Non-uniform refinement: Adaptive regularization improves single particle cryo-EM reconstruction

Ali Punjani[1,2,3,*] Haowei Zhang[1] David J. Fleet[1,2,*]

[1]Department of Computer Sciences, University of Toronto, Toronto, Ontario, Canada
[2]Vector Institute, Toronto, Ontario, Canada
[3]Structura Biotechnology Inc., Toronto, Ontario, Canada
[*]Corresponding Authors
Correspondence to Ali Punjani (alipunjani@cs.toronto.edu) and David J Fleet (fleet@cs.toronto.edu)

Abstract

Cryo-EM is widely used to study biological macromolecules that comprise regions with disorder, flexibility, or partial occupancy. For example, membrane proteins are often kept in solution with detergent micelles and lipid nanodiscs that are locally disordered. Such spatial variability negatively impacts computational 3D reconstruction with existing iterative refinement algorithms that assume rigidity. We introduce *non-uniform refinement*, an algorithm based on cross-validation optimization, which automatically regularizes 3D density maps during refinement to account for spatial variability. Unlike common shift-invariant regularizers, *non-uniform refinement* systematically removes noise from disordered regions while retaining signal useful for aligning particle images, yielding dramatically improved resolution and 3D map quality in many cases. We obtain high resolution reconstructions for multiple membrane proteins as small as 100 kDa, demonstrating increased effectiveness of cryo-EM for this class of targets critical in structural biology and drug discovery. *Non-uniform refinement* is implemented in the *cryoSPARC* software package.

1 Introduction

Single particle cryo-EM has transformed rapidly into a mainstream technique in biological research [1]. Cryo-EM images individual protein particles, rather than crystals, and has therefore been particularly useful for structural studies of integral membrane proteins, which are difficult to crystallize [2]. These molecules are critical for drug discovery, targeted by more than half of drugs today [3]. Membrane proteins pose challenges in cryo-EM sample preparation, imaging, and computational 3D reconstruction, as they are often of small size, appear in multiple conformations, have flexible subunits, and are embedded in a detergent micelle or lipid nanodisc [2]. These characteristics cause strong spatial variation in structural properties, like rigidity and disorder, across the target molecule's 3D density. Traditional cryo-EM reconstruction algorithms, however, are based on the simplifying assumption of a uniform, rigid particle.

We develop an algorithm that incorporates such domain knowledge in a principled way, improving 3D reconstruction quality and allowing single particle cryo-EM to achieve higher-resolution structures of membrane proteins. This expands the range of proteins that can be effectively studied, and is especially important for structure-based drug design [4, 5]. We begin by formulating a cross-validation regularization framework for single particle cryo-EM refinement, and use it to account for the spatial variability in resolution and disorder found in a typical molecular complex. The framework incorporates general domain knowledge about protein molecules, without specific knowledge of any particular molecule, and critically, without need for manual user input. Through this framework we derive a new algorithm called *Non-Uniform Refinement* which automatically accounts for structural variability, while ensuring that key statistical properties for validation are maintained to mitigate the risk of over-fitting during 3D reconstruction.

With a GPU accelerated implementation of non-uniform refinement in the *cryoSPARC* software package [6], we demonstrate improvements in resolution and map quality for a range of membrane proteins. We show results on a 48 kDa membrane protein in lipid nanodisc with a Fab bound, a 180 kDa membrane protein complex with a large detergent micelle, and a 245 kDa sodium channel complex with flexible domains. Non-uniform refinement is reliable and automatic, requiring no change in parameters between datasets, without reliance on hand-made spatial masks or manual labels.

Iterative Refinement and Regularization

In standard cryo-EM 3D structure determination [7, 6, 8], a generative model describes the formation of 2D electron microscope images from a target 3D protein density (Coulomb potential). According to the model, the target density is rotated, translated, and projected along the direction of the electron beam. The 2D projection is modulated by a microscope contrast transfer function (CTF), and corrupted by additive noise. The goal of reconstruction is to infer the 3D density map from particle images, without knowledge of latent 3D pose variables, i.e., the orientation and position of the particle in each image. Iterative refinement methods formulate inference as a form of maximum likelihood or maximum *a posteriori* optimization (e.g., [9, 10, 6, 11]). Such algorithms can be viewed as a form of block-coordinate descent or Expectation-Maximization [12], each iteration comprising an E-step, estimating the pose of each particle image, given the 3D structure, and an M-step, regularized 3D density estimation given the latent poses.

Like many inverse problems with noisy, partial observations, the quality of cryo-EM map reconstruction depends heavily on regularization. Regularization methods, widely used in computer science and statistics, leverage prior domain knowledge to penalize unnecessary model complexity and avoid over-fitting. In cryo-EM refinement, regularization is needed to mitigate the effects of imaging and sample noise so that protein signal alone is present in the inferred 3D density, and so accumulated noise does not contaminate latent pose estimates.

Existing refinement algorithms use an explicit regularizer in the form of a shift-invariant linear filter, typically obtained from Fourier Shell Correlation (FSC) (e.g., [15, 10, 16, 6, 17, 18, 19]). Such filters smooth the 3D structure using the same kernel, and hence the same degree of smoothing, at all locations. Since FSC captures the average resolution of the map, such filters presumably under- *and* over-regularize different regions, allowing noise accumulation in some regions, and a loss of resolvable detail in others. This effect should be pronounced with membrane proteins that have highly non-uniform rigidity and disorder across the molecule. As a motivating example, Fig. 1 shows a reconstruction of the TRPA1 membrane protein [13] with a relatively low density threshold to help visualize regions of significant noise which indicate over-fitting (e.g., the disordered micelle

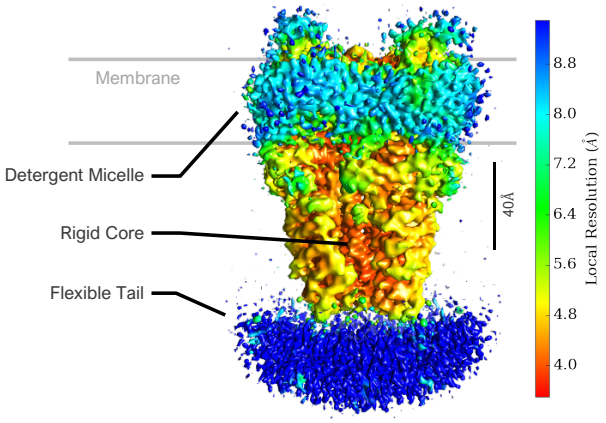


Figure 1: A 3D map from uniform refinement (in *cryoSPARC*) reveals spatial variations in resolution in a prototypical membrane protein (TRPA1 ion channel, EMPIAR-10024 [13]). Color depicts local resolution [14] as a proxy for local structure properties. Following the default FSC-based regularization and B-factor sharpening in cryoSPARC, the density map has been thresholded at a relatively low value to clearly visualize regions of significant noise. The core inner region (red) is more rigid and better resolved. The solvent facing region (yellow) is less well ordered. The detergent micelle (light blue) is largely disordered. A flexible tail at the bottom (blue) is blurred due to motion. In uniform refinement, a shift-invariant regularizer smooths all regions with the same kernel. This allows high frequency noise to accumulate in disordered regions (under-regularization), while discarding resolvable signal in rigid regions (over-regularization). Non-uniform refinement is designed to mitigate these problems.

and the flexible tail at and bottom of the protein). We hypothesize that under-fitting occurs in the core region where over-regularization attenuates useful signal. As such, accumulated noise and attenuated signal degrade pose estimates during refinement, limiting final structure quality. For inference problems of this type, the amount and form of regularization depends on *regularization parameters*. Correctly optimizing these parameters is often critical, but care must be taken to ensure that the optimization itself is not also prone to over-fitting.

2 Results

We next outline the formulation of an adaptive form of regularization, and with it, a new refinement algorithm called *Non-Uniform Refinement*. We discuss its properties and demonstrate its application on several membrane protein datasets.

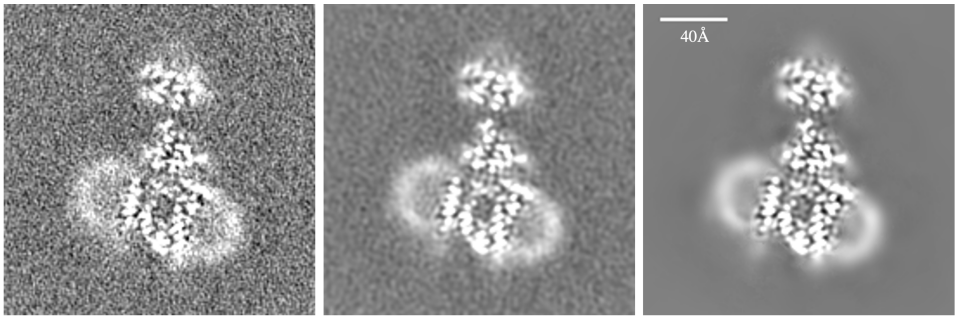
Adaptive Cross-Validation Regularization

With the aim of incorporating spatial non-uniformity into cryo-EM reconstruction, we formulate a family of regularizers denoted r_θ, with parameters $\theta(x)$ that depend on spatial position x. Given a 3D density map $m(x)$, the regularization operator, evaluated at x, is defined by

$$(r_\theta \circ m)(x) \;=\; \sum_{\xi} h(\xi; \theta(x))\; m(\xi - x)\;, \tag{1}$$

where $h(x; \psi)$ is symmetric smoothing kernel, the spatial scale of which is determined by parameter ψ.

This family provides greater flexibility than shift-invariant regularizers, but in exchange, requires making the correct choice of a new set of parameters, $\theta(x)$. We formulate the selection of the regularization parameters



Raw reconstruction | Shift-invariant regularization (FSC-based) | Adaptive regularization (Cross-validation optimal)

Figure 2: Images illustrate the difference between shift-invariant regularization and adaptive (cross-validation-optimal) regularization on a membrane protein (PfCRT, EMPIAR-10330 [23]). **Left:** A central slice through a raw reconstruction (M-step) of a half-map after 9 iterations. **Middle:** The same reconstruction is shown after uniform isotropic filtering (based on FSC between half-maps). **Right:** The same half-map reconstruction after adaptive regularization with the optimal CV regularizer. Adaptive regularization removes noise from the solvent background and nanodisc region, while preserving the high-resolution structure needed for particle alignments in the well-ordered protein region. It is particularly effective due to its implicitly small number of degrees of freedom and capacity to model sharp transitions between ordered and disordered regions.

as an optimization subproblem during refinement, for which we adopt a 2-fold cross-validation (CV) objective [20, 21]. The data are first randomly partitioned into two halves. On each of two trials, one part is considered training data and the other as held-out validation data. To find the regularizer parameters θ, we minimize the sum of the per-trial validation errors (e), measuring consistency between the model and the validation data; i.e.,

$$E(\theta) \;=\; e(r_\theta \circ m_1,\, m_2) + e(r_\theta \circ m_2,\, m_1) \tag{2}$$

where m_1 and m_2 are reconstructions from the two folds of the data. Similar objectives have been used for image de-noising [22]. We also introduce constraints on the parameters $\theta(x)$ to control degrees of freedom. The optimization problem is solved using a discretized search algorithm (see Methods).

The resulting cross-validation regularizer automatically identifies regions of a protein density with differing structural properties, optimally removing noise from each region. Fig. 2 illustrates the difference between uniform filtering (FSC-based) and the CV-optimal adaptive regularizer in non-uniform refinement. Shift-invariant regularization smooths all regions equally, while the adaptive regularizer preferentially removes noise from disordered regions, retaining high-resolution signal in well-structured regions that is essential for 2D-3D alignment.

Non-Uniform Refinement Algorithm

The non-uniform refinement algorithm takes as input a set of particle images and a low resolution *ab initio* 3D map. The data are randomly partitioned into two halves, each of which is used to independently estimate a 3D *half-map*. This "gold standard" refinement [19] allows use of FSC for evaluating map quality, and for comparison with existing algorithms. The key to non-uniform refinement is the adaptive cross-validation regularization, applied at each iteration of 3D refinement. The regularizer parameters $\theta(x)$ are estimated independently for each half-map (see Methods), adhering to the "gold standard" assumptions. In contrast, in conventional uniform refinement the two half-maps are not strictly independent since the regularization parameters, determined by FSC, are shared by both half-maps. Finally, the optimization and application of the adaptive regularizer cause

non-uniform refinement to be approximately two times slower than uniform refinement in cryoSPARC.

Validation with Membrane Protein Datasets

We experimentally compare non-uniform refinement against conventional uniform refinement with three membrane protein datasets, all processed in *cryoSPARC*. (See Supplementary Information for an additional membrane protein with no soluble region.) Both algorithms are given the same *ab-initio* structures. Except for the regularizer, all parameters and stages of data analysis, including the 2D-3D alignment and back-projection, are identical in uniform and non-uniform refinements. The default use of spatial solvent masking during 2D-3D alignment in *cryoSPARC* is used for both algorithms (see Supplementary Information). No manual masks are used, and no masking is used to identify or separate micelle or nanodisc regions. The same non-uniform refinement default parameters (other than symmetry) are used for all datasets.

When computing gold-standard FSC during the analysis of the reconstructed 3D density maps, for each dataset we use the same mask for uniform and non-uniform refinement. The masks are tested using phase randomization [15] to avoid FSC bias. Also, the same B-factor is used to sharpen both uniform and non-uniform refinement maps for each dataset. This consistency helps ensure that visible differences in 3D structure are due solely to algorithmic differences. To color maps using local resolution, we use a straightforward implementation of *Blocres* [14] for resolution estimation. No local filtering or local sharpening is used for visualization. Uniform and non-uniform refinement density maps are thresholded to contain the same total volume for visual comparison.

We also note that the FSC-based regularizer in conventional uniform refinement is equivalent to a shift-invariant regularizer optimized with cross-validation. That is, one can show that the optimal shift-invariant filter under cross-validation with squared error has a transfer function equivalent to the FSC curve. Thus, the experiments below capture the differences between adaptive and shift-invariant regularization.

STRA6-CaM: CV Regularization yields improved pose estimates and FSC

Zebrafish STRA6-CaM [24] is a 180 kDa C2 symmetric membrane protein complex comprising the 74 kDa STRA6 protein bound to calmodulin (CaM). STRA6 mediates the uptake of retinol in various tissues. We processed a dataset of 28,848 particle images of STRA6-CaM in a lipid nanodisc, courtesy of Oliver Clarke and Filippo Mancia [25]. Non-uniform refinement provides a substantial improvement in nominal resolution from $4.0\mathring{A}$ to $3.6\mathring{A}$ (Fig. 3A), indicating an improvement in the average signal-to-noise over the entire structure. However, different regions exhibit different resolution characteristics (Fig. 3C), as is often observed with protein reconstructions [14]. There is significant improvement in structural detail in most regions, while peripheral and flexible regions remain at low resolutions. Differences in structure quality, clearly visible in detail views of α-helices (Fig. 3D), are especially important during atomic model building, where in many cases, protein backbone and side-chains can only be traced with confidence in the non-uniform refinement map. Improvements in structure quality coincide with changes in particle alignments (Fig. 3B), approximately 3° on average. While disorder in the lipid nanodisc and non-rigidity of the CaM subunits are problematic for uniform refinement, adaptive regularization in non-uniform refinement reduces the influence of noise on alignments and produces improved map quality, especially in the periphery of the protein.

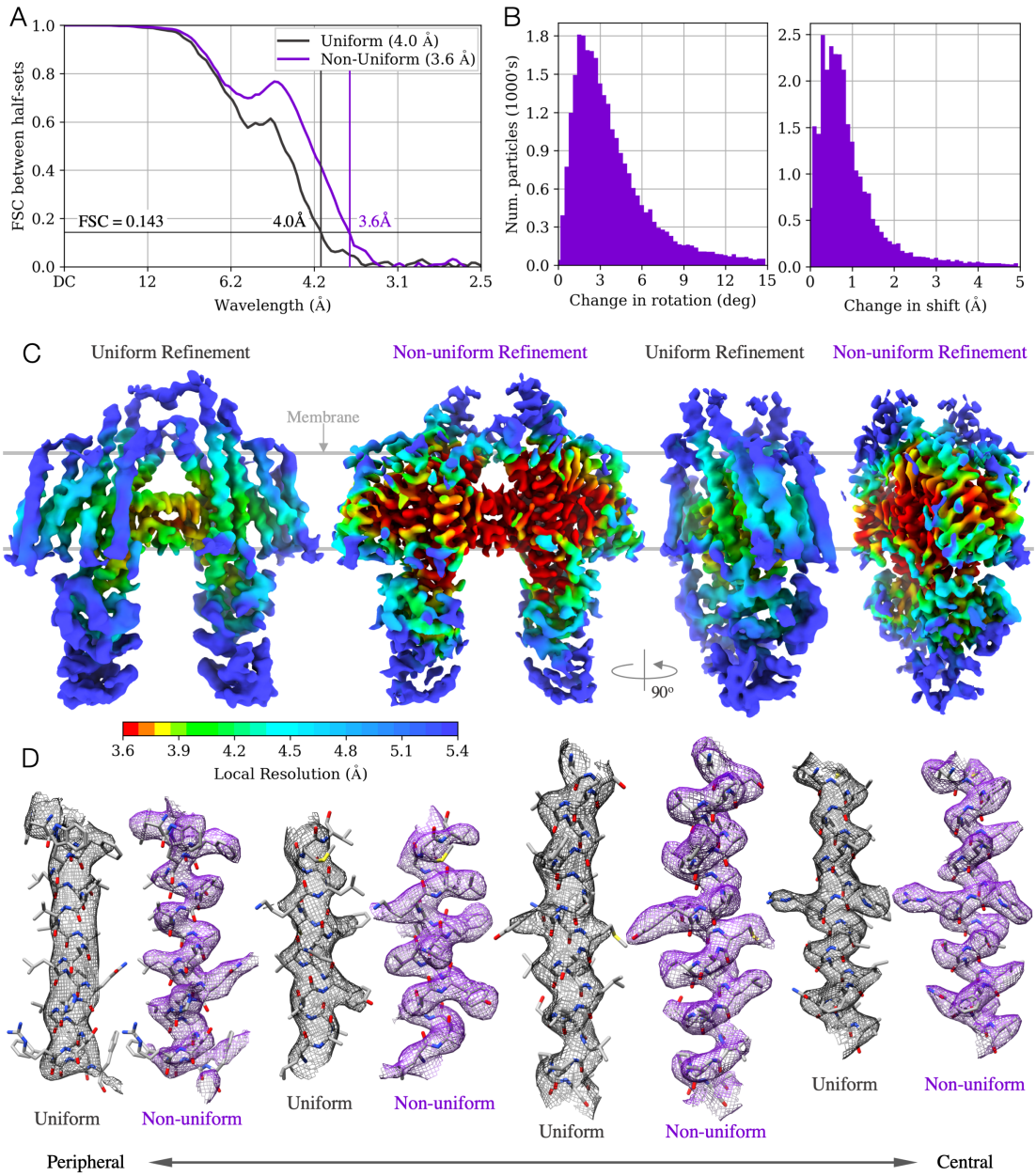


Figure 3: Results of uniform and non-uniform refinement from 28,848 particle images of **STRA6-CaM** [24] in lipid nanodisc. **A**: FSC curves computed with the same mask for both refinements. Numerical improvement from $4.0\mathring{A}$ to $3.6\mathring{A}$ reflects a spatially global improvement of signal. Improvement in some regions is even larger. **B**: Histograms of change in particle alignment pose and shift between uniform and non-uniform refinement. Optimal regularization in non-uniform refinement improves the ability to align particles over iterations. **C**: 3D density maps from uniform and non-uniform refinement are filtered using the corresponding FSC curves and sharpened with the same B-factor, $-140\mathring{A}^2$. No local filtering or sharpening is used, and thresholds are set to keep the enclosed volume constant. Map differences are due to algorithmic rather than visualization differences. Map color depicts local resolution (*Blocres* [14]), on a single color scale. **D**: Individual α-helical segments from the non-uniform map (purple) resolve backbone and side-chains while the uniform map (grey) does not in most cases. The left-most α-helix is peripheral while the right-most is central.

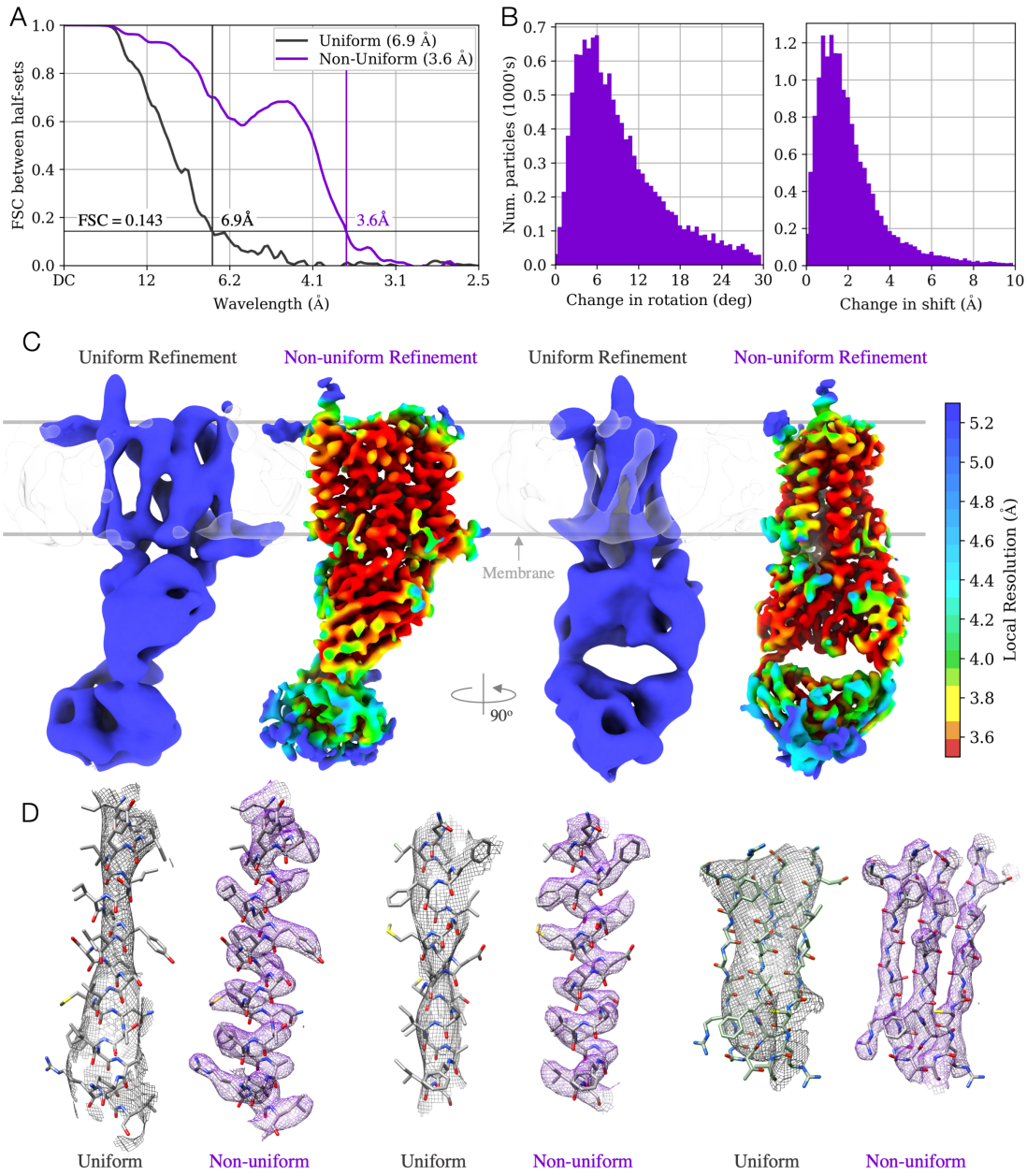


Figure 4: Results of uniform and non-uniform refinement from 16,905 particle images of **PfCRT** [23] in lipid nanodisc with a single Fab bound. **A**: FSC curves computed with the same mask show numerical improvement from $6.9\mathring{A}$ to $3.6\mathring{A}$, a dramatic global improvement of signal. **B**: Histograms of change in particle alignments between uniform and non-uniform refinement. Optimized adaptive regularization yields improved alignments through multiple iterations. Note that the x-axis limits are larger than in other figures. **C**: 3D density maps from uniform and non-uniform refinement, both filtered using the corresponding FSC curve, and sharpened with the same B-factor of $-100\mathring{A}^2$. No local filtering or sharpening is used, and thresholds are set to keep the enclosed volume constant. Density differences are thus due to algorithmic rather than visualization differences. Maps are colored by local resolution from *Blocres* [14], all on the same color scale. **D**: Individual α-helical segments from the non-uniform map (purple) resolve backbone and side-chains while α-helices are barely resolved in uniform (grey) map density and β-strands are not separated.

PfCRT: Enabling atomic model-building from previously unusable data

The *Plasmodium falciparum* chloroquine resistance transporter (PfCRT) is an asymmetric 48 kDa membrane protein [23]. Mutations in PfCRT are associated with the emergence of resistance to chloroquine and piperaquine as antimalarial treatments. We processed 16,905 particle images of PfCRT in lipid nanodisc with a Fab bound (EMPIAR-10330 [23]). For PfCRT, the difference in resolution (Fig. 4A) and map quality (Fig. 4C) between uniform and non-uniform refinement is striking.

Using uniform refinement, reaching $6.9\mathring{A}$, transmembrane α-helices are barely resolvable. Non-uniform refinement recovers signal up to $3.6\mathring{A}$ and provides a map from which an atomic model can be built with confidence. Transmembrane α-helices can be directly traced, including side-chains. In contrast, the uniform refinement map does not show helical pitch, and does not separate β-strands in the Fab domain. Indeed, an early version of the non-uniform refinement algorithm was essential for reconstructing the published high-resolution density map and model [23].

On the spectrum of proteins studied with cryo-EM, the PfCRT-Fab complex (100 kDa) is small. The lipid nanodisc ($\sim$50 kDa) also accounts for a large fraction of total particle molecular weight (see Fig. 2). We hypothesize that disorder in this relatively large nanodisc region leads to over- and under-regularization in uniform refinement. Most particle images exhibit orientation differences greater than 6° between the two algorithms (Fig. 4B), suggesting that a large fraction of particle images were grossly misaligned by uniform refinement.

$Na_v1.7$ Channel: Improvement in high-resolution features

$Na_V1.7$ is a voltage-gated sodium channel found in the human nervous system [26]. It plays a role in the generation and conduction of action potentials, and is targeted by toxins and therapeutic drugs (e.g., for pain relief). We processed data of $Na_V1.7$ bound to two Fabs, forming a 245 kDa C2-symmetric complex solublized in detergent [26]. Following pre-processing (see Methods), as in [26], we detected both active and inactive conformational states of the channel. We obtained reconstructions with resolutions better than the published literature for both, but here we focus on 300,759 particle images of the active state.

Compared to the preceding datasets, the $Na_V1.7$ complex has a higher molecular weight, and in relative terms, a smaller detergent micelle. However other regions are disordered or flexible, namely, a central 4-helix bundle, peripheral transmembrane domains, and the Fabs. For $Na_V1.7$, uniform refinement reaches $3.4\mathring{A}$ resolution and non-uniform refinement reaches an improved $3.1\mathring{A}$ (Fig. 5A). This result is also an improvement over the published result of $3.6\mathring{A}$ (EMDB-0341) [26]. where the authors performed all processing in *cisTEM* [10].

With non-uniform refinement, map quality is clearly improved in central transmembrane regions, while some flexible parts of the structure (Fab domains and 4-helix bundle) remain at intermediate resolutions (Fig. 5C). In detailed views of α-helices (Fig. 5D) closer to the periphery of the protein, improvement in map quality and interpretability is readily apparent in the non-uniform refinement map, allowing modelling of of side-chain rotamers with confidence. Central α-helices show less improvement, but map quality remains equal or slightly improved, indicating that reconstructions of protein regions without disorder are not harmed by using non-uniform refinement.

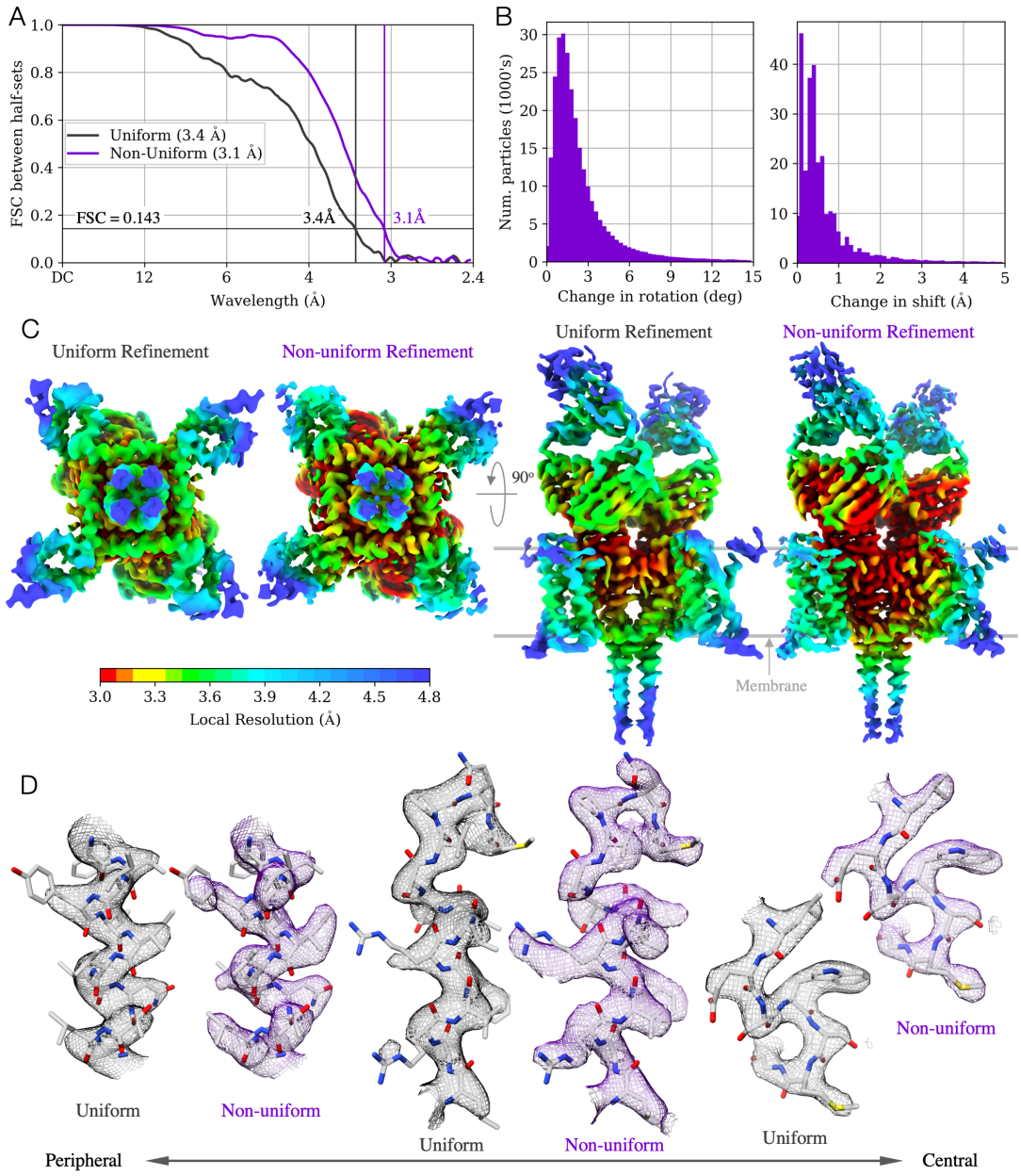


Figure 5: Results of uniform and non-uniform refinement on 300,759 particle images of **$Na_V1.7$** [26] in a detergent micelle with two Fabs bound. **A**: FSC curves indicate significant improvement in signal with non-uniform refinement, with numerical improvement from $3.4\mathring{A}$ to $3.1\mathring{A}$. **B**: Histograms of change in particle alignments between uniform and non-uniform refinement. Optimized adaptive regularization yields improved alignments through multiple iterations. **C**: The 3D density maps are filtered using corresponding FSC curves, sharpened with a B-factor of $-85\mathring{A}^2$. Thresholds are set to keep the enclosed volume constant. No local filtering or sharpening is used. Color depicts local resolution (*Blocres* [14]), on the same color scale. **D**: Individual α-helical segments from the non-uniform map (purple) are better resolved, allowing modelling of side-chain rotamers with confidence while uniform map (grey) often does not. Two α-helices (left, center), from the peripheral transmembrane domain, show larger improvement. The other (right), from the central core, is already well resolved by uniform refinement.

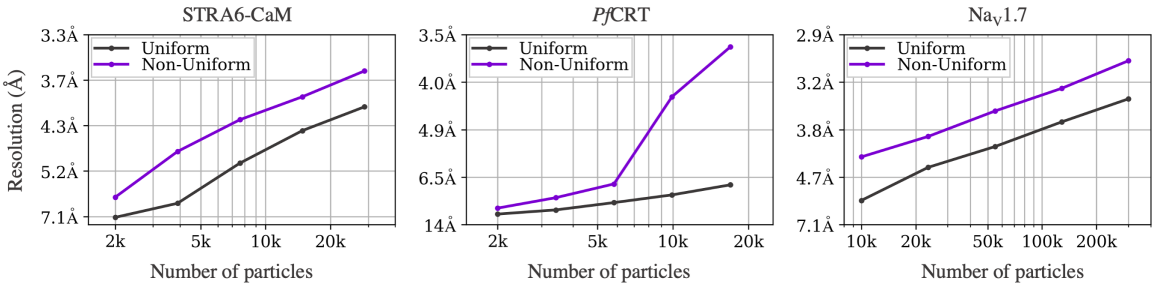


Figure 6: Plots of inverse resolution (frequency squared) versus number of particles on a log scale, for each of the STRA6-CaM [24], PfCRT [23] and Na$_v$1.7 [26] datasets. Each point represents an independent run of uniform (black) or non-uniform (purple) refinement, using a random sample from the entire dataset. For each dataset we use one mask for FSC at all dataset sizes. In all cases non-uniform refinement is more data efficient, reaching the same resolution with less than half the number of particles. The plots also show that non-uniform refinement resolution increases at an equal or greater rate with increasing particle numbers compared to uniform refinement. That is, the resolution gap persists over a wide range of dataset sizes.

Increased data efficiency

Examining map quality as function of dataset size further helps to explore data efficiency. Figure 6 plots inverse resolution versus the number of particles. Such plots are useful as they relate to image noise and computational extraction of signal [17]. Across all datasets, non-uniform refinement reaches the same resolution as uniform refinement with less than half the number of particle images. It has also been argued that higher curves indicate more accurate pose estimates [27]. Interestingly, the resolution gap between uniform and non-uniform refinement persists over a wide range of dataset sizes. While this resolution gap may decrease for much larger datasets than those considered here, the collection of more data alone may not allow uniform refinement to match the performance of non-uniform refinement for membrane protein targets, until resolution is saturated as one nears fundamental limits.

3 Discussion

With adaptive regularization and effective optimization of regularization parameters using cross-validation, non-uniform refinement achieves reconstructions of higher resolution and quality from single particle cryo-EM data. It is particularly effective for membrane proteins, which exhibit varying levels of disorder, flexibility, or occupancy. Here we focused on one specific family of adaptive regularizers (see Methods), but it is possible within the same framework to explore other families that may be even more effective. For example, one could look to the extensive denoising literature for different regularizers.

Spatial variability of structure properties and the existence of over- and under-regularization in uniform refinement are well-known. For instance, cryo-EM methods for estimating local map resolution once 3D refinement is complete leverage statistical tests on the coefficients of a windowed Fourier transform [14] or some form of wavelet transform [28, 29]. Although non-uniform refinement does not estimate resolution *per se*, the regularizer parameter $\theta(x)$ is related to a local frequency band-limit. As such it might be viewed as a proxy for local resolution, but with important differences. Notably, $\theta(x)$ is optimized with a cross-validation objective, and it does not depend on a specific definition of "local resolution" nor on an explicit resolution estimator.

Local resolution estimates [14, 30] or statistical tests [31] have also been used to adaptively filter 3D maps,

for visualization, to assess local map quality, or to aid molecular model building. The family of filters and frequency cut-offs are typically selected to maximize visual quality. They are not optimized for the local resolution estimator, nor is consideration given to the number of degrees of freedom in filter parameters, or the strict independence of half-maps, all critical for reliable regularization. Thus, while local resolution estimation followed by local filtering is useful for post-processing, its use for regularization during iterative refinement (e.g., see *EMAN2.2* documentation [9]), can yield over-fitting (see Methods). Map artifacts like spikes or streaks are especially problematic for datasets with junk particles, structured outliers, or small difficult to align particles. Non-uniform refinement couples an implicit resolution measure to the choice of regularizer, with optimization designed to control model capacity and avoid over-fitting of regularization parameters (see Methods).

Another related technique, used in *cisTEM* [10] and *Frealign* [7] and among the first to acknowledge and address under- and over-fitting, entails manual creation of a mask to label a local region one expects to be disordered (eg., detergent micelle), followed by low-pass filtering in that region to a pre-set resolution at each refinement iteration. While this shares the motivation for non-uniform refinement, it relies on manual interaction during refinement, often necessitating a tedious trial and error process that can be difficult to replicate.

SIDESPLITTER [32] is a recently proposed method to mitigate over-fitting during refinement. Based on estimates of signal-to-noise ratios from voxel statistics, it smooths local regions with low SNR more aggressively than indicated by a global FSC-based resolution. The method does not directly address under-fitting, nor does it maintain independence of half-maps, or control for the number of degrees of freedom in the local filter model. More generally, the cross-validation framework here is robust to different noise distributions and model misspecification, which can be problematic for methods with parametric noise models. Empirical results indicate that *SIDESPLITTER* mitigates overfitting, and shows improvement in map quality relative to uniform refinement.

Bases other than the Fourier basis may also enable local control of reconstruction. Wavelet bases have been used for local resolution estimation, but less commonly for reconstruction [33]. Kucukelbir et al [34] proposed the use of an adaptive wavelet basis and a sparsity prior. While similar in spirit to the goal of non-uniform refinement, this method has a single regularizer parameter for the entire 3D map, computed from noise in the corners of particle images. This may not capture variations in noise due to disorder, motion, or partial occupancy.

Non-uniform refinement has been successful in helping to resolve several new structures. Examples include multiple conformations of an ABC exporter [35], mTORC1 docked on the lysosome [36], the respiratory syncytial virus polymerase complex [37], a GPCR-G protein-β-arrestin megacomplex [38], and the SARS-CoV-2 spike protein [39].

Acknowledgements

We are extraordinarily grateful to Oliver Clarke, Filippo Mancia, and Yong Zi Tan for providing valuable cryo-EM data early in this project, and for sharing their experience as early adopters of non-uniform refinement and cryoSPARC. We thank the entire team at Structura Biotechnology Inc. that designs, develops, and maintains the cryoSPARC software system in which this project was implemented and tested. We thank John Rubinstein for comments on this manuscript. Resources used in this research were provided, in part, by the Province of Ontario, the Government of Canada through NSERC (Discovery Grant RGPIN 2015-05630 to DJF) and CIFAR, and companies sponsoring the Vector Institute.

Author Contributions

AP and DJF designed the algorithm, AP and HZ implemented the software and performed experimental work. AP and DJF contributed expertise and supervision. DJF and AP contributed to manuscript preparation.

Competing Financial Interests

AP is CEO of Stuctura Biotechnology Inc. which builds the cryoSPARC software package. DJF is an advisor to Stuctura Biotechnology Inc. Novel aspects of the method presented are described in a patent application (WO2019068201A1), with more details available at https://cryosparc.com/patent-faqs.

Correspondence to:

Ali Punjani (alipunjani@cs.toronto.edu) and David J Fleet (fleet@cs.toronto.edu)

References

[1] Y. Cheng, "Single-Particle Cryo-EM at Crystallographic Resolution," *Cell*, vol. 161, no. 3, pp. 450–457, 2015. 1

[2] Y. Cheng, "Membrane protein structural biology in the era of single particle cryo-EM," *Current Opinion in Structural Biology*, vol. 52, pp. 58–63, 2018. 1

[3] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?," *Nature Reviews Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006. 1

[4] J.-P. Renaud, A. Chari, C. Ciferri, W.-t. Liu, H.-W. Rémigy, H. Stark, and C. Wiesmann, "Cryo-EM in drug discovery: achievements, limitations and prospects," *Nature Reviews Drug Discovery*, vol. 17, no. 7, pp. 471–492, 2018. 2

[5] G. Scapin, C. S. Potter, and B. Carragher, "Cryo-EM for Small Molecules Discovery, Design, Understanding, and Application," *Cell Chemical Biology*, vol. 25, no. 11, pp. 1318–1325, 2018. 2

[6] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, "CryoSPARC: Algorithms for rapid unsupervised cryo-em structure determination," *Nature Methods*, vol. 14, pp. 290–296, 2017. 2, 16, 17, 20, 21

[7] N. Grigorieff, "FREEALIGN: High resolution refinement of single particle structures," *Journal of Structural Biology*, vol. 157, pp. 117–125, 2007. 2, 11, 16

[8] S. H. W. Scheres, "A Bayesian view on cryo-em structure determination," *Journal of Molecular Biology*, vol. 415, pp. 406–418, 2012. 2, 16

[9] J. M. Bell, M. Chen, P. R. Baldwin, and S. J. Ludtke, "High resolution single particle refinement in EMAN2.1," *Methods*, vol. 100, pp. 25–34, 2016. 2, 11, 16, 21

[10] T. Grant, A. Rohou, and N. Grigorieff, "*cis*TEM, user-friendly software for single-particle image processing," *eLife*, vol. 7, p. e35383, 2018. 2, 8, 11, 16

[11] J. Zivanov, T. Nakane, B. Forsberg, D. Kimanius, W. Hagen, and E. Lindahl, "New tools for automated high-resolution cryo-em structure determination in RELION-3," *eLife*, vol. 7, e42166, 2018. 2, 16

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977. 2, 16

[13] C. E. Paulsen, J. P. Armache, Y. Gao, Y. Cheng, and D. Julius, "Structure of the TRPA1 ion channel suggests regulatory mechanisms," *Nature*, vol. 520, pp. 511–517, 2015. 2, 3

[14] G. Cardone, J. B. Heymann, and A. C. Steven, "One number does not fit all: Mapping local variations in resolution in cryo-em reconstructions," *Journal of Structural Biology*, vol. 184, no. 2, pp. 226–236, 2013. 3, 5, 6, 7, 9, 10, 21

[15] S. Chen, G. McMullan, A. R. Faruqi, G. N. Murshudov, J. M. Short, S. H. W. Scheres, and R. Henderson, "High-resolution noise substitution to measure overfitting and validate resolution in 3d structure determination by single particle electron cryomicroscopy," *Ultramicroscopy*, vol. 135, pp. 24–35, 2013. 2, 5, 17

[16] G. Harauz and M. van Heel, "Exact filters for general geometry three dimensional reconstruction," *Optik*, vol. 73, pp. 146–156, 1986. 2, 17

[17] P. B. Rosenthal and R. Henderson, "Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy," *J. Molecular Biology*, vol. 333, no. 4, pp. 721–745, 2003. 2, 10, 17

[18] S. H. W. Scheres, "RELION: Implementation of a Bayesian approach to cryo-em structure determination," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519 – 530, 2012. 2, 16

[19] S. H. W. Scheres and S. Chen, "Prevention of overfitting in cryo-EM structure determination," *Nature methods*, vol. 9, pp. 853–854, sep 2012. 2, 4, 17, 18, 20

[20] G. M. Golub, M. Heath, and G. Wahba, "Generalized cross-validationas as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, p. 215–223, 1979. 4, 17

[21] G. Wahba, "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *The Annals of Statistics*, vol. 13, no. 4, pp. 1378–1402, 1985. 4, 17

[22] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," *Int. Conf. on Machine Learning*, 2018. 4, 18

[23] J. Kim, Y. Z. Tan, K. Wicht, *et al.*, "Structure and drug resistance of the plasmodium falciparum transporter pfcrt," *Nature*, Nov. 2019. 4, 7, 8, 10

[24] Y. Chen, O. B. Clarke, J. Kim, S. Stowe, Y.-K. Kim, Z. Assur, M. Cavalier, R. Godoy-Ruiz, D. C. von Alpen, C. Manzini, W. S. Blaner, J. Frank, L. Quadro, D. J. Weber, L. Shapiro, W. A. Hendrickson, and F. Mancia, "Structure of the stra6 receptor for retinol uptake," *Science*, vol. 353, p. aad8266, 2016. 5, 6, 10, 25

[25] O. Clarke. Personal Communication, 2018. 5

[26] H. Xu, T. Li, A. Rohou, C. P. Arthur, F. Tzakoniati, E. Wong, A. Estevez, C. Kugel, Y. Franke, J. Chen, C. Ciferri, D. H. Hackos, C. M. Koth, and J. Payandeh, "Structural basis of Nav1.7 inhibition by a gating-modifier spider toxin.," *Cell*, 2019. 8, 9, 10, 21

[27] S. Stagg, A. Noble, M. Spilman, and M. Chapman, "Reslog plots as an empirical metric of the quality of cryo-em reconstructions," *Journal of Structure Biology*, vol. 185, no. 3, pp. 418–426, 2014. 10

[28] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, "Quantifying the local resolution of cryo-em density maps," *Nature Methods*, vol. 11, pp. 63–65, 2014. 10

[29] J. L. Vilas, J. Gomez-Blanco, P. Conesa, R. Melero, J. M. de la Rosa-Trevín, J. Oton, J. Cuenca, R. Marabini, J. M. Carazo, J. Vargas, and C. O. S. Sorzano, "Monores: Automatic and accurate estimation of local resolution for electron microscopy maps," *Structure*, vol. 26, no. 2, pp. 337–344, 2018. 10

[30] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001. 10

[31] K. Ramlaul, C. M. Palmer, and C. H. S. Aylett, "A Local Agreement Filtering Algorithm for Transmission EM Reconstructions," *Journal of Structural Biology*, vol. 205, no. 1, pp. 30–40, 2019. 10

[32] K. Ramlaul, C. M. Palmer, and C. H. S. Aylett, "Mitigating Local Over-fitting During Single Particle Reconstruction with SIDESPLITTER," *bioRxiv*, 2020. 11

[33] C. Vonesch, L. Wang, Y. Shkolnisky, and A. Singer, "Fast wavelet-based single particle reconstruction in cryo-em," in *International Symposium on Biomedical Imaging*, pp. 1950–1953, 2011. 11

[34] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, "A Bayesian adaptive basis algorithm for single particle reconstruction," *Journal of Structural Biology*, vol. 179, pp. 56–67, 2012. 11

[35] S. Hofmann, D. Januliene, A. R. Mehdipour, C. Thomas, E. Stefan, S. Brüchert, B. T. Kuhn, E. R. Geertsma, G. Hummer, R. Tampé, and A. Moeller, "Conformation space of a heterodimeric ABC exporter under turnover conditions," *Nature*, vol. 571, no. 7766, pp. 580–583, 2019. 11

[36] K. B. Rogala, X. Gu, J. F. Kedir, M. Abu-Remaileh, L. F. Bianchi1, A. M. Bottino1, R. Dueholm1, A. Niehaus1, D. Overwijn1, A. C. Priso Fils1, S. X. Zhou1, D. Leary, N. N. Laqtom1, E. J. Brignole, and D. M. Sabatini, "Structural basis for the docking of mTORC1 on the lysosomal surface," *Science*, vol. 366, pp. 468–475, oct 2019. 11

[37] M. S. A. Gilman, C. Liu, A. Fung, I. Behera, P. Jordan, P. Rigaux, N. Ysebaert, S. Tcherniuk, J. Sourimant, J.-F. Eléouët, P. Sutto-Ortiz, E. Decroly, D. Roymans, Z. Jin, and J. S. McLellan, "Structure of the Respiratory Syncytial Virus Polymerase Complex," *Cell*, vol. 179, pp. 193–204.e14, sep 2019. 11

[38] A. H. Nguyen, A. R. B. Thomsen, T. J. Cahill, R. Huang, L.-Y. Huang, T. Marcink, O. B. Clarke, S. Heissel, A. Masoudi, D. Ben-Hail, F. Samaan, V. P. Dandey, Y. Z. Tan, C. Hong, J. P. Mahoney, S. Triest, J. Little, X. Chen, R. Sunahara, J. Steyaert, H. Molina, Z. Yu, A. des Georges, and R. J. Lefkowitz, "Structure of an endosomal signaling GPCR–G protein–β-arrestin megacomplex," *Nature Structural & Molecular Biology*, vol. 26, no. 12, pp. 1123–1131, 2019. 11

[39] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, and C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation," *Science*, vol. 367, pp. 1260–1263, Feb 2020. 11

Methods

Regularization in Iterative Refinement

In the standard cryo-EM 3D reconstruction problem set up (Section 2), the target 3D density map m is typically parameterized as a real-space 3D array with density at each voxel, in a Cartesian grid of box size N, and a corresponding discrete Fourier representation, $\hat{m} = Fm$. The goal of reconstruction is to infer the 3D densities of the voxel array, called model parameters. Representing the 3D density, 2D projections, observed images, and the noise model in the Fourier domain is common practice for computational efficiency, exploiting the well-known convolution and Fourier-slice theorems [9, 7, 6, 8]. The unobserved pose variables for each image are latent variables.

Iterative refinement methods (Algorithm 1), which provide state of the art results (e.g., [9, 10, 6, 11]), can be interpreted as variants of block-coordinate descent or the expectation-maximization algorithm [12]. In cryo-EM, and more generally in inverse problems with noisy, partial observations, a critical component that modulates the quality of the results is regularization.

One can regularize problems explicitly, using a prior distribution over model parameters, or implicitly, by applying a regularization operator to the model parameters during optimization. Iterative refinement methods tend to use implicit regularizers, attenuating noise in the reconstructed map at each iteration. In either case, the separation of signal from noise is the crux of many inference problems.

In the cryo-EM refinement problem, like many latent variable inverse problems, there is an additional interplay between regularization, noise buildup, and the estimation of latent variables. Retained noise due to under-regularization will contaminate the estimation of latent variables. This contamination is propagated to subsequent iterations and causes over-fitting.

This paper reconsiders the task of regularization based on the observation that common iterative refinement algorithms often systematically under-fit and over-fit different regions of a 3D structure, simultaneously. This causes a loss of resolvable detail in some parts of a structure, and the accumulation of noise in others. The reason stems from the use of frequency-space filtering as a form of regularization. Some programs, like *cisTEM* [10], use a strict resolution cutoff, beyond which Fourier amplitudes are set to zero before alignment of particle images to the current 3D structure. In *RELION* [18], regularization was initially formulated with an explicit

Algorithm 1 Iterative Refinement (Expectation-Maximization)

Require: Particle image dataset $\mathcal{D}$ and *ab initio* 3D map

1: Use smoothed *ab initio* 3D map array as the initial model parameters $m^{(0)}$
2: **while** not converged **do**
3: **E-step**: Given current estimate of model parameters, $m^{(t-1)}$, from step $t-1$, estimate (via marginalization or maximization) the latent variables: $z^{(t)} \leftarrow f(m^{(t-1)}, \mathcal{D})$
4: **M-step**: Given the latent variables $z^{(t)}$, compute raw estimates of the model parameter $\tilde{m}^{(t)}$ (without regularization): $\tilde{m}^{(t)} \leftarrow h(z^{(t)}, \mathcal{D})$
5: **Regularize**: Given noisy model parameters $\tilde{m}^{(t)}$, apply the regularization operator, r_θ, with regularization parameters θ: $m^{(t)} \leftarrow r_\theta(\tilde{m}^{(t)})$
6: **end while**

Gaussian prior on Fourier amplitudes of the 3D structure, with a hand-tuned parameter that controls Fourier amplitude shrinkage. Later versions of *RELION* [19] and *cryoSPARC*'s homogeneous (uniform) refinement [6] use a transfer function (or Wiener filter) determined by Fourier Shell Correlation (FSC) computed between two half-maps (e.g., [15, 16, 17]).

Such methods presume a Fourier basis and shift-invariance. Although well-suited to stationary processes, they are less well suited to protein structures, which are spatially compact and exhibit non-uniform disorder, e.g., due to motion or variations in signal resolution. FSC, for instance, provides an aggregate measure of resolution. To the extent that FSC under- or over-estimates resolution in different regions, FSC-based shift-invariant filtering will over-smooth some regions, attenuating useful signal, and under smooth others, leaving unwanted noise. To address these issues we introduce a family of adaptive regularizers that can, in many cases, find better 3D structures with improved estimates of the latent poses during refinement.

Cross-Validation Regularization

We formulate a new regularizer for cryo-EM reconstruction in terms of the minimization of a cross-validation objective [20, 21]. Cross-validation (CV) is a general principle that is widely used in machine learning and statistics for model selection and parameter estimation with complex models. In CV, observed data are randomly partitioned into a training set and a held-out validation set. Model parameters are inferred using the training data, the quality of which is then assessed by measuring an error function applied to the validation data. In k-fold CV, the observations are partitioned into k parts. In each of k trials, one part is selected as the held-out validation set, and the remaining $k-1$ parts comprise the training set. The per-trial validation errors are summed, providing the total CV error. This procedure measures agreement between the optimized model and the observations, without bias due to over-fitting. Rather, over-fitting during training is detected directly as an increase in the validation error. Importantly, formulating regularization in a cross-validation setting provides a principled way to design regularization operators that are more complex than the conventional, isotropic frequency-space filters. The CV framework is not restricted to a Fourier basis. One may consider more complex parameterizations, the use of meta-parameters, and incorporate cryo-EM domain knowledge.

Given a family of regularizers r_θ with parameters θ, the minimization of CV error to find θ is often applied as an *outer loop*. This requires the optimization of model parameters m to be repeated many times with different values of θ, a prohibitively expensive cost for problems like cryo-EM. Instead, one can also perform CV optimization as an *inner loop*, while optimization of model parameters occurs in the outer loop. Regularizer parameters θ are then effectively optimized on-the-fly, preventing under- or over-fitting without requiring multiple 3D refinements to be completed.

To that end, consider the use of 2-fold CV optimization to select the regularization operator, denoted $r_\theta(m)$ in the regularization step in Algorithm 1 (note that $k > 2$ is also possible). The dataset $\mathcal{D}$ is partitioned into two halves, $\mathcal{D}_1$ and $\mathcal{D}_2$, and two (unregularized) refinements are computed, namely m_1 and m_2. For each, one half of the data is the 'training set', and the other is held out for validation. To find the regularizer parameters θ we wish to minimize the total CV error E, i.e.,

$$\begin{aligned} \min_\theta E(\theta) &= \min_\theta \, e(r_\theta(m_1); \mathcal{D}_2) + e(r_\theta(m_2); \mathcal{D}_1) \\ &= \min_\theta \, \|r_\theta(m_1) - m_2\|^2 + \|r_\theta(m_2) - m_1)\|^2 \end{aligned} \tag{3}$$

where e is the negative log likelihood of the validation half-set given the regularized half-map. The second line

simplifies this expression by using the raw reconstruction from the opposite half-set as a proxy for the actual observed images. Note that assumptions for "gold standard" refinement [19] are not broken in this procedure (see Methods: Non-Uniform Refinement Algorithm). With the L2 norm, Eqn. 3 reduces to a sum of per-voxel squared errors, corresponding to white Gaussian noise between the half-set reconstructions. When the choice of θ causes r_θ to remove too little noise from the raw reconstruction, the residual error E will be unnecessarily large. If θ causes r_θ to over-regularize, removing too much structure from the raw reconstruction, then E increases as the structure retained by r_θ no longer cancels corresponding structure in the opposite half-map. As such, minimizing $E(\theta)$ provides the regularizer that optimally separates signal from noise.

We note that similar objectives have been used for general image de-noising [22] and adapted for cryo-EM images [40, 41, 42], however in these methods the aim was to learn a general neural network denoiser, while the goal here is to optimize regularization parameters on a single data sample. It is also worth noting this formulation can be extended in a straightforward way to compare each half-set reconstruction against images directly (dealing appropriately with the latent pose variables) or to use error functions corresponding to different noise models.

Regularization Parameter Optimization

The CV formulation in Eqn. 3 provides great flexibility in choosing the family of regularizers r_θ, taking domain knowledge into account. For non-uniform refinement, we wish to accommodate protein structures with spatial variations in disorder, motion, and resolution. Accordingly, we define the regularizer to be a space-varying linear filter. The filter's spatial extent is determined by the regularization parameter $\theta(x)$, which varies with spatial position. Here, we write the regularizer in operator form:

$$(r_\theta \circ m)(x) \;=\; \sum_{\xi} h(\xi; \theta(x))\; m(\xi - x)\;, \tag{4}$$

where $h(x;\psi)$ is symmetric smoothing kernel, the spatial scale of which is specified by ψ. In practice we let $h(x;\psi)$ be a 8^{th}-order Butterworth kernel. The 8^{th}-order kernel provides a middle ground between the relatively poor frequency resolution of the Gaussian kernel, and the sharp cut-off of the sinc kernel, which suffers from spatial ringing. We have experimented with different orders of Butterworth filters, and found that an 8^{th}-order kernel performs well on a broad range of particles.

When Eqn. 4 is combined with the CV objective for the estimation of $\theta(x)$, one obtains

$$\begin{aligned} \theta^* \;&=\; \arg\min_\theta E(\theta) \\ &=\; \arg\min_\theta \sum_x |\,(r_\theta \circ m_1)(x) - m_2(x)\,|^2 \;+\; |\,(r_\theta \circ m_2)(x) - m_1(x)\,|^2\;. \end{aligned} \tag{5}$$

With one regularization parameter at each voxel, i.e., $\theta(x)$, this reduces to a large set of decoupled optimization problems, one for each voxel. That is, for voxel x one obtains

$$\theta^*(x) \;=\; \arg\min_{\theta(x)} |\,(r_\theta \circ m_1)(x) - m_2(x)\,|^2 \;+\; |\,(r_\theta \circ m_2)(x) - m_1(x)\,|^2 \tag{6}$$

With this decoupling, $\theta(x)$ can transition quickly from voxel to voxel, yielding high spatial resolution. On the other hand, the individual sub-problems (6) are not well constrained since each parameter is estimated from data at a single location, so the parameter estimates are not useful. In essence, our regularizer design has two

competing goals, namely, reliable signal detection, and high spatial resolution (i.e. respecting boundaries between regions with different properties). Signal detection improves through aggregation of observations (e.g., neighboring voxels), while high spatial resolution prefers minimal aggregation (as in Eqn. 6).

To improve signal detection, we further constrain θ^* to be smooth. That is, although in some regions θ should change quickly (solvent-protein boundaries), in most regions we expect it to change slowly (solvent and regions of rigid protein mass). Smoothness effectively limits the number of degrees of freedom in θ, which is important to ensure θ itself does not overfit during iterative refinement (see Sec. 3). One can encourage smoothness in θ by explicitly penalizing spatial derivatives of θ in the objective (Eqn. 5), but this yields a Markov random field problem that is hard to optimize. Alternatively, one can express θ in a low-dimensional basis (e.g., radial basis functions), but this requires prior knowledge of the expected degree of smoothness. Instead, we adopt a simple but effective approach. Assuming that θ is smoothly varying, we treat measurements in the local neighborhood of x as additional constraints on $\theta(x)$. A window function can be used to give more weight to points close to x. We thereby obtain the following least-squares objective:

$$\min_{\theta(x)} \sum_{\xi} w_\rho(\xi - x) \left[\, | \, (r_{\theta(x)} \circ m_1)(\xi) - m_2(\xi) \, |^2 + | \, (r_{\theta(x)} \circ m_2)(\xi) - m_1(\xi) \, |^2 \, \right] \tag{7}$$

where $w_\rho(x)$ is positive and integrates to one, with spatial extent ρ. This allows one to estimate θ at each voxel independently, while the overlapping neighborhoods ensure that $\theta(x)$ varies smoothly.

This approach also provides a natural way to allow for variable neighbourhood sizes, where $\rho(x)$ depends on location x, so both rigid regions and transition regions are well modeled. Importantly, we want $\rho(x)$ to be large enough to reliably estimate the local power of the CV residual error in order to estimate $\theta(x)$ correctly, but small enough to enable rapid local transitions. A reasonable balance can be specified in terms of the highest frequency with significant signal power, which is captured by the regularization parameter $\theta(x)$ itself. In particular, for regularizers that are close to optimal, we expect the residual signal to have its power concentrated at wavelengths near $\theta(x)$. In this case a good measure of the local residual power is to aggregate the squared residual over a small number of wavelengths $\theta(x)$ [43]. Thus we can reliably estimate both $\theta(x)$ and $\rho(x)$ as long as $\rho(x)$ is constrained to be a small multiple of $\theta(x)$.

We therefore adopt a simple heuristic, i.e., that $\rho(x) > \gamma\,\theta(x)$ where γ, the adaptive window factor (AWF), is a constant. With this constraint we obtain the final form of the computational problem solved in non-uniform refinement to regularize 3D electron density at each iteration; i.e.,

$$\begin{aligned} \theta^*(x) \;=\; & \arg\min_{\theta(x)} \min_{\rho(x)} \sum_{\xi} w_{\rho(x)}(\xi - x) \left[\, | \, (r_{\theta(x)} \circ m_1)(\xi) - m_2(\xi) \, |^2 + | \, (r_{\theta(x)} \circ m_2)(\xi) - m_1(\xi) \, |^2 \, \right] \\ & s.t. \quad \rho(x) > \gamma\,\theta(x) \end{aligned} \tag{8}$$

We find that as long as γ is larger than 3 we obtain reliable estimates of the local power of the residual signal. For $\gamma < 2$, estimates of residual power are noisy and optimization of the regularization parameters therefore suffers. The algorithm is relatively insensitive to values of γ greater 3, but there is some loss in the spatial resolution of the adaptive regularizer as γ increases.

Non-Uniform Refinement Algorithm

Given a set of particle images and a low resolution *ab initio* 3D map, non-uniform refinement comprises three main steps, similar to conventional uniform (homogeneous) refinement (Alg. 1). The data are randomly

Algorithm 2 Regularization Step for Non-Uniform Refinement

Require: Particle image dataset $\mathcal{D}$ with pose estimates z

1: Randomly partition $\mathcal{D}$ into halves, $\mathcal{D}_1$ and $\mathcal{D}_2$, with corresponding poses z_1 and z_2
2: Reconstruct $\tilde{m}_1$ and $\tilde{m}_2$, the raw (noisy) 3D maps from each half-set
3: Estimate regulization parameters θ^* by solving Eqn. 8
4: Reconstruct a single map from $\mathcal{D}$, z, and apply the optimal regularizer r_{θ^*}

partitioned into two halves, each of which is used to independently estimate a 3D half-map. This "gold standard" refinement [19] allows use of FSC for evaluating map quality, and for comparison with existing algorithms. The alignment of particle images against their respective half-maps, and the reconstruction of the raw 3D density map (the E and M steps in Alg. 1) are also identical to uniform refinement.

The difference between uniform and non-uniform refinement is in the regularization step. First, in non-uniform refinement, regularization is performed independently in the two half-maps. As such, the estimation of the spatial regularization parameters in Algorithm 2 effectively partitions each half-dataset into quarter-datasets. We often refer to the raw reconstructions in Algorithm 2 as quarter-maps. The non-uniform refinements on half-maps are therefore entirely independent, satisfying the assumptions of a "gold-standard" refinement [19]. By contrast, conventional uniform refinement uses FSC between half-maps to determine regularization parameters at each iteration, thereby sharing masks and regularization parameters, both of which contaminate final FSC-based assessment because the two half-maps are no longer reconstructed independently.

Most importantly, non-uniform refinement uses Eqn. 8 to define the optimal parameters with which to regularize each half-set reconstruction at each refinement iteration. Figure 2 shows an example of the difference between uniform filtering (FSC-based) and the new CV-optimal regularizer used in non-uniform refinement. Uniform regularization removes signal and noise from all parts of the 3D map equally. Adaptive regularization, on the other hand, removes more noise from disordered regions, while retaining the high-resolution signal in well-structured regions that is critical for aligning 2D particle images in the next iteration.

In practice, for regularization parameter estimation, Equation 8 is solved on a discretized parameter space where a relatively simple discrete search method can be used (e.g., as opposed to continuous gradient-based optimization). The algorithm is implemented in Python within the *cryoSPARC* software platform [6], with most of the computation implemented on GPU accelerators. An efficient solution to Eqn. 8 is important in practice because this subproblem is solved twice for each iteration of a non-uniform refinement.

Finally, the tuning parameters for adaptive regularization are interpretable and relatively few in number. They include the order of the Butterworth kernel, the discretization of the parameter space, and the scalar relating $\rho(x)$ and $\theta(x)$, called the adaptive window factor (AWF). In all experiments, we use an 8^{nd}-order Butterworth filter, and a fixed AWF parameter $\gamma = 3$. We discretize the regularization parameters into 50 possible values, equispaced in the Fourier domain to provide greater sensitivity to small scale changes at finer resolutions. We find that non-uniform refinement is approximately two times slower than uniform refinement in our current implementation.

Over-fitting of regularizer parameters

As mentioned in Discussion, local resolution estimates or local statistical tests have commonly been used to adaptively filter 3D maps. While these methods are generally satisfactory as one-time post-processing steps for

visualization, in our experience they can lead to severe over-fitting when used iteratively within refinement as a substitute for regularization. (Supplementary Figure 1 illustrates one example with the $Na_V1.7$ dataset.) During iterative refinement, small mis-estimations of local resolution at a few locations (due to high estimator variance [14]) cause subtle over- or under-fitting, leaving slight density variations. Over multiple iterations of refinement, these errors can produce strong erroneous density that contaminate particle alignments and the local estimation of resolution itself, creating a vicious cycle. A related technique using iterative local resolution and filtering was described briefly in EMAN2.2 documentation [9] and may suffer the same problem. The resulting artefacts (e.g., streaking and spikey density radiating from the structure) are particularly prevalent in datasets with junk particles, structured outliers, or small particles that are already difficult to align. To mitigate these problems, the approach we advocate couples an implicit resolution measure to a particular choice of local regularizer, with optimization explicitly designed to control model capacity and avoid over-fitting of regularizer parameters.

Data preprocessing

Experimental results for STRA6-CaM and *Pf*CRT datasets were computed directly from particle image stacks, with no further preprocessing. The original data for the $Na_V1.7$ protein comprise 25,084 raw microscope movies (EMPIAR-10261) from a Gatan K2 Summit direct electron detector in counting mode, with a pixel size of $0.849\mathring{A}$. We processed the dataset through motion correction, CTF estimation, particle picking, 2D classification, *ab-initio* reconstruction and heterogeneous refinement in *cryoSPARC* v2.11 [6]. A total of 738,436 particles were extracted and then curated using 2D classification yielding 431,741 particle images (pixel size $1.21\mathring{A}$, box size 360 pixels). As in [26], we detected two discrete conformations corresponding to the active and inactive states of the channel, with 300,759 and 130,982 particles. We obtained reconstructions with resolutions better than the published literature for both states, but for the results in this work we focus solely on the active state (refined with C2 symmetry yielding 601,518 asymmetric units).

Data Availability

No new datasets were created in this study. The raw datasets analyzed in this study were either downloaded from the EMPIAR repository (EMPIAR-10024, EMPIAR-10330, EMPIAR-10261) or were provided by authors of other studies, cited in the main text. Density maps and atomic coordinates from EMDB-0341 and PDB-6N4Q were used for visualization. All other results and outputs of data analysis in this study are available from the corresponding author on reasonable request.

Code Availability

The *cryoSPARC* software package is freely available in executable form for non-profit academic use at www.cryosparc.com.

References

[40] T. Bepler, A. Noble, and B. Berger, "Topaz-Denoise: General deep denoising models for cryoEM," *bioRxiv*, 2019. 18

[41] D. Tegunov and P. Cramer, "Real-time cryo-electron microscopy data preprocessing with warp," *Nature Methods*, vol. 16, no. 11, pp. 1146–1152, 2019. 18

[42] T.-O. Buchholz, M. Jordan, G. Pigino, and F. Jug, "cryoCARE: Content-aware image restoration for cryo-transmission electron microscopy data," *IEEE International Symposium on Biomedical Imaging*, 2019. 18

[43] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, pp. 95–120, 1994. 19

Supplementary Information for

Non-uniform refinement: Adaptive regularization improves single particle cryo-EM reconstruction

The follow sections provide supplementary information concerning potential over-fitting of regularization parameters, and the sensitivity of the non-uniform refinement algorithm to variations of default parameter values and filter kernels. We also include more details concerning the masking used during refinement in the experiments, and we include results from one more experimental dataset, namely, a 90 kDa membrane protein with no soluble domain.

Over-fitting regularization parameters

As mentioned in Discussion, local resolution estimates or local statistical tests have commonly been used to adaptively filter 3D maps. While these methods are generally satisfactory as one-time post-processing steps for visualization, in our experience they can lead to severe over-fitting when used iteratively within refinement as a substitute for regularization. During iterative refinement, small mis-estimations of local resolution at a few locations, e.g., due to noise in local resolution estimates, cause subtle over- or under-fitting, leaving slight density variations. Over multiple iterations of refinement, these errors can produce strong erroneous density that contaminate particle alignments and the local estimation of resolution itself, creating a vicious cycle. Supplementary Figure 7 provides an example of such over-fitting.

The resulting artefacts (e.g., streaking and spikey density radiating from the structure) are particularly prevalent in datasets with junk particles, structured outliers, or small particles that are already difficult to align. To mitigate these problems, the approach we advocate couples an implicit resolution measure to a particular choice of local regularizer, with optimization explicitly designed to control model capacity and avoid over-fitting of regularizer parameters.

Parameter sensitivity in Non-Uniform Refinement

Here we examine the sensitivity of non-uniform refinement to the adaptive window factor (the AWF parameter), which controls the spatial support used in estimating the regularizer parameters, and the family of filters used for the adaptive regularization. We consider values of the AWF parameter between 1 and 10. And we consider a number of low-pass filter kernels, including different orders of Butterworth filter, an isotropic Gaussian kernel (the parameter of which is the standard deviation), and a Sinc filter (the parameter of which is the cut-off frequency beyond which amplitudes are reduced to zero). We consider these variations in estimating the density map from the STRA6-CaM dataset. Beyond the parameters specified, all other parameters of the algorithm were held fixed at the default values used in all other experiments in the paper.

Supplementary Figure 8 shows FSC curves for density maps computed from STRA6-CaM data for different values of AWF (left), and different types of filter kernel (right). The results demonstrate that non-uniform refinement is not highly sensitive to these parameter choices on this dataset. For all experiments in the paper we used default choices of AWF=3 and an 8th-order Butterworth filter. We find these choices produce good results across many datasets, including STRA6-CaM. It is possible that one could vary these choices for each dataset to obtain

marginal improvements in reconstruction resolution. Nevertheless, one advantage of the non-uniform refinement algorithm, as shown in these plots, is that manual tuning is not required to achieve high quality reconstructions in practice.

Masks used during refinement

In conventional uniform refinement algorithms it is common practice to apply a real-space mask to the 3D reference density at each iteration of refinement. This plays an essential role during refinement in removing noise from the solvent regions of the map, thereby facilitating more accurate particle image alignments to the 3D map, and hence the overall resolution of the final 3D map. In the experiments in this paper we used the standard dynamic masking routine in *cryoSPARC* for both uniform and non-uniform refinements. This ensures a fair comparison between the two algorithms. For this dynamic masking routine, a soft, shaped mask is generated at each iteration by thresholding the 3D density map, followed by dilation and then application of soft padding. Supplementary Figure 9 shows examples of the masks used for uniform refinement of the STRA6-CaM, PfCRT and $Na_v1.7$ channel reconstructions. In all cases, the threshold is set to 20% of the maximum density, with dilation of 6Å and soft cosine padding of 8Å.

90 kDa Membrane Protein with no soluble region.

In addition to the protein complexes discussed in the main body of the paper, here we include results on a C3 symmetric membrane protein with a molecular weight of 90 kDa and no soluble domain, and no Fabs bound. The protein mass is surrounded by a large detergent micelle, making the pose estimation of particle images difficult. The dataset, courtesy of Oliver Clarke, contains 42,740 particle images (pixel size $1.05\mathring{A}$, box size 200 pixels, C3 symmetry yielding 128,200 asymmetric units). It is part of an on-going, unpublished study, used here with permission, but we do not name the protein or show the entire 3D map.

Supplementary Figure 10A shows FSC curves from uniform and non-uniform refinements, with improvements in global 3D map resolution from $3.9\mathring{A}$ to $3.6\mathring{A}$. One can also see that some local regions improve in quality more than the numerical resolutions might suggest. Supplementary Figure 10C shows densities for three transmembrane α-helices, with two (left, center) being closer to the periphery of the protein near the micelle, and one (right) being more central. The peripheral α-helices show significant improvement in the resolved detail of the 3D map, allowing tracing of backbone and side-chains in the non-uniform refinement map. By comparison, model building would be difficult with the corresponding uniform refinement map. The improvement of the central α-helix is modest, and was chosen to show that some parts of the uniform refinement map are resolved at its nominal $3.9\mathring{A}$ resolution. Supplementary Figure 10B shows large differences in particle alignments between uniform and non-uniform refinement.

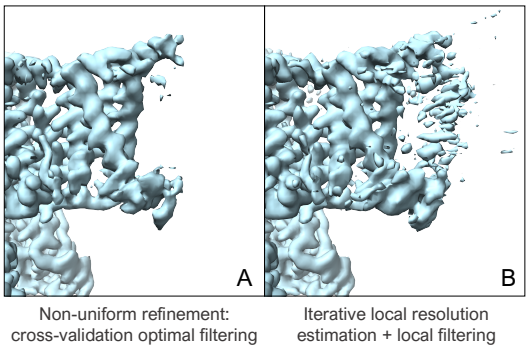


Figure 7: Side view of a peripheral transmembrane region of the $Na_V1.7$ channel showing that naive forms of local filtering during iterative refinement can lead to significant over-fitting. **A**: Unsharpened density map at the 8^{th} iteration of non-uniform refinement as presented in this work. **B**: Unsharpened density map at the 8^{th} iteration of a simpler iterative refinement scheme where local resolution estimation and local filtering are performed at each iteration. Both maps are thresholded at the same level. The map in **B** has acquired clear artefacts of over-fitting, with incorrect high-resolution density dominating the peripheral transmembrane region of the protein. This artefactual density continues to strengthen over subsequent iterations, contaminating local resolution estimates and leading to errors in particle alignments.

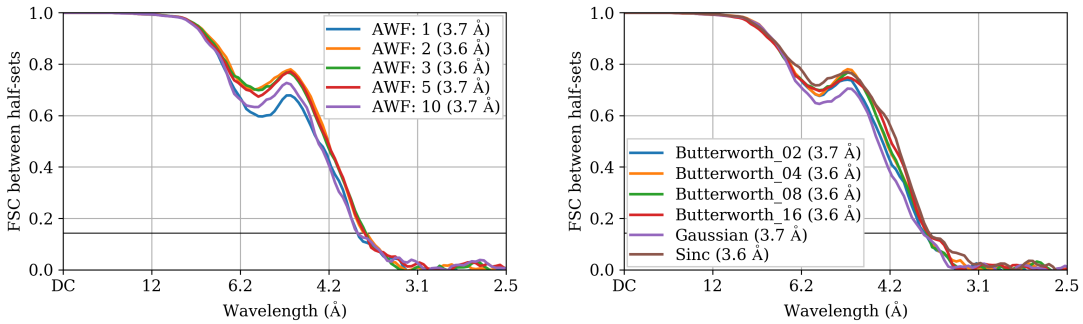


Figure 8: Gold-Standard FSC curves for non-uniform refinement reconstructions computed on the STRA6-CaM [24] dataset. **Left:** Curves for reconstructions computed with varying values of Adaptive Window Factor (AWF). Values of 2, 3, and 5 yield very similar resolution and FSC curves, while values of 1 and 10 show slight decrease in resolution. All results are computed with an 8th-order Butterworth filter. **Right:** Curves for reconstructions computed with varying types of filter using in adaptive filtering. Gaussian and low-order Butterworth filters perform slightly worse compared to high-order Butterworth (8th and 16th order) and Sinc filters. All results are computed with AWF=3.

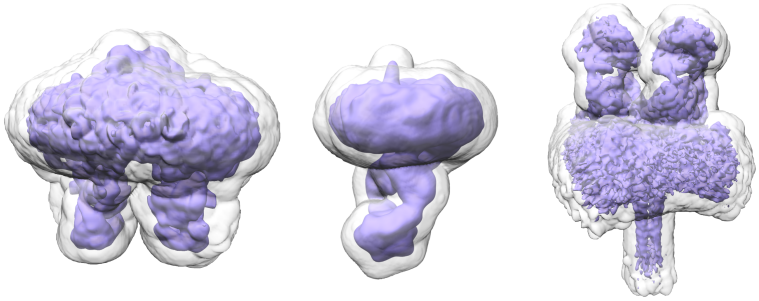


Figure 9: 3D density maps for STRA6-CaM (left), PfCRT (middle), and the $Na_v1.7$ channel (right) shown with spatial masks used during 2D-3D alignment during refinement iterations. In standard cryoSPARC dynamic masking, a soft, shaped mask is generated at each iteration by thresholding the 3D density map, followed by dilation and then application of soft padding.

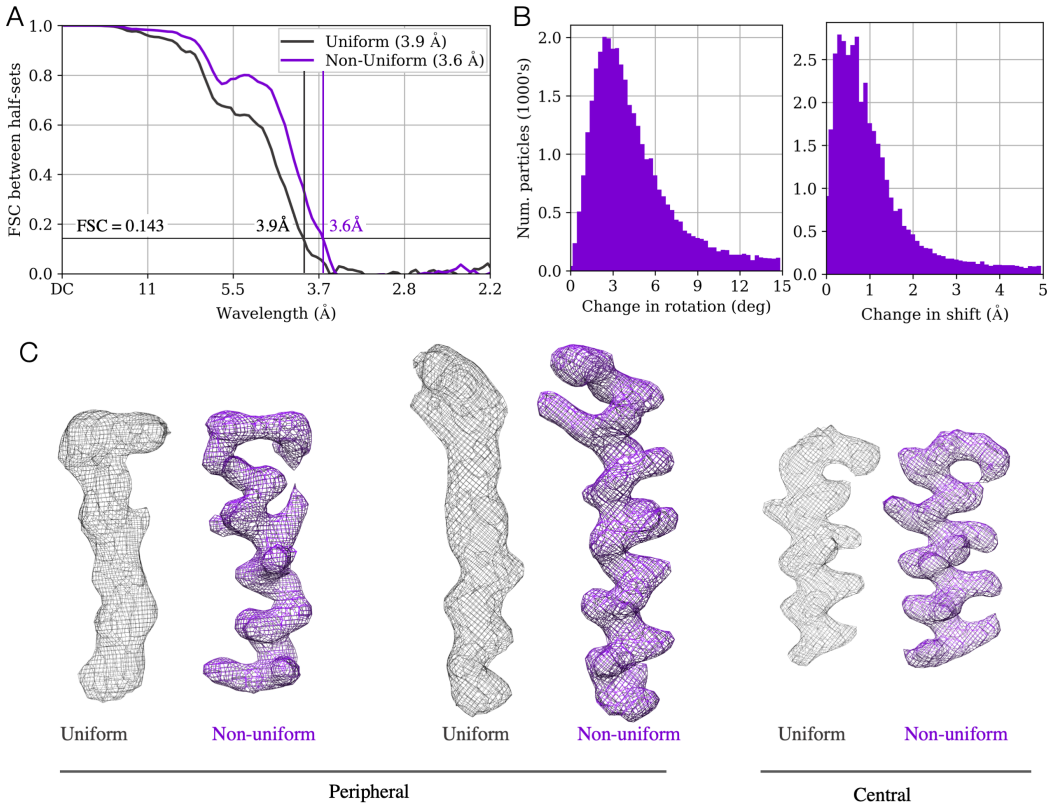


Figure 10: **C3 Symmetric Membrane Protein:** Results of uniform and non-uniform refinement on a dataset of 42,740 particle images of a C3 symmetric 90 kDa membrane protein with no soluble domains (data courtesy of Oliver Clarke). **A**: FSC curves computed using the same mask show numerical improvement from $3.9\mathring{A}$ to $3.6\mathring{A}$, indicating improved global average resolution. **B**: Histograms of change in particle alignments between uniform and non-uniform refinement. Optimal regularization in non-uniform refinement improves the ability to align particles over iterations, causing these changes. **D**: Density map detail for transmembrane α-helices. Two (left, center) are peripheral, near the micelle. One (right) is central within the protein. Both maps are filtered using the corresponding FSC curve, sharpened with the same B-factor of $-180\mathring{A}^2$, no local filtering or sharpening is used, and thresholds are set to keep enclosed volume constant. Density differences are thus due to algorithmic rather than visualization differences. Non-uniform (purple) map has clear density for backbone and side-chains while for peripheral α-helices, helical pitch is barely resolved in uniform (grey) map density. The central α-helix (right) improves less dramatically.