfMaps)
    x^initial, pLDDT^ref = OpenFold_Inference(m_0^cluster_profile)
    pseudo-B^ref = pLDDT_to_pseudo-B(pLDDT^ref)
    if ReflectionMTZ then
        MR_solution, x^ref = Phasertng_MR(x^initial, ReflectionMTZ)
        E_e, D_obs = Phasertng_Preprocess(ReflectionMTZ)
        return x^ref, E_e, D_obs, pseudo-B^ref, Type: Crystallographic
    else if HalfMaps then
        x^ref = CryoEM_Dock(x^initial, HalfMaps)
        E_e, D_obs = CryoEM_ParameterEstimation(HalfMaps)
        return x^ref, E_e, D_obs, pseudo-B^ref, Type: Cryo-EM
    else
        return Error: Missing required data (ReflectionMTZ or HalfMaps)
    end if
end function
```

Refinement Algorithm

ROCKET optimizes a linear bias (with scales $\mathbf{w}$ and offsets $\mathbf{b}$) that it applies to the input MSA cluster profile ($\mathbf{m}_0^{\text{cluster_profile}}$) to maximize agreement between an OpenFold prediction ($\mathbf{x}^{\text{prediction}}$)

and experimental data (E_e). This agreement is quantified by a data log-likelihood gain (LLG) target, $\mathcal{L}_{\text{LLG}}$, which is combined with an optional positional restraint, $\mathcal{L}_{\text{L2}}$, to yield an overall objective function $\mathcal{L}$ (see below).

Algorithm 2 ROCKET Refinement

```
function REFINEMENT(E_e, D_obs, N_iter, x^ref, pseudo-B^ref, m_0^cluster_profile, lr_mul, lr_add, ω_L2)
    w = ones_like(m_0^cluster_profile)                          ▷ Initialize multiplicative bias
    b = zeros_like(m_0^cluster_profile)                         ▷ Initialize additive bias
    optimizer = adam({w : lr_mul, b : lr_add})
    for iter = 1 ... N_iter do
        m^cluster_profile = w ⊙ m_0^cluster_profile + b
        x^prediction, pLDDT = OpenFold_Inference(m^cluster_profile)
        pseudo-B = pLDDT_to_pseudo-B(pLDDT)
        x^aligned = Weighted_Kabsch(x^prediction, x^ref, pseudo-B^ref)
        x^RBR = Rigid_Body_Refinement(x^aligned, pseudo-B, E_e, D_obs)
        L_LLG = LLG(x^RBR, pseudo-B, E_e, D_obs)                ▷ Log-likelihood-Gain Targets
        L_L2 = Weighted_L2(x^RBR, x^ref, pseudo-B^ref)          ▷ Weighted L2 loss
        L = L_LLG + ω_L2 · L_L2
        L.backward()
        optimizer.step()
    end for
    return x^RBR, w, b
end function
```

During each ROCKET iteration, we apply a linear bias to the initial cluster profile matrix and then run OpenFold inference to obtain a new prediction, along with its pLDDT confidence values. Our current implementation estimates atomic B-factors at every iteration from pLDDT confidence scores using a previously established heuristic [2], without explicitly refining them. Specifically, we convert pLDDT values into equivalent RMSD values, using the empirical relation from Baek et al. [2], and then to corresponding pseudo B-factors:

$$\text{pseudo-B} = \frac{8\pi^2}{3}\left[1.5 \times e^{0.7-0.01\times\text{pLDDT}}\right]^2 \tag{1}$$

At every iteration, we align the newly predicted model to the reference model using a weighted Kabsch alignment (with pseudo B-factors serving as the weights), followed by rigid-body refinement (elaborated in a separate section below). We then compute the experimental log-likelihood-gain (LLG) using the aligned coordinates and the pseudo B-factors.

In practice, ROCKET runs in two phases: an "adventurous" phase 1 and a "fine-tuning" phase 2. We apply different learning rates for the multiplicative ($\mathbf{w}$) and additive ($\mathbf{b}$) components of our MSA profile bias, and these rates vary across phases. In phase 1, we use higher learning rates (by default, $\text{lr}_{\text{mul}} = 1.0$ and $\text{lr}_{\text{add}} = 0.05$) and a default low-resolution cutoff of 3 Å. We also incorporate a weighted C_α mean squared distance between the reference model and latest prediction as an L2 regularization term that quantifies initial model confidence. The weights are computed using the same scheme we employ for the weighted Kabsch alignment discussed further below:

$$\mathcal{L}_{\text{L2}} = w(\text{pseudo-B}) \cdot \left(\mathbf{x}^{\text{aligned}} - \mathbf{x}^{\text{ref}}\right)^2 \tag{2}$$

with the confidence-based weights $w(\text{pseudo-B})$ defined in Equation 13 and the final loss for backpropagation:

$$\mathcal{L} = \mathcal{L}_{\text{LLG}} + \omega_{\text{L2}}\mathcal{L}_{\text{L2}} \tag{3}$$

By default, the L2 loss weight is $\omega_{\text{L2}} = 10^{-11}$. We run phase 1 for three independent traces, each consisting of 100 iterations, and select the model with the best LLG score to proceed to phase 2. The aim of phase 2 is to further fine-tune the structure. By default, we set phase 2 to run for 500 iterations with both lr_{mul} and lr_{add} to 10^{-3} and remove the L2 loss term ($\omega_{\text{L2}} = 0$). An early stop occurs if the LLG score does not improve by more than 0.1 for 50 consecutive iterations.

The pseudo-B approximation can limit accuracy by not capturing finer structural details. Moreover, geometric validation indicates that outputs from the iterative optimization have more bond outliers and steric clashes than stricter refinement protocols typically allow (Fig. S3a). To address these limitations, we append a short standard local-refinement step using phenix.refine [21] after iterative OpenFold inference. Analogous to AMBER relaxation in standard AF2 pipelines [1], this step further optimizes geometry (Fig. S3) and displacement parameters, polishing the final model's overall quality.

Implementation

ROCKET is implemented in PyTorch 1.12.1 as an extension of the OpenFold system. It currently uses the monomer version of OpenFold with AF2 model_1 weights to maintain consistency with AF2's data splits. This allows for prediction and refinement of crystallographic datasets containing a single chain in the asymmetric unit or one domain (at a time) in a cryo-EM complex.

Crystallographic Log-Likelihood-Gain Targets

For crystallographic datasets, we employ the log-likelihood-gain on intensity (LLGI) target introduced in [66], where for acentric and centric reflections:

$$\begin{aligned}\mathrm{LLGI_a}\left(E_\mathrm{e};E_\mathrm{C}\right) &= \ln\left[\frac{p_\mathrm{a}\left(E_\mathrm{e};E_\mathrm{C}\right)}{p_\mathrm{a}\left(E_\mathrm{e}\right)}\right] \\ &= \ln\left[p_\mathrm{a}\left(E_\mathrm{e};E_\mathrm{C}\right)\right] - \ln\left[p_\mathrm{a}\left(E_\mathrm{e}\right)\right]\end{aligned} \tag{4}$$

$$\begin{aligned}\mathrm{LLGI_c}\left(E_\mathrm{e};E_\mathrm{C}\right) &= \ln\left[\frac{p_\mathrm{c}\left(E_\mathrm{e};E_\mathrm{C}\right)}{p_\mathrm{c}\left(E_\mathrm{e}\right)}\right] \\ &= \ln\left[p_\mathrm{c}\left(E_\mathrm{e};E_\mathrm{C}\right)\right] - \ln\left[p_\mathrm{c}\left(E_\mathrm{e}\right)\right]\end{aligned} \tag{5}$$

with $p_\mathrm{a}\left(E_\mathrm{e}\right) = 2E_\mathrm{e}\exp\left(-E_\mathrm{e}^2\right)$, $p_\mathrm{c}\left(E_\mathrm{e}\right) = \left(\frac{2}{\pi}\right)^{1/2}\exp\left(-\frac{E_\mathrm{e}^2}{2}\right)$, and:

$$\begin{aligned}p_\mathrm{a}\left(E_\mathrm{e};E_\mathrm{C}\right) =& \frac{2E_\mathrm{e}}{1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2}\exp\left[-\frac{E_\mathrm{e}^2+\left(D_\mathrm{obs}\sigma_\mathrm{A}E_\mathrm{C}\right)^2}{1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2}\right] \\ &\times I_0\left(\frac{2D_\mathrm{obs}\sigma_\mathrm{A}E_\mathrm{e}E_\mathrm{C}}{1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2}\right)\end{aligned} \tag{6}$$

$$\begin{aligned}p_\mathrm{c}\left(E_\mathrm{e};E_\mathrm{C}\right) =& \left[\frac{2}{\pi\left(1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2\right)}\right]^{1/2}\exp\left[-\frac{E_\mathrm{e}^2+\left(D_\mathrm{obs}\sigma_\mathrm{A}E_\mathrm{C}\right)^2}{2\left(1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2\right)}\right] \\ &\times\cosh\left(\frac{D_\mathrm{obs}\sigma_\mathrm{A}E_\mathrm{e}E_\mathrm{C}}{1-D_\mathrm{obs}^2\sigma_\mathrm{A}^2}\right)\end{aligned} \tag{7}$$

and $p(x;y)$ denotes the conditional probability of x given y.

As defined in [66], E_e is the "effective" observed normalized amplitude, E_C is the normalized structure factor amplitude calculated from the predicted model in a differentiable manner using SFCalculator with solvent correction [38], D_{obs} encodes the reduction in correlation between true and "effective" normalized structure factors arising from experimental error, and σ_A is a resolution-dependent factor that encodes the reduction in correlation between the true and calculated normalized structure factors arising from model error.

We refine σ_A in resolution bins at every ROCKET iteration using the Newton-Raphson optimization method [67]. The log-likelihood gain of the observed effective amplitudes, E_e, given the calculated amplitudes, E_C, is maximized by refining σ_A. The derivative of the log-likelihood gain with respect to σ_A for each resolution bin is given by:

$$\frac{\partial\,\mathrm{LLG}}{\partial\sigma_A} = \sum_i \frac{\partial\ln p\left(E_{\mathrm{e},i};E_{\mathrm{C},i},\sigma_A\right)}{\partial\sigma_A} \tag{8}$$

where i is an index over observations. The second derivative is given by:

$$\frac{\partial^2\,\mathrm{LLG}}{\partial\sigma_A^2} = \sum_i \frac{\partial^2\ln p\left(E_{\mathrm{e},i};E_{\mathrm{C},i},\sigma_A\right)}{\partial\sigma_A^2} \tag{9}$$

We obtain the updated value of σ_A using a Newton step:

$$\sigma_A^{\text{new}} = \sigma_A - \frac{\partial^2 \mathcal{L}_{\text{LLG}}}{\partial \sigma_A^2}^{-1} \frac{\partial \mathcal{L}_{\text{LLG}}}{\partial \sigma_A} \tag{10}$$

The value of σ_A is constrained within $[0.015, 0.99]$ to maintain physical relevance and stability during refinement. For all LLG and σ_A estimates we use a working set of reflections (we find that using the working set for σ_A refinement does not lead to any meaningful overfitting with ROCKET, Fig. S11). We keep a test set of reflections for final R-free calculation.

Log-Likelihood Gain Target and Noise Modeling for Cryo-EM

For cryo-EM data, we follow the method outlined in [65] to dock the initial prediction into the experimental map. We model the signal and noise in Fourier space to account for directional and resolution-dependent variations in data quality. The signal is derived from correlations between Fourier terms of the experimental half-maps, while the noise is deduced from their differences [56].

For a single Fourier term, the log-likelihood gain (LLG) is given by:

$$\begin{aligned} \text{LLG}_{\text{cryo}} = & \frac{2 \cdot D_{\text{obs}} \cdot \sigma_A \cdot E_{\text{e}} \cdot E_{\text{calc}} \cdot \cos(\Delta\phi)}{1 - D_{\text{obs}}^2 \sigma_A^2} \\ & - \frac{D_{\text{obs}}^2 \sigma_A^2 (E_{\text{e}}^2 + E_{\text{calc}}^2)}{1 - D_{\text{obs}}^2 \sigma_A^2} \\ & - \ln(1 - D_{\text{obs}}^2 \sigma_A^2) \end{aligned} \tag{11}$$

where E_{e} and E_{calc} are the normalized amplitudes of the observed and calculated Fourier terms, $\Delta\phi = \phi_{\text{calc}} - \phi_{\text{obs}}$ is the phase difference between these Fourier terms, and D_{obs} and σ_A are analogous to their crystallographic counterparts.

We compute σ_A for each resolution bin as in [56], using observed normalized amplitudes (E_{e}), calculated normalized amplitudes (E_{calc}), and the phase difference $\Delta\phi = \phi_{\text{calc}} - \phi_{\text{obs}}$.

Weighted Kabsch Alignment and Rigid Body Refinement

As stated above, we use the initial prediction from OpenFold to run Phaser (or EM_placement for cryo-EM) for molecular replacement. For every iteration, we align the OpenFold model to the reference molecular replacement solution before computing the relevant LLG score. This alignment is performed by first solving the following minimization problem with the Kabsch algorithm [68]:

$$L(C) = \frac{1}{2} \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i^{\text{ref}} - C\mathbf{x}_i^{\text{prediction}} \right\|^2 \tag{12}$$

where C denotes the rotation-translation matrix, $\mathbf{x}^{\text{ref}}$ and $\mathbf{x}_i^{\text{prediction}}$ are corresponding atomic coordinates of atom i in the reference and moving model, respectively, and w_i are positional weights. Only α-carbon (C_α) atoms are included in the alignment, and their weights are determined empirically from the pseudo-B values of the reference structure (the first, unconditioned OpenFold prediction). Specifically, for each residue:

$$w(\text{pseudo-B}) = \begin{cases} 1.0 & \text{pseudo-B} \leqslant 11.5 \\ 1.0 - 0.5 \times \frac{\text{pseudo-B} - 11.5}{40.0 - 11.5} & 40.0 \geqslant \text{pseudo-B} > 11.5 \\ 0.5 \exp(-\sqrt{\text{pseudo-B} - 40.0}) & \text{pseudo-B} > 40.0 \end{cases} \tag{13}$$

Translation vectors are determined by the vector difference of weighted center of mass of C_α atoms in the reference and moving models, then the rotation matrices are estimated with Kabsch algorithm. Once the alignment is completed, a subsequent rigid-body refinement is performed through gradient optimization of the LLG target:

$$C^* = \arg\max_C \text{LLG}(E_o, \mathbf{x}, C) \tag{14}$$

where $\mathbf{x}$ is the model coordinates after Kabsch alignment. When the predicted aligned error (PAE) matrix from AF2 can be used automatically by Phaser to split proteins into domains, ROCKET performs domain-specific alignments to make the best use of gradient information from the LLG. We used this, for example, for the refinement of the dataset related to PDB ID 7RAW in Fig. 3c.

Multi-Chain Dataset Handling

ROCKET can readily handle monomeric protein predictions in its current form. We have also demonstrated refinement of crystallographic datasets that contain two chains in the asymmetric unit for the kinase datasets (PDB ID 3PYY in Fig. 3 and PDB ID 7DT2 in Fig. S9). For these cases, we included the second chain present in the asymmetric unit as a fixed Fourier contribution in the likelihood calculation but excluded it from refinement. However, general multi-chain refinement would require integrating ROCKET with a multimeric model such as OpenFold-Multimer.

Given the current size limit of about 500 residues (on 40GB GPUs) for backpropagating through OpenFold, we also wanted to illustrate how larger cryo-EM complexes may be refined in parts. For the groEL example in Fig. 3c (PDB ID 9C0B), we only refined the domain of the groEL subunit that requires the most rebuilding. We did so by inserting a 9-glycine loop in the region connecting the two domains of the subunit, and treating the resulting single entity as the target for prediction.

Data Visualization

Visualization of PDB structures and experimental densities was done with PyMOL [69] and Moorhen [70].

Code Availability

Release of all ROCKET code and target functions is in preparation.

Acknowledgments

The following funding is acknowledged: Wellcome Trust (grant No. 209407/Z/17/Z to RJR); National Institutes of Health (NIH), National Institute of General Medical Sciences (P01-GM063210 to RJR and TCT; R35-GM150546 to MA; and DP2-GM141000 to DRH); Biotechnology and Biological Sciences Research Council (grant No. BB/Y009398/1 to RJR); a graduate fellowship, Eric and Wendy Schmidt Center to MHL. We would also like to thank K.M. Dalton, R. Gaudet, T.J. Lane, S. Eddy, and S. Petti for insight and discussions, and C. Floristean for help with the OpenFold codebase.

References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[2] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

[3] Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function, and Bioinformatics* **89**, 1711–1721 (2021).

[4] Pereira, J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics* **89**, 1687–1699 (2021).

[5] Karelina, M., Noh, J. J. & Dror, R. O. How accurately can one predict drug binding modes using AlphaFold models? *eLife* **12**, RP89386 (2023).

[6] Terwilliger, T. C. *et al.* AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods* **21**, 110–116 (2024).

[7] He, X.-h., Li, J.-r., Shen, S.-y. & Xu, H. E. AlphaFold3 versus experimental structures: assessment of the accuracy in ligand-bound G protein-coupled receptors. *Acta Pharmacologica Sinica* 1–12 (2024).

[8] Buttenschoen, M., M. Morris, G. & M. Deane, C. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science* **15**, 3130–3139 (2024).

[9] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

[10] Durant, G., Boyles, F., Birchall, K. & Deane, C. M. The future of machine learning for small-molecule drug discovery will be driven by data. *Nature Computational Science* **4**, 735–743 (2024).

[11] Frank, C. *et al.* Scalable protein design using optimization in a relaxed sequence space. *Science* **386**, 439–445 (2024).

[12] Stroud, R. M. & Fauman, E. B. Significance of structural changes in proteins: Expected errors in refined protein structures. *Protein Science* **4**, 2392–2404 (1995).

[13] Cruickshank, D. W. J. Remarks about protein structure precision. *Acta Crystallographica Section D: Biological Crystallography* **55**, 583–601 (1999).

[14] Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H. & Fraser, J. S. Ligand binding remodels protein side-chain conformational heterogeneity. *eLife* **11**, e74114 (2022).

[15] Brändén, G. & Neutze, R. Advances and challenges in time-resolved macromolecular crystallography. *Science* **373**, eaba0954 (2021).

[16] Hekstra, D. R. Emerging Time-Resolved X-Ray Diffraction Approaches for Protein Dynamics. *Annual Review of Biophysics* **52**, 255–274 (2023).

[17] Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS letters* **594**, 3243–3261 (2020).

[18] Nogales, E. & Mahamid, J. Bridging structural and cell biology with cryo-electron microscopy. *Nature* **628**, 47–56 (2024).

[19] McCoy, A. J., Sammito, M. D. & Read, R. J. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallographica Section D: Structural Biology* **78**, 1–13 (2022).

[20] Millán, C. *et al.* Assessing the utility of CASP14 models for molecular replacement. *Proteins: Structure, Function, and Bioinformatics* **89**, 1752–1769 (2021).

[21] Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology* **75**, 861–877 (2019).

[22] Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica Section D: Biological Crystallography* **53**, 240–255 (1997).

[23] Yamashita, K., Wojdyr, M., Long, F., Nicholls, R. A. & Murshudov, G. N. GEMMI and Servalcat restrain REFMAC5. *Acta Crystallographica Section D: Structural Biology* **79**, 368–373 (2023).

[24] Langer, G. G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols* **3**, 1171–1179 (2008).

[25] Terwilliger, T. C. Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. *Acta Crystallographica Section D: Biological Crystallography* **59**, 1174–1182 (2003).

[26] Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D: Biological Crystallography* **64**, 61–69 (2008).

[27] The wwPDB Consortium. EMDB—the Electron Microscopy Data Bank. *Nucleic Acids Research* **52**, D456–D465 (2024).

[28] Headd, J. J. *et al.* Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D: Biological Crystallography* **68**, 381–390 (2012).

[29] Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences* **96**, 5482–5485 (1999).

[30] Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics* **77**, 89–99 (2009).

[31] Stahl, K., Graziadei, A., Dau, T., Brock, O. & Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology* **41**, 1810–1819 (2023).

[32] Zhang, Y. *et al.* Distance-AF: Modifying Predicted Protein Structure Models by Alphafold2 with User-Specified Distance Constraints (2023).

[33] Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).

[34] Terwilliger, T. C. *et al.* Accelerating crystal structure determination with iterative AlphaFold prediction. *Acta Crystallographica Section D: Structural Biology* **79**, 234–244 (2023).

[35] Levy, A. *et al.* Solving Inverse Problems in Protein Space Using Diffusion-Based Priors (2024). ArXiv:2406.04239 [cs].

[36] Maddipatla, A. *et al.* Inverse problems with experiment-guided AlphaFold (2025). ArXiv:2502.09372 [q-bio].

[37] Ahdritz, G. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods* **21**, 1514–1524 (2024).

[38] Li, M., Dalton, K. & Hekstra, D. SFCalculator: connecting deep generative models and crystallography. *BioRxiv* 10.1101/2025.01.12.632630 (2025).

[39] Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).

[40] Bryant, P. & Noé, F. Improved protein complex prediction with AlphaFold-multimer by denoising the MSA profile. *PLOS Computational Biology* **20**, e1012253 (2024).

[41] Xie, T., Song, Z. & Huang, J. Conditioned Protein Structure Prediction. *PRX Life* **2**, 043001 (2024).

[42] Monteiro da Silva, G., Cui, J. Y., Dalgarno, D. C., Lisi, G. P. & Rubenstein, B. M. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature Communications* **15**, 2464 (2024).

[43] Stein, R. A. & Mchaourab, H. S. SPEACH_af: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS computational biology* **18**, e1010483 (2022).

[44] Budziszewski, G. R., Snell, M. E., Wright, T. R., Lynch, M. L. & Bowman, S. E. J. High-Throughput Screening to Obtain Crystal Hits for Protein Crystallography. *Journal of Visualized*

Experiments: JoVE **193**, e65211 (2023).

[45] Boby, M. L. *et al.* Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **382**, eabo7201 (2023).

[46] Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642–646 (2021).

[47] Fearon, D. *et al.* Accelerating Drug Discovery With High-Throughput Crystallographic Fragment Screening and Structural Enablement. *Applied Research* **4**, e202400192 (2025).

[48] Dalton, K. M., Greisman, J. & Hekstra, D. A unifying bayesian framework for merging x-ray diffraction data. *Nature Communications* 7764 (2022).

[49] Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods* **18**, 176–185 (2021).

[50] Kimanius, D. *et al.* Data-driven regularization lowers the size barrier of cryo-EM structure determination. *Nature Methods* **21**, 1216–1221 (2024).

[51] CryoREAD provides fully automated DNA–RNA structure modeling for cryo-EM maps. *Nature Methods* **20**, 1637–1638 (2023).

[52] Croll, T. I. *ISOLDE* : a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D Structural Biology* **74**, 519–530 (2018).

[53] Wang, X., Zhu, H., Terashi, G., Taluja, M. & Kihara, D. Diffmodeler: Large macromolecular structure modeling in low-resolution cryo-em maps using diffusion model. *bioRxiv* 10.1101/2024.01.20.576370 (2024).

[54] He, J., Lin, P., Chen, J., Cao, H. & Huang, S.-Y. Model building of protein complexes from intermediate-resolution cryo-EM maps with deep learning-guided automatic assembly. *Nature Communications* **13**, 4066 (2022).

[55] McCoy, A. J. *et al.* *Phaser* crystallographic software. *Journal of Applied Crystallography* **40**, 658–674 (2007).

[56] Read, R. J., Millán, C., McCoy, A. J. & Terwilliger, T. C. Likelihood-based signal and noise analysis for docking of models into cryo-EM maps. *Acta Crystallographica Section D* **79**, 271–280 (2023).

[57] Wang, W., Gong, Z. & Hendrickson, W. A. AlphaFold-guided molecular replacement for solving challenging crystal structures. *Acta Crystallographica Section D: Structural Biology* **81**, 4–21 (2025).

[58] van Beusekom, B. *et al.* Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Science* **27**, 798–808 (2018).

[59] Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

[60] Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. *BioRxiv* 10.1101/2021.10.11.463937 (2021).

[61] Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

[62] Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *BioRxiv* 10.1101/2021.10.04.463034 (2021).

[63] Zheng, H., Hou, J., Zimmerman, M. D., Wlodawer, A. & Minor, W. The future of crystallography in drug discovery. *Expert opinion on drug discovery* **9**, 125–137 (2014).

[64] McCoy, A. J. *et al.* *Phasertng*: directed acyclic graphs for crystallographic phasing. *Acta Crystallographica Section D* **77**, 1–10 (2021).

[65] Millán, C., McCoy, A. J., Terwilliger, T. C. & Read, R. J. Likelihood-based docking of models into cryo-EM maps. *Acta Crystallographica Section D* **79**, 281–289 (2023).

[66] Read, R. J. & McCoy, A. J. A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error. *Acta Crystallographica Section D: Structural Biology* **72**, 375–387 (2016).

[67] Fletcher, R. *Practical Methods of Optimization* 2nd edn (1987).

[68] Lawrence, J., Bernal, J. & Witzgall, C. A purely algebraic justification of the Kabsch-Umeyama algorithm. *Journal of research of the National Institute of Standards and Technology* **124**, 1 (2019).

[69] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8 (2015).

[70] Moorhen: A web browser molecular graphics program based on the coot desktop program (2025). URL https://github.com/moorhen-coot/Moorhen.

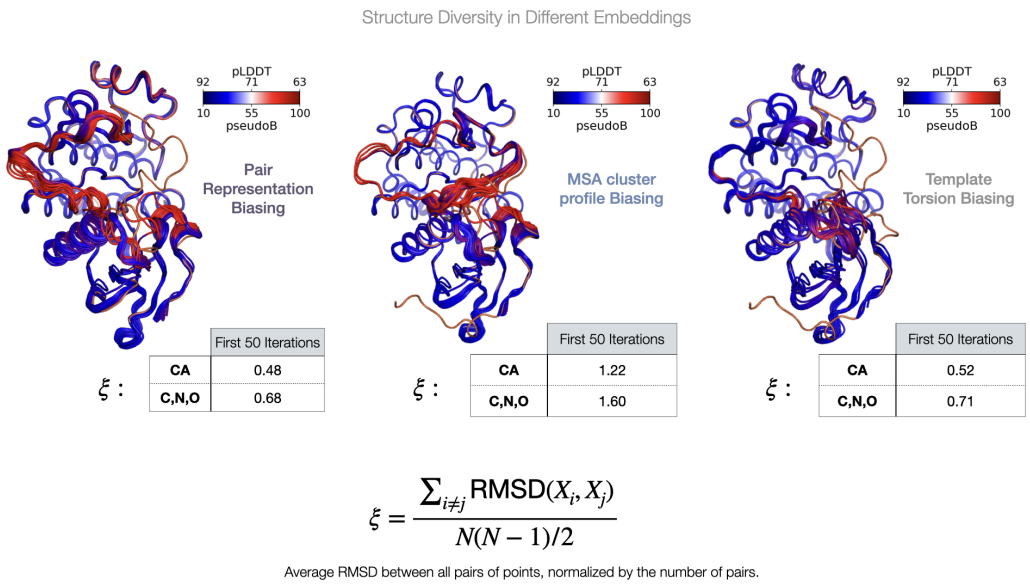
[71] Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293–315 (2018).

Supplementary Table

Code	n_res	TM_score	Spacegroup	Resolution (Å)	RMSD(AF2, deposited) (Å)
7E3Z	261	0.974	P 1 21 1	1.45	1.29
7VNX	223	0.996	P 61	1.80	0.75
7DNT	250	0.990	P 21 21 2	2.50	1.74
7O51	207	0.994	P 41 21 2	2.20	0.90
7TFQ	292	1.000	P 21 21 2	1.75	0.48
7RAW	251	0.975	I 21 3	2.10	0.89
7FIU	293	0.943	P 21 21 21	1.84	1.77
7SEZ	221	0.990	C 2 2 21	1.70	1.75
7RM7	228	0.938	P 21 21 21	1.03	0.54
7QDV	100	0.994	P 43 21 2	1.90	0.84
7TRW	217	0.994	P 62	2.28	1.96
7T26	144	0.999	P 41 21 2	1.14	0.67
7S3L	269	0.997	P 43 21 2	2.60	1.56
7DMS	94	0.424	P 21 21 21	1.96	0.56
7AOJ	179	0.910	P 61	1.63	1.60
7U2R	245	0.999	P 61 2 2	1.85	0.54
7KZH	202	0.979	P 3 2 1	2.49	0.62
7EJG	237	0.998	P 61	1.68	1.11
7T7Y	154	0.990	P 2 21 21	1.81	0.68
7EDC	249	0.989	C 1 2 1	1.95	1.46
7EYJ	95	0.999	P 62	1.38	0.64
7RPY	256	0.957	P 32 2 1	1.67	0.84
7ECD	272	0.997	I 41	2.60	1.31
7UNN	260	0.902	P 21 21 21	1.45	0.58
7TBS	238	0.945	I 2 2 2	1.96	0.62

Table 1 Table of high-resolution crystallographic benchmark dataset used for ROCKET. n_res indicates the number of residues.

Supplementary Figures



$$\xi = \frac{\sum_{i \neq j} \mathrm{RMSD}(X_i, X_j)}{N(N-1)/2}$$

Average RMSD between all pairs of points, normalized by the number of pairs.

Fig. S1 ROCKET Optimization in Different AlphaFold2 Embeddings. We performed data-guided inference-time optimization, as described in the main text, in alternative embeddings. We found that optimizing MSA cluster profiles provides the highest structural diversity along search trajectories. We calculated an average RMSD (ξ) between all pairs of structures, normalized by the number of pairs for the first 50 structures in a ROCKET phase 1 refinement run. Diversity is much higher when biasing the MSA cluster profile ($\xi = 1.22$ versus $\xi = 0.48$ and $\xi = 0.52$ when biasing the pair and template torsion representations, respectively).

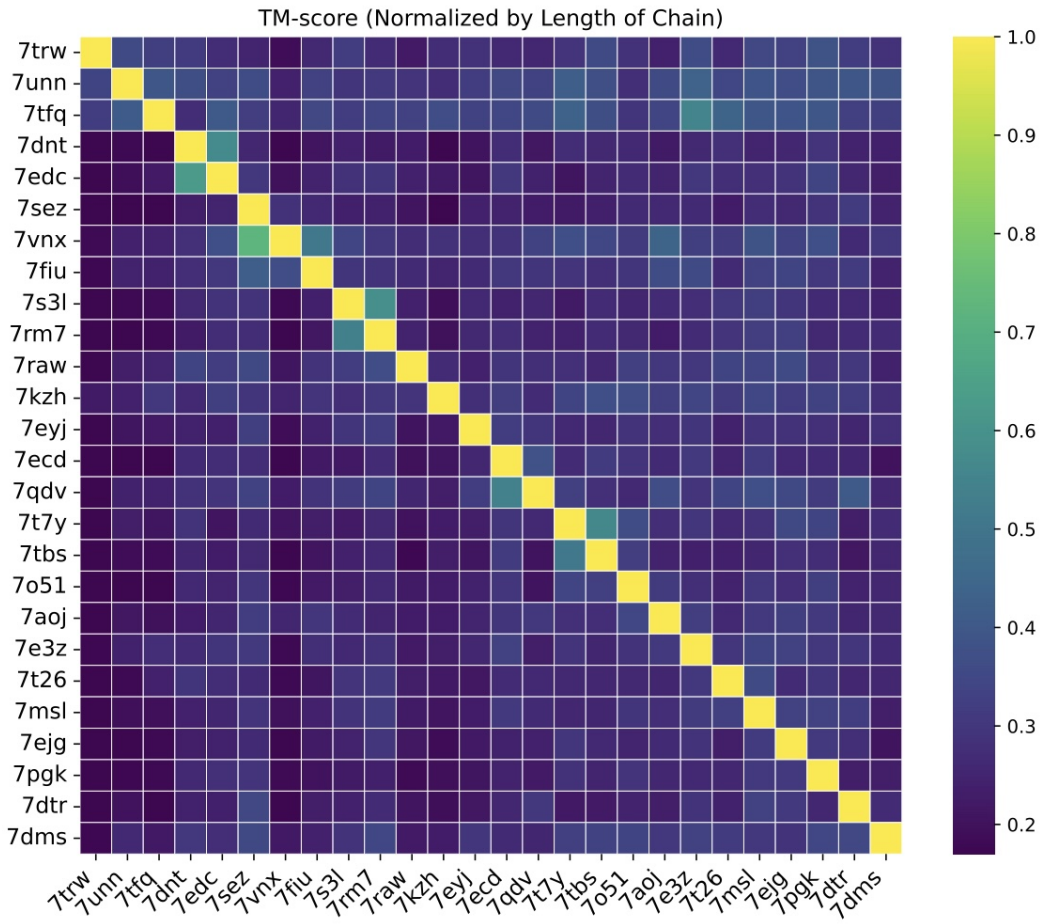


Fig. S2 Fold Diversity in High-Resolution Benchmark Dataset. Template Modeling (TM) scores for the 27 structures present in our crystallographic benchmark dataset, identified through their PDB ID. The TM-score evaluates how well two structures align in three-dimensional space, independent of sequence identity, with TM-score closer to 1 indicating very high similarity in the two folds. The low TM-scores between structures reflect their fold diversity. These structures were all released after the training of the AF2 weights used here and were solved by the single-wavelength anomalous diffraction (SAD) method, suggesting that it was challenging to find structural homologs in the PDB for molecular replacement. We therefore expect limited leakage from the original AF2 training set.

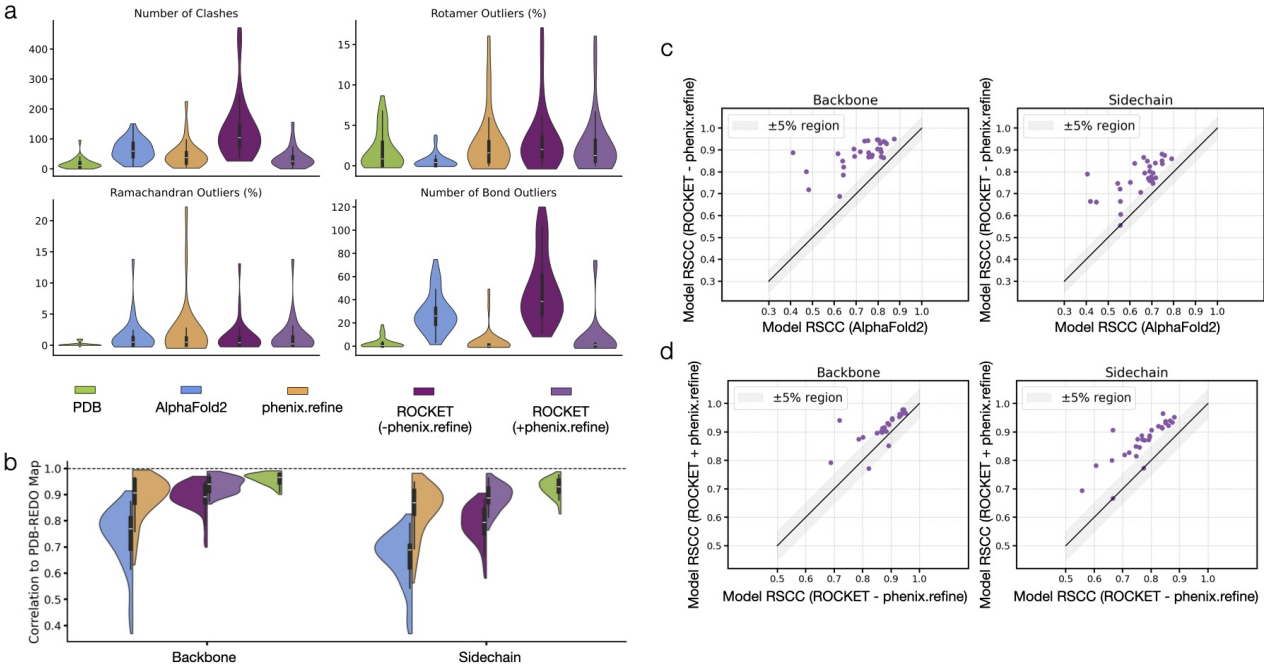


Fig. S3 Refinement Results for High-Resolution Crystallographic Datasets. (a) Geometric validation after refinement by different methods, carried out using MolProbity [71] for the 27 high-resolution crystallographic test cases. (b) Real-space Pearson correlation coefficient (RSCC) values for backbone and sidechains across initial AF2 models, phenix.refine, ROCKET refinements without or with the phenix.refine step. All residues were included, including those for which AF2 and ROCKET report low confidence. RSCC values for human-deposited models are also shown. Note that PDB-REDO maps are expected to have favorable phase bias for these models, which will artificially increase reported RSCC. (c-d) Breakdown of the incremental RSCC improvement from the initial AF2 prediction to the ROCKET structures (c), and before and after the phenix.refine step (d).

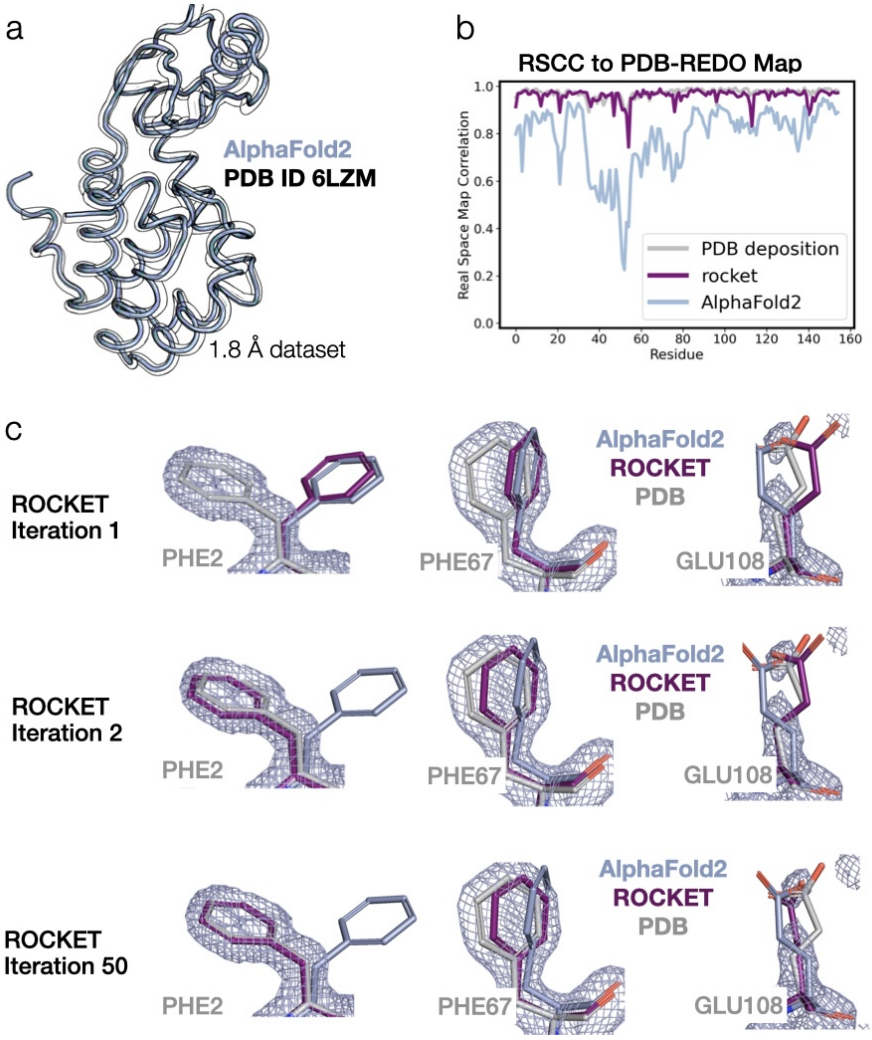


Fig. S4 MSA Cluster Profile Fine-Tuning Enables Both Backbone and Sidechain Refinement. Example of ROCKET refinement for a bacteriophage T4 lysozyme dataset (related PDB ID 6LZM), which differs from the AF2 prediction, particularly in the orientation of its upper domain (a). The PDB REDO structure is shown in transparent outlines. (b) ROCKET can refine the prediction to the same quality as the deposited structure. (c) When the data support a clear conformation (*e.g.*, for PHE2 and PHE67), ROCKET is able to find different rotameric configurations and gradually fine-tune the sidechain position. For GLU108, where the data are noisy, ROCKET explores different conformations and, in this case, settles for placing a carboxylate oxygen atom in the available density.

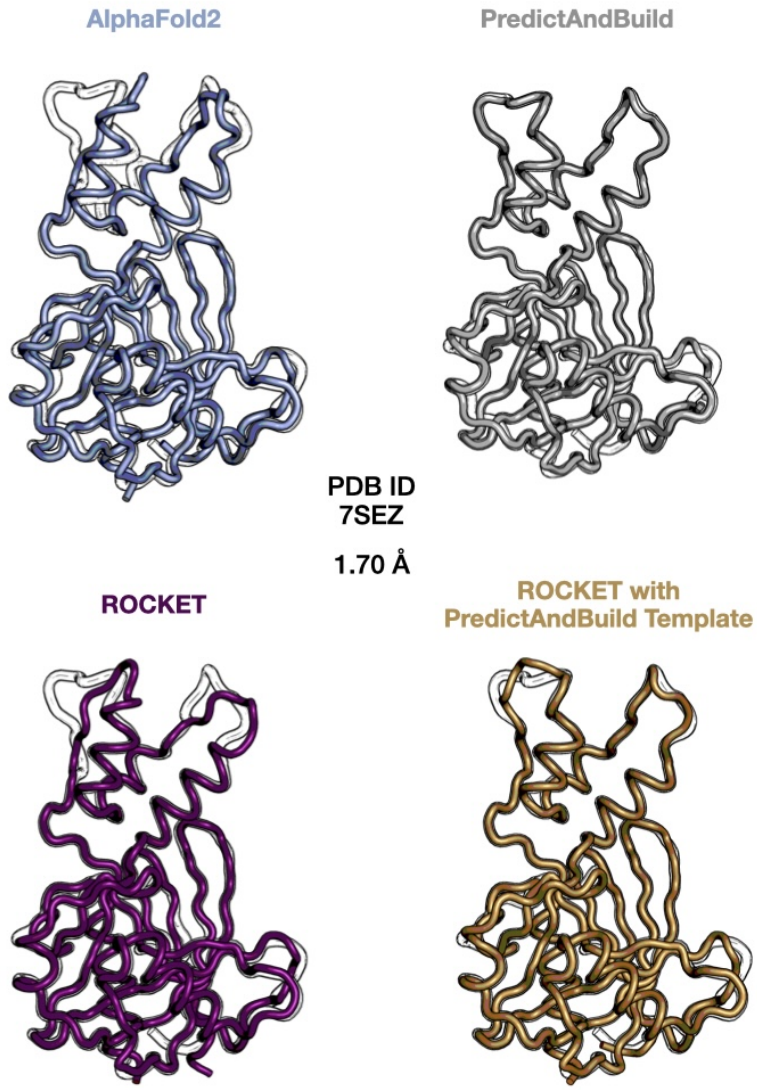


Fig. S5 Combining ROCKET with PredictAndBuild. We discuss in the main text and further in Fig. S9 that ROCKET can struggle to flip small loops containing long sidechain residues. The refinement of the Vaccinia Virus decapping enzyme D9 (PDB ID 7SEZ) is an example of this, where ROCKET is unsuccessful in modeling two small loops in the upper domain, while PredictAndBuild converges to the correct backbone. We show the complementarity of the two approaches by running ROCKET with the PredictAndBuild structure provided as a template during OpenFold inference and improving the final output.

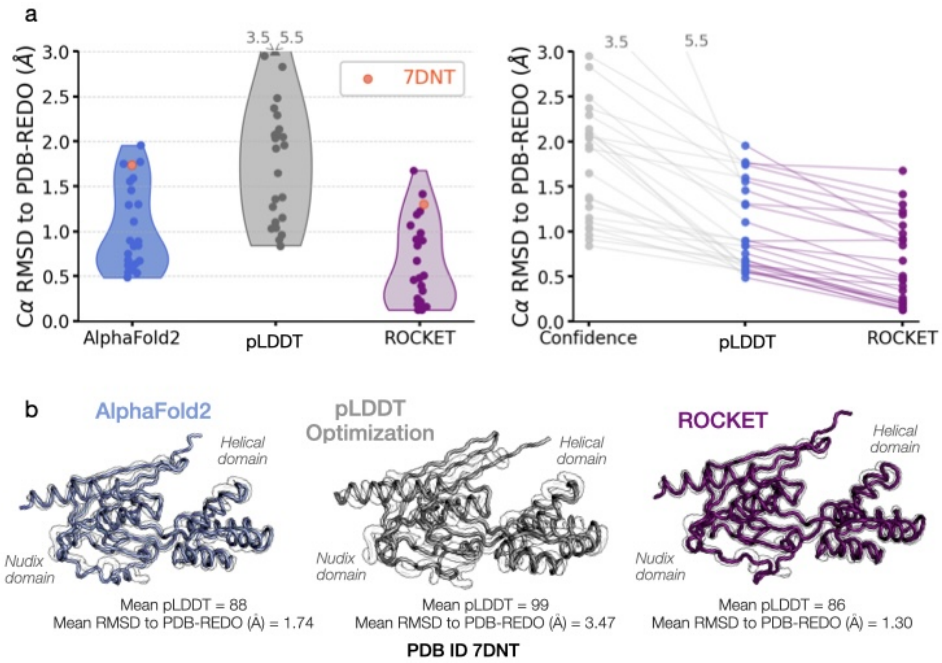


Fig. S6 Comparison with Confidence Maximization. To assess the added value of experimental data integration, we compare ROCKET's inference-time optimization to the results that can be obtained by maximizing AF2 model confidence. AF2 model confidence maximization has previously been used to explore alternate conformations for a given sequence, especially in the context of protein design. Depending on the task, different studies have optimized AF2's local confidence metric (predicted local-distance difference test – pLDDT) [11, 60, 61], its confidence of relative positioning between residues (predicted aligned error matrix – PAE matrix) [60], or interface confidence for complexes [11, 40]. We conducted a search in the MSA cluster profile space that maximizes pLDDT and found that AF2 does not easily produce experimentally observed conformations without further experimental information. (a) Cα RMSD values between PDB-REDO models and structures from standard AF2 inference, pLDDT maximization, and ROCKET's data-likelihood maximization. ROCKET consistently improves the match between AF2 predictions and experimental structures, while pLDDT maximization alone does not achieve comparable accuracy. (b) Example of the viral mRNA-decapping enzyme g5rp (PDB ID 7DNT) illustrating ROCKET's refinement capability. The initial AF2 prediction shows domain misalignment and secondary structure inaccuracies compared to the experimentally resolved conformation. ROCKET refines these discrepancies, while pLDDT maximization does not converge to the correct structure.

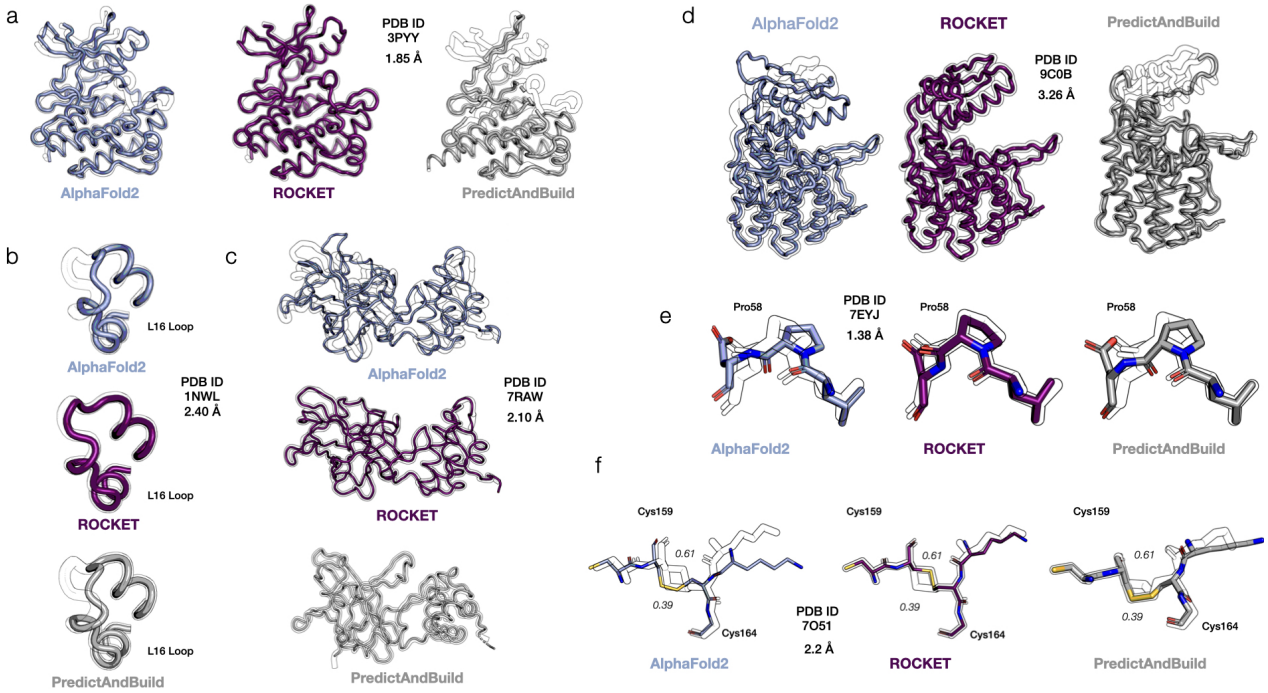


Fig. S7 ROCKET Overcomes Structural Barriers Inaccessible to PredictAndBuild. Showing the same examples as in Fig. 3 of the main text, but with the relevant PredictAndBuild refinements. PDB-REDO models are shown in transparent outlines. PredictAndBuild often chops low-confidence regions and flexible loops (see (a) and (d)), leaving them for manual rebuilding. ROCKET can handle these automatically. Two examples of bond rearrangements applied by ROCKET that PredictAndBuild does not access are also shown. The first is a peptide flip in the refinement of *E. coli* nucleoside phoshorylase (PDB ID 7YEJ, (e)). The second is a switch in a disulphide bond in the structure of thaumatin from *T. daniellii* (PDB ID 7AOJ, (f)). Two alternate conformations are present in the deposited structure, with refined occupancies of 0.31 and 0.69. ROCKET builds the conformation with highest occupancy. Both ROCKET and PredictAndBuild can tackle domain rearrangements (c) through separate alignment of the two domains to the data.

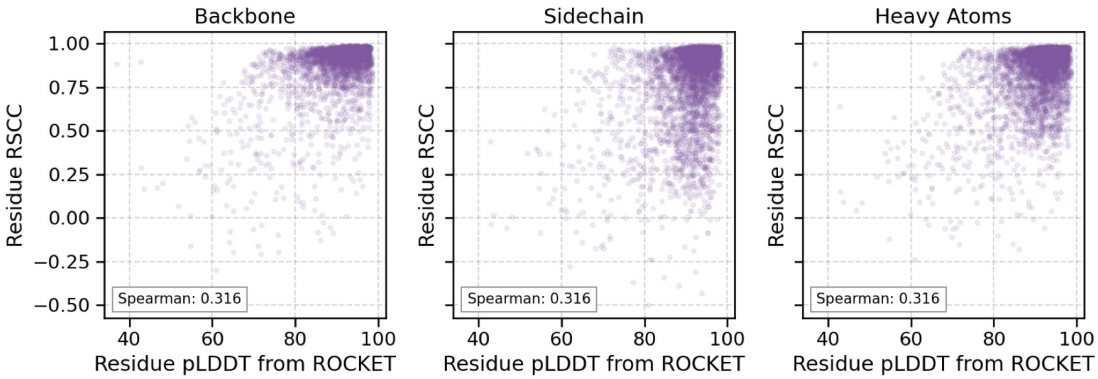


Fig. S8 Relationship between ROCKET pLDDT and Fit to Experimental Density. Residue pLDDT at the end of ROCKET refinement for all 27 high resolution test cases are plotted against real-space correlation coefficients (RSCC) between the map calculated from the ROCKET model and the PDB-REDO map. Results are broken down by different atom types.

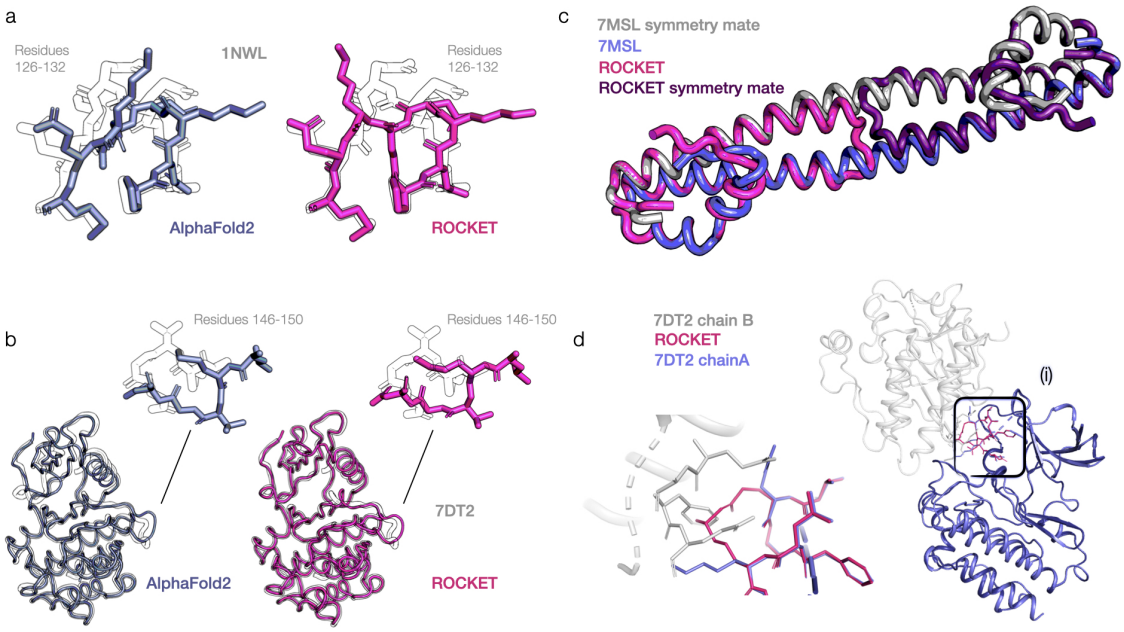


Fig. S9 Current Shortcomings in ROCKET Model Building. (a-b) We notice that ROCKET can fail to flip small loops (3-4 residues in length) that contain long sidechain residues. (c-d) Due to OpenFold's lack of awareness of crystal contacts, ROCKET may struggle to converge to certain lattice-dependent conformations (PDB ID 7SML, a dimer with crystallographic symmetry), or to account for the presence of another chain in the asymmetric unit (PDB ID 7DT2).

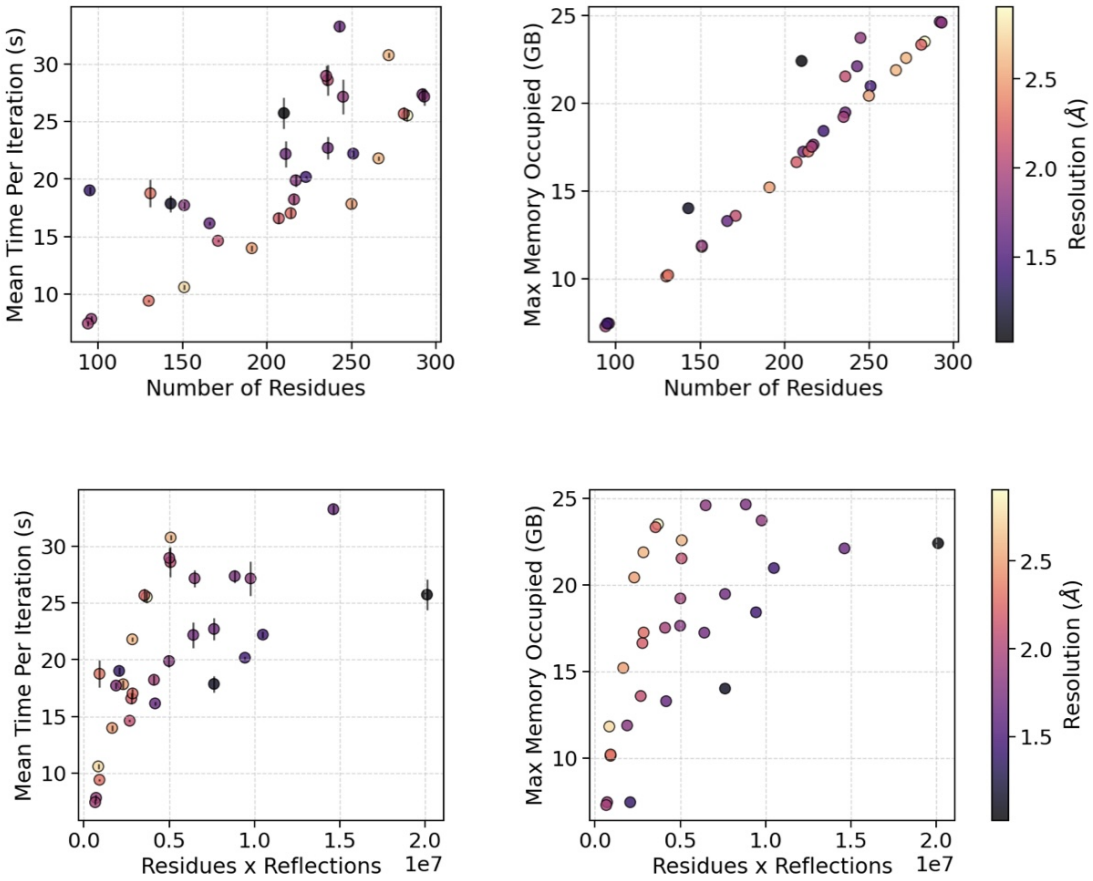


Fig. S10 Memory and Computation Time Requirements for ROCKET on a Nvidia 40 GB A100 GPU.

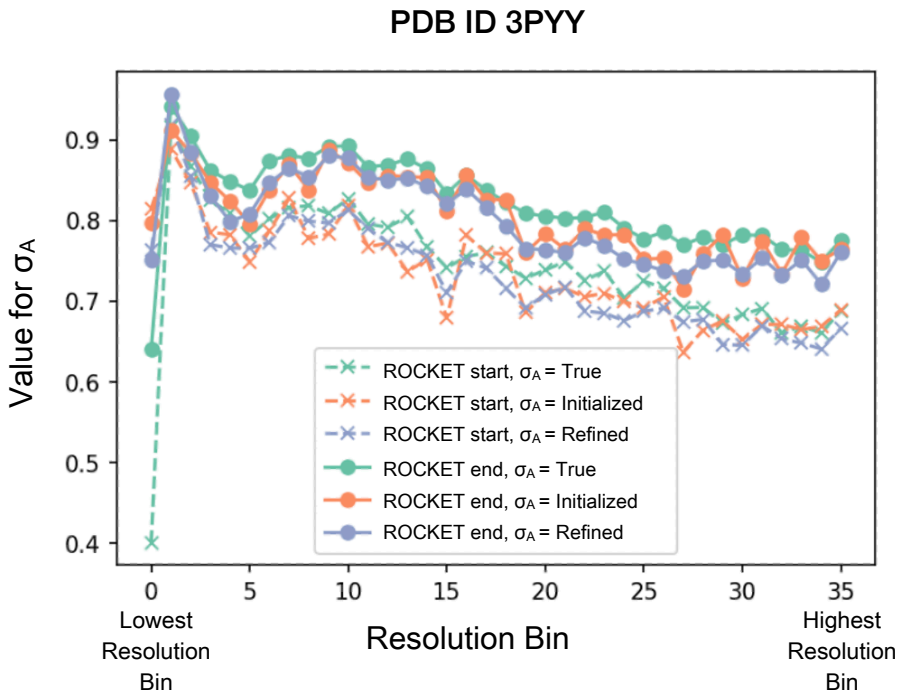


Fig. S11 Refinement of Crystallographic σ_A with the Working Set of Reflections. We find that refining σ_A values with ROCKET using the working set of reflections does not lead to meaningful overfitting. Here we plot σ_A values for each resolution bin for the starting (ROCKET start) and final (ROCKET end) iterations of ROCKET refinement for the PDB ID 3PYY dataset of the c-Abl kinase. In orange, we show the initialized values for the iteration and, in blue, the values after the σ_A refinement described in Methods. For comparison, we also plot, in green, the σ_A values that can be computed using the PDB REDO model phases and that we use as the basis for "true" σ_A comparison.