A MAP-ONLY FRAMEWORK FOR PREDICTING RESIDUE PLACEMENT CONFIDENCE DURING CRYO-EM MODEL BUILDING

by

SARTHAK MOHANTY, B.S.

Undergraduate Thesis

Presented to the Faculty of

The University of Texas at San Antonio

for the degree of

BACHELOR OF SCIENCE IN BIOCHEMISTRY

THE UNIVERSITY OF TEXAS AT SAN ANTONIO

College of Sciences

Department of Chemistry

August 2026

COMMITTEE MEMBERSHIP

TITLE:	A Map-Only Framework for Predicting Residue Placement Confidence During Cryo-EM Model Building
AUTHOR:	Sarthak Mohanty
DATE SUBMITTED:	August 7th, 2026
FACULTY SUPERVISOR:	Hadi Arman, Ph.D. Professor, Department of Chemistry The University of Texas at San Antonio
COMMITTEE MEMBER:	Maria Falzone, Ph.D. Assistant Professor, Department of Biochemistry and Structural Biology The University of Texas Health Science Center at San Antonio
COMMITTEE MEMBER:	Audrey Lamb, Ph.D. Chair, Department of Chemistry The University of Texas at San Antonio

DEDICATION

Dedicated to my younger self, who decided to go through this challenging life experience, and to everyone that helped me learn to see the world as fundamental particles and their interactions. Your guidance shaped this work, and the physician-scientist I hope to become.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my thesis committee and the friends, family, and mentors who shaped this work.

Dr. Hadi Arman served as my official thesis supervisor, and his mentorship extended far beyond the pages of this work. His influence shaped my development as a scientist, a teacher, and a researcher across every role I held at UTSA. He helped me understand the beginning of structure, and more importantly its biological importance.

Dr. Maria Falzone provided the intellectual home in which this thesis grew. Her willingness to take a chance on an undergraduate with big questions and limited cryo-EM experience, and to mentor me with the seriousness of a graduate student, is something I will carry with me for the rest of my career. This work is as much a product of her scientific vision as it is my own.

Dr. Audrey Lamb, whose Biochemistry II course first showed me that enzymes are not just mechanisms to memorize but stories about molecular logic. I am grateful for your careful reading of this work and for planting seeds I did not know were growing.

Dr. Christopher Sandford taught me that science without communication is incomplete. Through Stereochemistry in Synthesis and Chemical Communications, he pushed me to think more clearly, write more precisely, and teach more intentionally. If this thesis reads well, a meaningful share of the credit belongs to him.

Margot, my 2022 MacBook Pro, who I would be remiss not to acknowledge, who endured countless crashed calculations, thermal throttling, and fan speeds that could only be described as aeronautical. When she crashed, she started back up and never gave up, even when she probably should have.

Several friends and friend groups supported me during this thesis:

- David Volk, Stefan Nashawati, Legend Kwate, Jude Aden, and Heather Green for several supportive roles throughout the last two semesters of my degree and who gave constant feedback on drafts.

- Rubi Cordero and Jacqlyn Cervantes for helping me love and, more importantly, succeed in Dr. Reeves' Organic Chemistry I and II.

- Amber Hall, Xiara Rodriguez, Kamila Rivas-kadour, and Hector Martinez who convinced me to see it through in Dr. Audrey Lamb's class when the easiest option was to quit. I did end up seeing it through and passing.

Mazzie, my 2019 Mazda CX-3 who got totaled when I was out of town. He was with me for my entire college journey, save the last few weeks.

My grandmother for raising me to ask curious questions and whose love continues to inspire my pursuit of knowledge.

All large-scale map processing and cohort batch jobs were run on the UTSA Advanced Research Computing (ARC) cluster; I am grateful to ARC staff for maintaining the compute environment and documentation. No external funding supported this work.

August 2026

ABSTRACT

Structural biology seeks to determine the three-dimensional arrangement of atoms within biological macromolecules to understand their functions. Recent advances in cryogenic electron microscopy (cryo-EM) have enabled the routine determination of macromolecular structures at near-atomic resolution without the need for crystallization, making cryo-EM a central technique in modern structural biology. Despite these advances, the reconstructions produced by cryo-EM are not yet atomic models. Rather, they are three-dimensional density maps from which an atomic model must be subsequently constructed through model building.

Depending on the resolution of the map, the process of model building can be interpretive. A growing number of reconstructions now reach true atomic resolution, where the density is unambiguous and requires little interpretation. However, many reconstructions still fall within the 2.5–4 Å average-resolution range, where the challenges described below are present throughout. Although some regions of a density map contain well-defined features that permit unambiguous atomic placement, other regions exhibit weaker or less informative densities arising from molecular flexibility, conformational heterogeneity, limited local signal, or reconstruction uncertainty. Consequently, not all regions of a cryo-EM map can be interpreted with equal confidence, and identifying these differences is an essential component of accurate structure determination.

During cryo-EM model building, researchers must decide where to place atomic structures within the density map, often before any model exists (from crystallography or NMR). Local resolution is a measure that builders routinely consult when deciding how much of a map to interpret; however, it primarily reflects reconstruction reproducibility rather than how well the density constrains atomic positions. These are distinct questions, and the extent to which they diverge depends on the local-resolution estimator used. I introduce the reliability score, a map-only metric

computed from the local gradient magnitude of the normalized half-map average, which requires no fitted model. From the local gradient magnitude, a percentile-ranked reliability score was derived and binned into omit/caution/build terciles to provide actionable placement guidance. Validated against the Q-score, a cryo-EM-native, model-to-map measure of placement confidence, the reliability score tracks per-residue placement confidence across a cohort of 55 maps and identifies hard-to-place residues with a median per-map AUC of 0.82. It tracks EMRinger slightly more closely than it tracks the Q-score (cohort-median $\rho = +0.62$ versus $+0.53$). Against the coordinate corrections published by MEDIC, an independent residue-level detector of model errors, residues assigned to the omit zone are strongly enriched for subsequent correction. On the same structures, four local-resolution estimators predict placement confidence less consistently and differ markedly from one another in how well they do so. The advantage over the strongest estimator is small, and it does not survive reduction of both predictors to coarse labels. The utility of this tool weakens above roughly 4 Å, where the half-map average no longer varies finely enough to resolve placement difficulty at the residue scale, and is not supported above roughly 6 Å global resolution; therefore, the 2.5–4 Å band remains the primary operating regime for this metric.

The framework is implemented as cryoRIDGE (Reliability Inferred from Density Gradient Energy), a Python package that computes the reliability score and its omit/caution/build zones from a pair of unsharpened half-maps. This work argues that the map-only, gradient-magnitude-based metric termed reliability provides a more consistent basis for ranking placement difficulty during model building than local resolution, whose usefulness for this purpose varies substantially with the estimator chosen.

Keywords: cryo-EM, structural biology, model building, map quality, local resolution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

ΔR^2	Incremental held-out variance explained
AUC	Area under the ROC curve
ρ	Spearman rank correlation coefficient
$5 \times 5 \times 5$	Cubic voxel neighborhood used for local statistics
V	Local gradient magnitude of the half-map average
Å	Angstrom (unit of length; 10^{-10} m)
χ_1	First side-chain dihedral angle (EMRinger)
CC	Windowed half-map cross-correlation (Pearson)
Q	Q-score (per-residue model-to-map resolvability)

LIST OF ABBREVIATIONS

ARC	Advanced Research Computing (UTSA)
AUC	Area Under the Curve
BlocRes	Local-resolution estimator based on local Fourier shell correlation in a moving window
CC	Cross-Correlation
cryo-EM	cryogenic Electron Microscopy
cryoRIDGE	Reliability Inferred from Density Gradient Energy
EMDB	Electron Microscopy Data Bank
EMRinger	Side-chain rotamer validation score for cryo-EM models
FSC	Fourier Shell Correlation
FSC-Q	Hybrid local FSC / model-to-map per-residue score
LOMO	Leave-One-Map-Out (cross-validation)
MEDIC	Model Error Detection in Cryo-EM (coordinate-correction benchmark)
MonoRes	Monogenic-signal local-resolution estimator
PDB	Protein Data Bank
Q-score	Atom/residue resolvability score from map–model agreement
RELION	REgularised LIkelihood OptimisatioN (single-particle processing)
ResMap	Local-resolution estimator based on local signal-to-noise testing
ROC	Receiver Operating Characteristic
UTSA	The University of Texas at San Antonio

Chapter 1

BACKGROUND

1.1 What Is Cryo-EM? How Are Maps Made?

Single-particle cryogenic electron microscopy (cryo-EM) is a widely used technique that allows structural biologists to determine the three-dimensional structure of proteins and other large molecules at near-atomic resolution, including molecules that are too large, flexible, or difficult to crystallize for traditional X-ray methods.[1,2] In a typical experiment, a purified sample is rapidly frozen in a thin layer of ice and imaged with an electron microscope. Because individual molecules are extremely sensitive to radiation damage, each image must be taken at very low electron doses; the resulting dataset comprises millions of noisy, two-dimensional projection images recorded from many random orientations. Computational algorithms then iteratively align and average these particle images to build a three-dimensional reconstruction called a density map, which reflects the electron scattering potential of the sample.

1.2 What Is Model Building? Why Does It Matter?

The density map is not yet a structure, but a three-dimensional grid of numbers, where each grid point holds a value representing local density (fig. 1.1). These grid points are called voxels, an analog to a pixel in 2D space. When turning this map into an atomic model, a builder or software must decide where to place each amino acid, fitting the known amino acid sequence of the protein obtained from the gene sequence into the density. This process is called model building, and the resulting atomic model is the basis for most scientific interpretations. Accordingly, residues placed where the density does not adequately constrain them can support conclusions about molecular function or mechanism that the data do not justify.

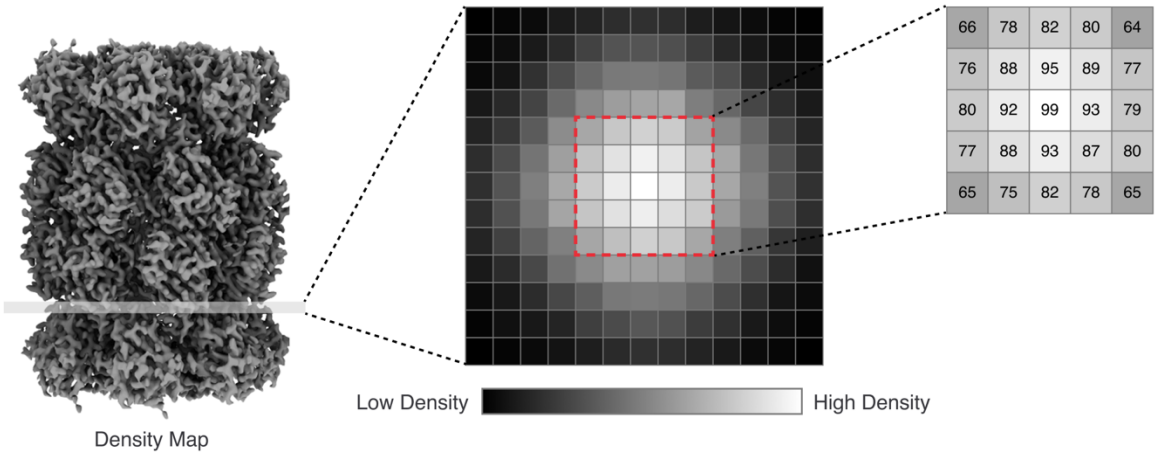

Figure 1.1: *Cryo-EM density maps are arrays of voxel intensities, shown here from the whole reconstruction of a T20S proteasome (EMD-6287) down to individual voxel values. On the left is a three-dimensional surface rendering of a reconstructed density map at 2.8 Å global resolution with a single two-dimensional slice through the volume. In the center, the slice is displayed as a grid of voxels, where each square is one grid point and its brightness encodes the local density, with brighter voxels corresponding to higher density. The dashed red box outlines a 5×5 region of voxels selected for magnification. Right: The magnified 5×5 block is ultimately a discrete array of numbers rather than a continuous surface. This voxel array formed the basis for all downstream calculations.*

Not all regions of this density map are of sufficient resolution for model building. Some parts are rigid and well-ordered, and therefore produce sharp, unambiguous density. Other parts, including flexible loops, mobile domains, or regions that are known to adopt different conformations in different environments around the protein, all produce smeared, weak, or ambiguous density, even when the overall map is of high quality. A model builder placing residues in these regions faces uncertainty about whether the map encodes enough information to constrain atomic positions.

The stakes for placement are higher than they appear. In practice, most researchers who use a deposited PDB model never inspect the underlying density to judge whether a given placement is justified; the atomic coordinates are taken at face value,[3] so a modeling error propagates silently into every downstream interpretation of the data. Every modeled residue is a claim about the structure of a molecule that downstream biological interpretation depends on directly, whether or not the density supports this claim. This is especially consequential for side chains, particularly longer side chains that can adopt several rotamers and mediate many specific contacts of a protein, where an incorrect rotamer or identity can change the inferred hydrogen bonds, ion coordination, or interface interactions. A common class of errors is the register shift. This occurs when the overall backbone trace through a region of density is essentially correct, but the residue identities are systematically offset by a few positions along the protein sequence. Register shifts in cryo-EM are one of the most difficult problems to identify and correct because the misplaced residues can still fit reasonably well into the density and occupy positions that appear plausible. Standard validation metrics do not reliably flag these errors, and the resulting model reports incorrect residues at key functional positions (i.e., active sites, ion coordination sites, and protein-protein interfaces). Therefore, a register shift of several residues can lead to an incorrect assignment of the amino acids responsible for a binding interaction or catalytic function, which can misdirect experiments for years. Register shifts are not the specific error this work sets out to detect, but they illustrate the general failure

mode it targets: density that is locally ambiguous yet plausible enough to be accepted without challenge. There are documented instances of cryo-EM models deposited and published only for independent groups to later find errors in them.[4,5] A recent validation tool, MEDIC, reported over 100 curated, error-typed corrections across 12 deposited structures.[4] Errors of this kind are therefore not isolated exceptions, at least among structures that have been independently re-examined. In this work, I used 2.5–4 Å as the atomic-building regime: the band that overlaps most of this difficult range while excluding resolutions at which neither the reliability score nor Q-score retains discriminative power.

In cryo-EM, the reported resolution of a reconstruction is a single global value that reflects the average resolution of the entire density map. This is determined by the Fourier Shell Correlation (FSC), which measures the agreement between two independently reconstructed half-maps as a function of spatial frequency. The resolution is taken as the spatial frequency at which this correlation falls below a fixed threshold, conventionally 0.143.[6] Because it is a single number computed over the entire map, it describes the average quality of the reconstruction, rather than the quality at any given position. During post-processing, the map is typically sharpened and then low-pass filtered to this global value; thus, the same nominal resolution is uniformly applied across the entire map volume.

Because a single global value cannot capture how map quality varies from one region to another, the most common assessment of local trustworthiness in the field is local resolution,[7,8] which estimates the different resolution values across the map.

1.3 Existing Approaches for Map Quality Assessment

Several tools estimate local resolution by analyzing how map quality varies spatially, and these estimates are widely used to guide sharpening[9] and, most importantly, to provide model

builders with a sense of where to trust density.[10,11] Unlike global resolution, which is a single number summarizing the overall quality of the reconstruction, local resolution estimators produce a value for each region of the map to reflect the density detail of the area.

Four implementations are compared in this work — BlocRes, ResMap, MonoRes, and the local-resolution estimate produced by RELION postprocessing — and each differs in the specific algorithm used to estimate the local frequency content.[10–12] BlocRes[13] computes local resolution by comparing the power spectra of overlapping sub-volumes between the two half-maps. ResMap[11] uses a hypothesis-testing framework based on the local signal-to-noise ratio. MonoRes[14] uses a monogenic signal approach that avoids the need to define local sub-volumes. RELION[15] derives its estimate from the same soft-masked FSC post-processing used to determine global resolution. Despite these algorithmic differences, all four tools answer the same question: "at each location in the map, up to what spatial frequency does the density contain correlated signal in the half maps?" which is the information content of the map, and these tools are well-validated for their intended purpose: to guide local sharpening[9] and to summarize overall map quality.[10,11,13]

The central hypothesis of this work is that local resolution is not the best tool for guiding accurate atomic placement. The limitation motivating this hypothesis is that local resolution methods measure how well the map resolves the signal, which is not necessarily correlated with the feasibility of constraining atomic placement.[8,16,17] A region may have moderate local resolution (i.e., half-maps agree up to a certain frequency) while still containing strong local gradient features (sharp, well-defined boundaries where density intensity changes steeply from one voxel to the next) that tightly constrain the true atomic position. Furthermore, local resolution estimates require interpretation from a builder and are not a straightforward set of instructions defining where an atomic model can be reliably built. Currently, no map-only, residue-scale approach is framed around placement rather than frequency content. Local resolution is available before a model exists and can be sampled at

the residue scale, but it answers a different question and requires interpretation before it becomes a placement decision, which is a particular obstacle for inexperienced structural biologists. This is the gap that this work addresses: a map-only, pre-model metric that ranks placement difficulty at the residue scale before any model has been built, giving builders a basis for deciding placement confidence. The goal is to build a tool that reports per-residue placement confidence directly from the half-maps, before any atomic model exists, and to benchmark it against local resolution as the current standard for guiding model building.

1.4 Gradient Magnitude & the Reliability Score

To address this gap, I developed and evaluated a new metric for characterizing local map quality. Throughout this work, the developed metric is denoted as the reliability score and is computed from the half-map average (fig. 1.2). The reliability score measures the gradient magnitude of the density, that is, how sharply the density changes across neighboring voxels as a proxy for how well-constrained a residue position would be. Importantly, the reliability score is a real-space quantity computed directly from the density values on the voxel grid, whereas local resolution is derived from the Fourier shell correlation and is therefore defined in Fourier (frequency) space. The reliability score is derived by ranking each voxel's gradient magnitude as a percentile throughout the map, yielding a normalized 0–1 score that is then binned into terciles (omit/caution/build) to provide builders with actionable guidance. To evaluate this approach, I compared per-residue reliability scores against two established post-model validators, the Q-score and EMRinger, across a cohort of deposited maps and benchmarked the reliability score against four independent local-resolution estimators.

Because the reliability score is derived only from the half-map average with no reference to the atomic model, there is no guarantee that it corresponds to anything that a model builder would

recognize as easy or difficult to place after model building. To test this correspondence, I used the Q-score, a model-to-map measure of placement confidence introduced by Pintilie et al.[16] The Q-score is computed for each atom by comparing the density values around the position of that atom to the density expected for a well-resolved atom at that resolution, and these per-atom scores are aggregated to yield a per-residue value. A Q-score below 0.5 is commonly used as a threshold[16,18,19] to indicate a residue with low placement confidence. Because Q-scores are computed directly from the model and map together, they are used as a cryo-EM-native measure of placement confidence and are routinely reported in the Electron Microscopy Data Bank (EMDB) for all recently deposited structures.

EMRinger,[20] which assesses model quality from a different structural signature than the Q-score, was used for comparison in this work. Rather than comparing the density values directly, EMRinger scans the $\chi 1$ dihedral angle of each long side chain and asks whether the Cγ atom's peak density falls at one of the discrete rotameric angles that well-ordered amino acids are expected to adopt. A side chain whose density peak falls at a non-rotameric angle is flagged as poorly fit, independent of whether the surrounding backbone density is otherwise strong. Because the EMRinger signal comes from the sidechain rotamer geometry rather than from a raw density-value comparison at each atom, it captures a partially independent aspect of model-to-map agreement from the Q-score, which is why both are reported as separate post-model comparators against the reliability score later in this thesis.

In practical terms, the reliability score is intended to identify where model building will be difficult; it is benchmarked against local resolution and against an established model-validation measure (the Q-score) in the Results. Across a cohort spanning a wide range of global resolutions and structural classes, the per-residue reliability score tracks the Q-score more closely than local resolution does, although the size of that advantage depends on the local-resolution estimator used.

Therefore, local resolution alone is an incomplete guide to where a builder should expect placement to be difficult.

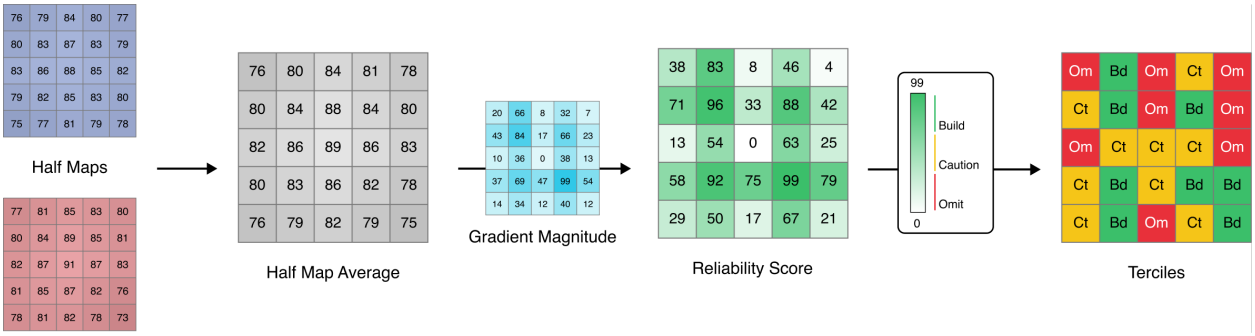

Figure 1.2: *Overview of the reliability score pipeline. The two unsharpened half-maps were averaged voxel-wise (the depositor-sharpened map was excluded), the gradient magnitude of the averaged density was computed, ranked by percentile (0–1), and binned into terciles (omit/caution/build).*

1.5 Aims of This Work

This work addresses four related questions that build on each other.

1. Which properties of the half-map average best predict the regions where a model builder will struggle to confidently place residues?

This motivated the choice of local gradient as the core signal and was examined through a systematic comparison of map features, including, but not limited to, local variance, gradient magnitude, and multiscale statistics.

2. Does the local gradient and reliability score, computed without any atomic model, correlate with the established model validation approaches derived from the final deposited model? If so, under what conditions?

This is the primary validation question, and the answer establishes a resolution regime consistent with the Q-score context and the structural contexts in which the reliability score is

most useful during building (i.e., heterogeneous reconstructions for assemblies with significant local variation).

3. How do the reliability score and local resolution each correlate with the Q-score and EMRinger at the residue scale, and does the reliability score provide model-building guidance that local resolution does not?

4. Do omit/caution/build zones provide actionable, ranked guidance in practice? Are omit zones meaningful for residues that the deposited model confirms are difficult to place?

Together, these four questions define the scope of this thesis and the motivation for this work. The framework developed to answer these questions is implemented as cryoRIDGE, a Python package that computes the reliability score and its omit/caution/build zones directly from a pair of unsharpened half-maps. The central claim is that, within the atomic building regime, the reliability score captures distinct placement-relevant signals that local resolution calculated with some programs does not provide. The practical value of the reliability score is that it flags areas of the map that may require more scrutiny, providing a quantitative evaluation from the half-maps alone before any atomic model exists. By doing so, the treatment of ambiguous density becomes more consistent across builders, replacing a judgment call that varies between individuals with a quantity computed identically for every map.

Chapter 2

METHODS

2.1 Data

A cohort of 71 cryo-EM half-map pairs was obtained from the EMDB. Of these, 55 entries with deposited protein models suitable for Q-score analysis formed the analysis cohort used throughout this thesis (Table A.1). This analysis cohort comprises a core cohort of 33 maps assembled first and an expansion cohort of 22 maps added subsequently to broaden the range of structural classes; the two terms are used in this sense throughout. Maps were selected to span a range of global resolutions (approximately 2–7 Å), molecular masses, and structural classes, including membrane proteins, soluble assemblies, and heterogeneous complexes. For each entry, both half-maps and the deposited primary map were downloaded. Across the 55-map analysis cohort, the isotropic voxel size ranged from 0.53 to 1.71 Å (median 1.04 Å); the production 5×5×5 window therefore corresponds to a median physical cube of ˜5.2 Å on a side (range ˜2.7–8.6 Å), or ˜4.7 Å on the reference map EMD-49450 (0.930 Å/voxel). Because the window is defined in voxels rather than Å, its physical size scales with each map's sampling; all within-map percentile rankings remain well-defined. The deposited contour level was used to define a binary mask distinguishing signal from solvent noise for all downstream calculations (the sensitivity to this choice is assessed in Table **??**). Per-residue Q-scores were computed for all 55 analysis-cohort maps using the qscore Python package[16] with the deposited primary map and structure, matching the standard map-model validation practice. The reliability score and other half-map statistics were computed on the half-map average.

The 16 entries without a deposited model suitable for Q-score analysis were validated using the windowed half-map cross-correlation (CC; see Windowed Half-Map Cross-Correlation) only. BlocRes local resolution estimates were computed for the 33 core-cohort maps using Bsoft blocres[13]

with a box size of 15 voxels at an FSC threshold of 0.143,[6] which was applied inside the depositor contour mask. Q-scores used eight radial samples per atom, and residues were matched by the deposited auth chain and sequence number and correlated with the reliability score at in-mask Cα positions.

2.2 Mathematical Formulas

The half-map average is defined as: $\overline{\rho} = \frac{1}{2}(h_1 + h_2)$ where h_1 and h_2 are the two unsharpened half-maps. The average is then globally z-scored over all voxels in the map box:

$$\widetilde{\rho} = \frac{\overline{\rho} - \mu_{\overline{\rho}}}{\sigma_{\overline{\rho}} + \varepsilon}$$

where $\mu_{\overline{\rho}}$ and $\sigma_{\overline{\rho}}$ are the mean and standard deviation of $\overline{\rho}$ over the full box, and ε guards against division by zero. This is a single affine transform applied uniformly to every voxel, so it rescales all gradients by a common factor and does not change within-map voxel rankings; it is included so that the gradient magnitude is on a comparable scale across maps and to match the analysis pipeline.

The gradient magnitude of $\widetilde{\rho}$ is estimated at each voxel by central differences. For a voxel at position i, j, k, the partial derivative in the x-direction is approximated as:

$$\frac{\partial \overline{\rho}}{\partial x}\Big|_{i,j,k} \approx \frac{\overline{\rho}(i+1,\, j,\, k) - \overline{\rho}(i-1,\, j,\, k)}{2\Delta x}$$

with analogous expressions in y and z, where Δx is the voxel spacing in Å. The local gradient magnitude is defined as half the mean squared gradient over a local 5×5×5 neighborhood:

$$V_{i,j,k} = \frac{1}{2}\left\langle \|\nabla \widetilde{\rho}\|^2 \right\rangle_W = \frac{1}{2\,|W|} \sum_{W(i,j,k)} \left[\left(\frac{\partial \widetilde{\rho}}{\partial x}\right)^2 + \left(\frac{\partial \widetilde{\rho}}{\partial y}\right)^2 + \left(\frac{\partial \widetilde{\rho}}{\partial z}\right)^2 \right]$$

where $\left\langle \|\nabla\widetilde{\rho}\|^2 \right\rangle_W$ denotes the mean over a 5×5×5 box W centered at (i, j, k), implemented as a uniform box filter, and $|W| = 125$ is the number of voxels in the window.

Although referred to throughout as the gradient magnitude for brevity, this quantity is a windowed mean squared gradient, that is, an energy density rather than the magnitude of a single gradient vector. Local Dirichlet energy density is a quantity widely used in physics, image processing, and geometric analysis to measure how rapidly a field changes over space.[21] Regions with high Dirichlet energy contain sharp spatial transitions, whereas regions with low Dirichlet energy are smooth. The local gradient magnitude is the local Dirichlet energy density, $\frac{1}{2}\|\nabla\widetilde{\rho}\|^2$, of the normalized half-map average. Global z-scoring is a monotone transform, and squaring is monotone on non-negative values, so both preserve the pointwise ordering of the magnitude of $\|\nabla\widetilde{\rho}\|$; after window averaging, it tracks mean squared gradient energy.

The reliability score R is derived by ranking gradient magnitude values across all in-mask voxels within a map and normalizing to the unit interval:

$$R_{i,j,k} = \frac{rank\,(V_{i,j,k}) - 1}{N - 1}$$

where N is the total number of in-mask voxels and ranks are assigned in ascending order (the lowest gradient magnitude receives rank 1). Because R is a within-map percentile of the gradient magnitude, the monotone z-score and squaring operations leave it unchanged. Voxels are then assigned to terciles:

$$\text{Zone}_{i,j,k} = \begin{cases} \text{Omit} & \text{if } R_{i,j,k} \leq 0.33 \\ \text{Caution} & \text{if } 0.33 < R_{i,j,k} \leq 0.66 \\ \text{Build} & \text{if } R_{i,j,k} > 0.66 \end{cases}$$

2.3 Computing the Gradient Magnitude

For each map, the half-map average $\overline{\rho}$ was computed from the two unsharpened half-maps. The depositor-sharpened primary map was deliberately excluded from this calculation, as sharpening choices vary across depositors and can confound the local gradient magnitude. $\overline{\rho}$ was then z-scored over the full map box to give $\widetilde{\rho}$. The gradient of $\widetilde{\rho}$ was estimated by central differences (a three-point stencil per axis), and the per-voxel squared gradient magnitude $\|\nabla\widetilde{\rho}\|^2$ was averaged over a 5×5×5 box (uniform filter) and halved to give the local gradient magnitude. The box filter suppresses voxel-level noise in the raw gradient; the unwindowed variant (w = 1) is substantially noisier (Table **??**).

A Sobel gradient operator was tested as an alternative to central differences; the difference in the downstream Q-score correlation was negligible at window size 5 (Table **??**), and central differences were retained as the simpler formulation. The window size was swept across a range of neighborhood sizes (Table **??**); w = 5 was near-optimal and was used throughout.

Note that two distinct length scales are involved: central differences set the differentiation stencil (a three-point estimate per axis), while the 5×5×5 box sets the smoothing window applied to $\|\nabla\widetilde{\rho}\|^2$. These are independent and the window does not widen the gradient stencil.

2.4 Computing the Reliability Score & Terciles

Within each map, the gradient magnitude values of all in-mask voxels were ranked by percentile, yielding a normalized reliability score from 0 to 1 for each voxel. This percentile ranking was computed independently per map, making the reliability score a within-map relative measure rather than an absolute cross-map quantity. Voxels were then assigned to one of three terciles based on their reliability score: the lowest third was labeled omit, the middle third labeled caution, and

the highest third labeled build. The cut points are within-map quantiles of the gradient magnitude rather than fixed gradient-magnitude values; because R is itself a within-map percentile, the zone boundaries fall at R = 0.33 and R = 0.66 in every map regardless of that map's absolute quality; no absolute threshold is applied. For per-residue analysis, gradient magnitude and reliability score values were sampled at the Cα position of each in-mask residue using a 2 Å sphere, taking the mean value within that sphere. The 2 Å isotropic sphere was used for BlocRes aggregation, FSC-Q, cross-metric heatmaps, and placement-utility tables. For the headline ρ(Q, reliability score) cohort statistics, nearest-voxel sampling was used.

2.5 Spearman's Rank Correlation Coefficient (ρ)

Throughout this thesis, the Spearman rank correlation coefficient (ρ) was used to measure the association between two variables.[22] Unlike the Pearson correlation, which assumes a linear relationship and is sensitive to outliers, Spearman's ρ measures monotonic relationships (that is, whether one variable tends to increase as the other increases, or to decrease as the other increases), regardless of whether that relationship is linear. This was done by replacing the raw values with their ranks and computing the correlation between the ranks. Note that ρ denotes the Spearman coefficient throughout; the same symbol is conventionally used for electron density, which in this work appears only with an explicit spatial argument or with an overbar denoting the half-map average.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)}$$

where d_i is the difference between the ranks of paired observations x_i and y_i. A coefficient of +1 indicates a perfect positive association of ranks, meaning every increase in one variable corresponds to an increase in the other; a coefficient of –1 indicates a perfect negative association. In

practice, a coefficient of 0.1–0.3 is considered weak, meaning the variables are barely associated and neither can be reliably predicted from the other. A coefficient of 0.3–0.7 is considered moderate, indicating a real but partial association in which one variable constrains the other without determining it; a coefficient of 0.7–0.9 is considered strong, indicating a highly reliable trend. Correlations in the moderate band are routinely treated as substantive findings in other quantitative fields; one published example outside structural biology, a study of upward lightning propagation and peak stroke current, reported a coefficient of 0.64.[23] Spearman's ρ is used throughout this work rather than Pearson's because the relationship between the reliability score and placement metrics such as Q-scores is not assumed to be linear,[19] and the per-residue distributions of both quantities contain outliers that would distort a Pearson estimate.

2.6 Validation Against Q-Scores

Per-residue reliability score values were correlated against the per-residue Q-scores computed with the qscore package (see Data) using Spearman's rank correlation coefficient (ρ), computed independently for each map. Cohort-level summaries report the median ρ across all maps with finite correlation values. Resolution sensitivity was assessed by binning maps by global FSC resolution and computing median ρ(Q, reliability score) within each bin; a 0.5 Å sweep was additionally performed to identify the empirical resolution ceiling. All Q-score correlations are reported at in-mask Cα positions only.

2.7 Validation Against EMRinger

Per-residue EMRinger scores were computed for each cohort map with a deposited atomic model using Phenix 2.0, run on the deposited primary map and coordinates. Because EMRinger evaluates only residues carrying a γ-carbon on the χ1 dihedral, the per-map residue set is smaller

than that for the Q-score. EMRinger values were matched to reliability score values at the same in-mask Cα positions and correlated by Spearman ρ independently for each map, with cohort-level summaries reported as the median ρ across the maps returning a finite value. For the ROC analyses, the low-EMRinger positive class was defined by a within-map median split rather than an absolute cutoff, because EMRinger scores are not calibrated to a fixed interpretive threshold in the way that $Q < 0.5$ is.

2.8 Comparison Against MEDIC Coordinate Corrections

Residue-level coordinate corrections published by the MEDIC study were used as an independent, model-based validation set. The deposited maps and half-maps for the entries analyzed in this work were downloaded from the EMDB, and the reliability score and its terciles were computed from the half-maps alone, without reference to the deposited coordinates, MEDIC outputs, or any Q-score. Corrected residues were then matched to reliability score values at their Cα positions by author residue numbering. Two blocks were analyzed. The first compares deposited low-resolution models against independent high-resolution re-determinations, a residue being counted as corrected when its Cα shift exceeds 1 Å after alignment; the second compares deposited structures against the MEDIC-rebuilt versions of the same entries. In both blocks, enrichment is then tested within the reliability-score zones computed on the low-resolution map. Two further Block 1 pairs were excluded as incompatible. Enrichment was quantified as the odds ratio for a corrected residue falling in the omit tercile versus elsewhere; because residues within a map are not independent, cluster-robust confidence intervals with the map as the clustering unit are reported alongside the unadjusted odds ratio. Discrimination on the continuous score was summarized as the per-map ROC AUC for ranking corrected against uncorrected residues and reported as a cohort median.

2.9 Comparison With Local Resolution

To compare the success of the reliability score and local resolution, Spearman ρ(reliability score, locres) was computed per residue for each map and for each of the four programs and was summarized as a cohort median. The association between each local-resolution estimator and the post-model validators was assessed in the same way, computing ρ(Q, locres) and ρ(EMRinger, locres) per residue and taking cohort medians. Because the local resolution is reported in Å, where a lower value is better, these correlations are sign-aligned (multiplied by –1) wherever they are displayed alongside the reliability score, so that a higher correlation always indicates a closer agreement between the local resolution and the validator. The raw, unaligned values are reported in the Appendix.

Four local-resolution (locres) estimators were used for comparison: BlocRes (Bsoft), ResMap, RELION, and MonoRes. Each was run on the same unsharpened half-map pairs inside the depositor contour mask and produced a per-voxel resolution field in Å, which was sampled at in-mask Cα positions using the same 2 Å sphere applied to the reliability scores. BlocRes parameters are given in Data. Local-resolution volumes were produced using RELION 4.0 postprocessing. ResMap 1.1.5 was run using a patched headless fork (sarthaktexas/resmap-headless, commit 3173de0, derived from kuixu/ResMap) with compatibility fixes for Python 3.9+ and NumPy 1.24+. MonoRes was run using Xmipp 6.0.0 on a local workstation. Coverage differed among programs, and two distinct counts are reported. The first is the number of maps for which a program produced a usable per-voxel volume: BlocRes 33, MonoRes 31, ResMap 53 (EMD-11638 and EMD-52525 produced flat or empty outputs), and RELION 54. The second is the number of maps yielding a finite Spearman ρ(Q, locres) at in-mask Cα positions, which is the n reported alongside every cohort-median correlation: 47 for ResMap, 54 for RELION, 33 for BlocRes, and 28 for MonoRes. The six-map gap

for ResMap comprises entries on which ResMap ran successfully, but for which the residue-level Q correlation is undefined; the two counts are not strict subsets of one another, since one map with an unusable flat field still returns a finite correlation. Paired tests use whichever subset has both quantities finite, which is often 47 and sometimes smaller.

2.10 Receiver Operating Characteristic Curves & Area Under the Curve

To assess whether the reliability score can correctly identify hard-to-place residues, I used receiver operating characteristic (ROC) analysis.[24] A ROC curve is constructed by varying the decision threshold across all possible values of a continuous predictor (the reliability score), and at each threshold, two quantities are computed: the true positive rate (the fraction of genuinely hard-to-place residues, defined as Q $<$ 0.5,15 that the predictor correctly flags) and the false positive rate (the fraction of easy-to-place residues incorrectly flagged).

$$AUC = P\left(s\left(r_{\text{hard}}\right) > s\left(r_{\text{easy}}\right)\right)$$

where r_{hard} and r_{easy} are a randomly drawn Q$<$0.5 and Q$\geq$0.5 residue, and $s(\cdot)$ is the predictor score. Plotting these two rates against each other across all thresholds produces a curve; a perfect predictor produces a curve hugging the top-left corner of the plot, flagging all hard-to-place residues before flagging any easy ones, while a predictor no better than random chance produces a diagonal line (i.e., a straight diagonal from (0,0) to (1,1)). The Area Under the Curve (AUC) summarizes this curve as a single number between 0 and 1: an AUC of 0.5 corresponds to random chance, an AUC below 0.5 indicates a predictor that is systematically anti-correlated with the outcome (and would perform better than chance if its sign were reversed), and an AUC of 1.0 corresponds to perfect discrimination. In this thesis, AUC is computed per map and summarized as a cohort median, and the reliability score is compared directly to ResMap local resolution as

an alternative predictor on the same structures. The primary head-to-head comparison scores both predictors continuously; median-split ResMap flags and the wider-cohort continuous comparison are reported as secondary context. The corresponding values are given in the Results.

2.11 Statistical Testing and Uncertainty

Because per-residue observations within a map are not independent, all significance testing was performed at the level of one summary value per map, never pooling residue-level observations across maps, which would pseudo-replicate the sample and inflate the apparent significance. Paired comparisons between two predictors (for example, reliability score versus a local-resolution comparator on the same maps) used the Wilcoxon signed-rank test on the per-map paired differences, testing whether the median paired difference exceeded zero. Uncertainty on medians and paired differences was estimated by percentile bootstrap resampling of maps (5,000 resamples). Effect sizes for paired comparisons are reported as Cohen's d_z, which is the mean paired difference divided by the standard deviation of the paired differences across maps.

The Wilcoxon signed-rank test was used to evaluate this null hypothesis nonparametrically: for each map, the paired difference (e.g., AUC(reliability score) – AUC(ResMap)) was computed, the absolute values of these differences were ranked, and the ranks were summed separately for the positive and negative differences. Because the paired differences cannot be assumed to be normally distributed, a test operating on ranks rather than raw values was appropriate given the modest and variable cohort sizes involved. Comparisons where a directional hypothesis was specified in advance, that is, the reliability score was expected to outperform rather than merely differ from a given comparator, were reported as one-sided p-values, noted explicitly wherever they apply; a one-sided p-value is half the corresponding two-sided value for the same data, so the two should not be compared directly.

A statistically significant difference does not, on its own, indicate how large that difference is: with enough maps, even a negligible average gap can be significant. Cohen's d_z addresses this by expressing the mean paired difference in units of its own standard deviation across maps, making it comparable across metrics with different scales (AUC, correlation, and balanced accuracy). Following standard conventions, $d_z \approx 0.2$ is considered a small effect, 0.5 a medium effect, and 0.8 or above a large effect; effects below roughly 0.2 are treated as negligible in this thesis regardless of whether the associated p-value is small.

2.12 Operational Placement Utility

To test whether the reliability score provides useful guidance before a model exists, three analyses were performed. The continuous measures are computed on all 55 maps with a finite ρ(Q, reliability score) wherever the comparator allows; the operational enrichment and rank-recovery tables are computed on the 33-map core cohort, and head-to-head comparisons that require BlocRes are restricted to the core and intersection panels of 33 and 31 maps. First, the fraction of hard-to-place residues ($Q < 0.5$) falling in the omit zone was computed per map and compared to the background fraction of low-Q residues across all in-mask Cα positions. If the omit zone captures hard-to-place residues at a higher rate than the background, it is enriched for placement difficulty. Second, ROC curves and AUC were computed as described above, with local resolution included as a direct comparator on the same structures. Third, rank recovery was assessed by asking whether residues ranked lowest by reliability score were also ranked lowest by Q-score. This tested whether the ordering of placement difficulty is correct even without applying any threshold. All three analyses were computed per map and summarized as cohort medians.

2.13 Leave-One-Map-Out Cross-Validation

To assess whether the reliability score's placement utility generalizes to unseen maps rather than reflecting properties specific to the training cohort, a leave-one-map-out (LOMO) cross-validation procedure was applied. In each fold, one map was held out as the test case, while the remaining maps were used to establish any cross-map quantity entering the comparison, including the training-fold median used for the binary split variant; the tercile boundaries themselves are within-map percentiles and require no training data. The procedure was iterated once per cohort map: the held-out map contributed no residues to any quantity estimated on the training folds, its reliability score was recomputed from its own half-maps, and its residues were scored using thresholds derived only from the other maps. Two families of statistics were recorded for each fold. The first treats the reliability score as a continuous predictor, computing the area under the ROC curve for classifying $Q < 0.5$, together with the Spearman correlation between the score and Q-score at in-mask Cα positions. The second converts the score to a discrete flag, either the omit tercile or a below-training-median split and computes the balanced accuracy against the same $Q < 0.5$ label. Balanced accuracy is defined as $BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$. Both are reported because they disagree in an informative way, as described in the Results section. In practice, because the tercile boundaries are fixed at the 33rd and 66th percentiles of the within-map reliability score distribution rather than fit to any cross-map training data, LOMO tests the generalization of the operational thresholds rather than overfitting of learned parameters. For each held-out map, the AUC was computed independently using the reliability score as the continuous predictor and $Q < 0.5$ as the positive class, exactly as in the non-cross-validated analysis. Rank recovery was similarly computed per held-out map as Spearman ρ between the reliability score and Q-score at in-mask Cα positions. The held-out AUC and rank-recovery distributions are reported on the panel on which each comparison is defined

rather than on the full cohort: the continuous reliability score against ResMap on 32 maps (EMD-52525 is skipped for sparse ResMap coverage), the training-fold median split on the same 32 maps, the multi-estimator comparison on the 27-map subset described below, and the semi-prospective omit-zone analysis on the 33-map core cohort. Each summarizes how consistently the reliability score discriminates hard-to-place residues on maps that have not been tuned. Local resolution was evaluated under the same LOMO procedure for direct comparison, using the within-map median local-resolution value as the classification threshold in each fold. To avoid resting the comparison on a single estimator, this was repeated for BlocRes, ResMap, and MonoRes on the subset of 26 maps for which all three estimators returned complete per-voxel volumes. Maps with missing ResMap or MonoRes coverage were excluded from this subset rather than being carried forward with partial data.

2.14 Windowed Half-Map Cross-Correlation

The windowed half-map cross-correlation (CC), r, is a local reproducibility proxy computed at each voxel by measuring the Pearson correlation between the two half-map density values across a 5×5×5 cubic neighborhood. For each voxel at position (i, j, k), local means, variances, and the cross-moment are estimated via uniform filtering over the 125-voxel window:

$$\mu_{h_1} = \langle h_1 \rangle_w , \quad \mu_{h_2} = \langle h_2 \rangle_w$$

$$\sigma^2_{h_1} = \left\langle h_1^2 \right\rangle_w - \mu^2_{h_1}, \quad \sigma^2_{h_2} = \left\langle h_2^2 \right\rangle_w - \mu^2_{h_2}$$

$$\operatorname{cov}(h_1, h_2) = \langle h_1 h_2 \rangle_w - \mu_{h_1} \mu_{h_2}$$

The windowed half-map cross-correlation at each voxel is then expressed as:

$$r_{i,j,k} = \frac{cov\left(h_1, h_2\right)}{\sqrt{\sigma_{h_1}^2 \cdot \sigma_{h_2}^2}}$$

where $\langle \cdot \rangle_w$ denotes the local mean over the 5×5×5 window centered at (i, j, k), implemented via scipy.ndimage.uniform_filter with mode="nearest". The raw half-map amplitudes are used without normalization. At in-mask Cα positions, r is point-sampled at the nearest voxel with no additional sphere averaging.

This quantity is distinct from BlocRes local resolution and from FSC-based estimators: it is a sliding-window reproducibility proxy that measures local half-map agreement at the density amplitude level rather than in frequency space. It is also distinct from the reliability score, which is computed on the windowed squared gradient of the z-scored half-map average rather than the correlation between the two half-maps directly. Both the reliability score and the windowed half-map CC use the same 5×5×5 smoothing scale, but they measure different physical properties of the density. The reliability score measures how much the average density changes across space, while r measures how consistently the two half-maps agree on density values within a local neighborhood. The real-space Pearson formulation follows the CC definition used in VESPER for map comparison tasks,[25] applied here between half-maps rather than between two independent maps.

Three alternative formulations were considered and rejected. First, windowed Fourier shell correlation (the approach used by BlocRes and RELION)[13] was not used because a 5×5×5 window of 125 voxels was too small for meaningful Fourier decomposition. The frequency resolution is insufficient at this scale, and the resulting local FSC estimates are noisy and resolution-dependent, complicating their interpretation as a simple reproducibility proxy. Second, the cosine similarity between local half-map windows (used as a real-space FSC analog in recent robust resolution estimation work)[12] was considered but not adopted because it is sensitive to the relative magnitudes of the half-map amplitudes in ways that Pearson's CC, which normalizes by local variance, is not; since

the two half-maps may differ in overall amplitude due to reconstruction-specific scaling, Pearson's CC is the more robust choice for cross-map reproducibility measurement. Third, the global CC used in map-model validation tools such as phenix.map_model_cc[17] was not applicable here because it collapses spatial variation into a single number and cannot produce the per-voxel reproducibility map needed for residue-level comparison with the reliability score. Therefore, the windowed real-space Pearson CC is an appropriate choice for this work: it is locally defined, amplitude-normalized, computationally manageable at a 5×5×5 scale, and directly comparable to the reliability score at the same spatial resolution.

One important distinction from local resolution estimators is that the windowed half-map CC is not a resolution estimate and should not be interpreted as one. It measures whether the two half-maps agree on density values within a local neighborhood, which is a necessary but not sufficient condition for high local resolution. A region can have high CC (both half-maps agree on a smooth, featureless density) while still being at a low effective resolution. This is the exact distinction that motivates the axis-separation argument of this thesis. The windowed half-map CC is included as a comparator to test how much of the placement axis a reproducibility proxy computed in real space can recover; as reported in the Results, it recovers most of it, and that result is itself informative about what the placement signal consists of (see Discussion, What the Reliability Score Does & Does Not Offer).

2.15 Software & Reproducibility

All analyses were implemented in Python as the cryoridge package, published on the Python Package Index and installable with pip install cryoridge. Data and outputs are archived at Zenodo (10.5281/zenodo.20618526). The full reproduction commands are provided in the Appendix (Software & Reproducibility).

Chapter 3

RESULTS

3.1 The Reliability Score Correlates With Q-Score Across a Diverse Cohort

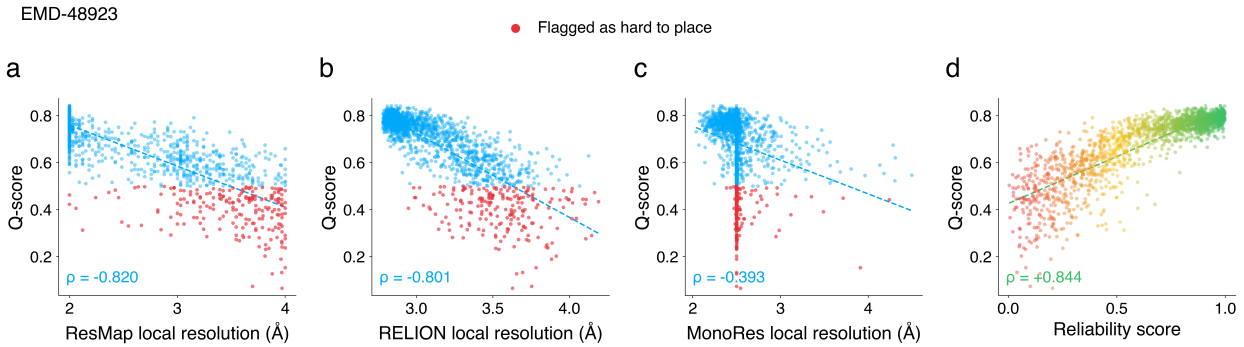

Figure 3.1: *The proposed reliability score outperforms local resolution as a predictor of per-residue model placement confidence in EMD-48923, a strong, illustrative example. Each point represents a residue, with the Q-score (a model-to-map measure of placement confidence) on the y-axis. (a–c) Q-score versus three local-resolution estimators: ResMap (Spearman ρ = –0.82), RELION (ρ = –0.80), and MonoRes (ρ = –0.39), the last of which is only weakly related to the Q-score. (D) Q-score versus the reliability score (percentile rank of V), where the relationship is stronger than for any of the three local-resolution estimators (ρ = +0.84). Red points mark residues flagged as hard to place, and the reliability score tracks placement confidence more closely than local resolution does.*

To assess how well the reliability score generalizes across cryo-EM maps, the per-residue reliability score was correlated with the Q-score. The Q-score is a post-model validator (map–model agreement), and the reliability score is computed from the half-map averaged density only and does not use the deposited coordinates. The test of agreement is whether map-only sharpness ranks placement difficulty similarly to atom-level resolvability measured after building. fig. 3.1 illustrates this comparison in a single map: in EMD-48923, the reliability score tracks the Q-score more closely (ρ = +0.84) than any of the three local resolution estimators, the strongest of which is ResMap at ρ = –0.82. Across 55 maps with finite per-residue ρ(Q, reliability score), the score showed a predominantly positive Spearman correlation with the Q-score at the in-mask

Cα positions. The cohort-median ρ(Q, reliability score) was +0.53, indicating a moderate positive association between the pre-model gradient magnitude and post-model placement confidence across structures spanning a wide range of resolutions, sizes, and structural classes. This correlation was not uniform across the cohort; the per-map ρ(Q, reliability score) ranged from slightly negative to +0.89 (Table 3.1) and was strongly modulated by global resolution, following a graded pattern rather than a single cutoff. Maps at or below 2.5 Å showed the strongest association (n = 6, median $\rho \approx 0.71$); the 2.5–4 Å atomic-building regime, which is the primary focus of this thesis, showed the main reported association (n = 42 maps with a finite ρ in this band, median ρ(Q, reliability score) = +0.54), and MgtA (EMD-49450, 2.73 Å), a P-type ATPase magnesium transporter used throughout this thesis as the reference map for single-map illustrations, achieved ρ = +0.82, which may reflect MgtA combining rigid transmembrane helices with more mobile cytoplasmic domains.[26] This is the regime where placement confidence varies sharply from residue to residue and where a map-only predictor offers the most to a builder. Maps in the 4–6 Å range showed a substantially weaker, more heterogeneous but still positive association (n = 5, median $\rho \approx 0.29$), and above approximately 6 Å, the correlation was not supported by the available data (n = 2, median $\rho \approx$ 0.02). This falloff reflects the progressive loss of side-chain density, and at coarser resolution even backbone features become unresolved: the local gradient contrast between well- and poorly ordered residues compresses toward the noise floor, and the reliability score loses the structural signal on which it depends (fig. 3.2). There is no established target value for this correlation, and ρ = +1.0 is not the goal: the reliability score is computed without any model, while the Q-score depends on the modeled coordinates; therefore, the two cannot be expected to agree perfectly, even where the map is unambiguous. The meaningful benchmark is comparative: whether the reliability score ranks placement difficulty better than the alternatives available to a builder before any model exists.

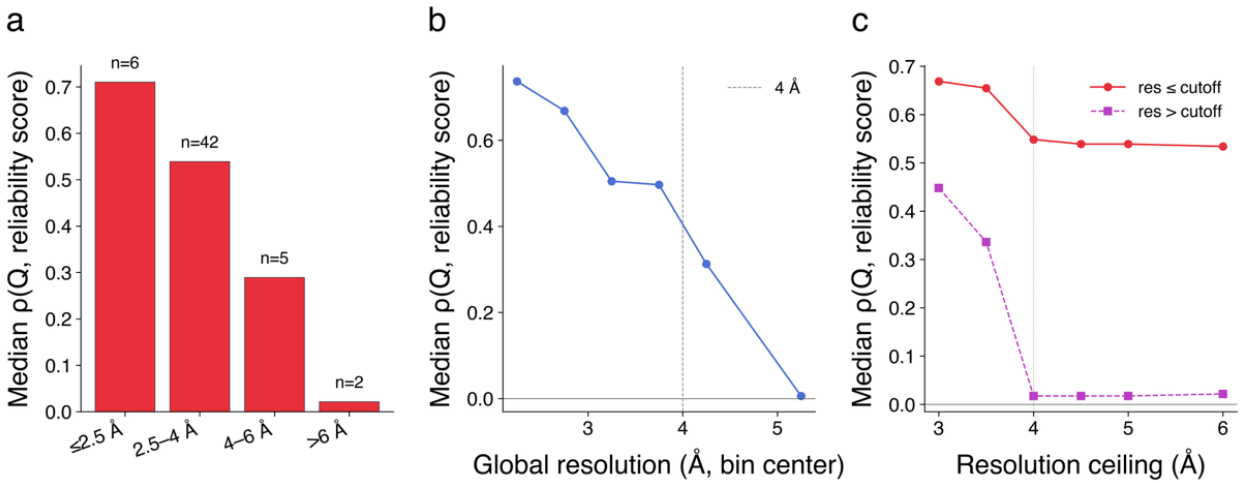

Figure 3.2: *Resolution sensitivity of ρ(Q, reliability score) across the cohort. (A) Median ρ(Q, reliability score) binned by global resolution, (B) 0.5 Å sweep of median ρ(Q, reliability score) from 2–6 Å, and (C) Cumulative comparison of maps at or below versus above each resolution ceiling of 4 Å.*

FSC-Q[27] represents a hybrid approach that combines local resolution estimation with model-to-map agreement and produces a per-residue score that compares the local FSC with the fit of the atomic model at that position. Similar to Q-scores, FSC-Q requires a fitted atomic model and is therefore a post-building validation tool rather than a pre-model guide for building atomic models. I examined FSC-Q as an additional comparison point on a subset of maps and found that FSC-Q and the reliability score were only weakly correlated (cohort-median $\rho \approx +0.05$; Table **??**), consistent with the two metrics measuring largely independent quantities. The key distinction between FSC-Q and the approach in this thesis is the same as that for Q-scores: FSC-Q is available only after a model has been built, whereas the reliability score is available beforehand. Independently of this comparison, the scope of the reliability score as a placement-confidence metric is the atomic-building resolution regime, where model-building decisions are both the most consequential and least visually obvious from the map alone.

Table 3.1: *Per-map Spearman correlation between Q-score and each map-only or local-resolution predictor across the 55-map analysis cohort*

EMDB	Protein	Res (Å)	n (Cα)	Reliability	ResMap	RELION	BlocRes	MonoRes
EMD-11638	Apoferritin	1.22	160	+0.40	+0.00	—	+0.13	—
EMD-61596	HSV-2 gB (+Fab16F9)	2.18	798	+0.68	+0.68	+0.56	+0.21	—
EMD-50549	Aerolysin WT in SMALP	2.20	2781	+0.74	+0.59	+0.64	—	—
EMD-47552	YnaI mechanosensitive channel (DDPC nanodisc)	2.30	2270	+0.83	+0.83	+0.86	—	—
EMD-24120	70S pre-translocation (E. coli)	2.33	5405	–0.08	–0.09	–0.14	–0.03	—
EMD-51569	Human muscle nAChR in nanodisc	2.48	2364	+0.88	+0.84	+0.79	—	—
EMD-52525	Complex III (mitochondrial)	2.52	437	+0.46	—	+0.11	+0.09	+0.09
EMD-37504	GLP-1R + Gs (ligand-free)	2.54	1027	+0.69	+0.36	+0.52	+0.27	+0.33
EMD-51902	Human myoferlin in nanodisc	2.56	1309	–0.06	–0.09	+0.04	—	—
EMD-71338	Human 80S hibernating class4	2.57	13006	+0.88	+0.86	+0.80	—	—
EMD-48923	MgtA (E2.Mg.BeF3)	2.59	1674	+0.84	+0.82	+0.80	+0.47	+0.39
EMD-71435	Human 80S + EBP1 in situ	2.67	12294	+0.88	+0.87	+0.82	—	—
EMD-41596	MsbA outward-facing	2.68	1069	+0.54	+0.09	+0.43	+0.16	+0.13
EMD-29275	Human pre-60S state L1 composite	2.70	7086	+0.41	—	+0.42	—	—
EMD-49450	MgtA (E2P+E1)	2.73	2928	+0.82	+0.82	+0.79	+0.44	+0.12
EMD-62841	Thermophile spliceosome ILS	2.80	12074	+0.89	+0.89	+0.90	+0.54	+0.58
EMD-71331	Human 80S P-E in situ	2.80	11577	+0.81	+0.80	+0.74	—	—
EMD-6287	T20S proteasome	2.80	5768	+0.64	+0.26	+0.37	+0.19	+0.41
EMD-51377	Human SLC45A4 in detergent	2.80	259	+0.07	+0.02	+0.16	—	—
EMD-25418	70S post-translocation (E. coli)	2.90	5835	+0.69	+0.62	+0.57	+0.15	+0.15
EMD-29262	Human pre-60S state G composite	2.90	6081	+0.49	—	+0.51	—	—
EMD-41756	LRRK2 CTD (+GZD824)	2.90	760	+0.67	+0.61	+0.40	+0.19	–0.03

EMDB	Protein	Res (Å)	n (Cα)	Reliability	ResMap	RELION	BlocRes	MonoRes
EMD-39332	Human 20S proteasome assembly intermediate	2.91	3485	+0.51	—	+0.37	—	—
EMD-53474	Human NTCP in nanodisc	3.00	636	+0.38	+0.36	+0.15	—	—
EMD-44471	Human nucleosome 3.0 Å	3.00	729	+0.41	+0.39	+0.30	+0.06	–0.01
EMD-53483	Rat SLCO2A1 in detergent-lipid micelle	3.00	428	–0.14	–0.18	–0.21	—	—
EMD-54253	70S + EF-G hybrid H1	3.00	6558	+0.68	+0.65	+0.56	—	—
EMD-48534	MgtA (E2P.Mg)	3.07	864	+0.72	+0.71	+0.63	+0.05	+0.47
EMD-33734	SARS-CoV-2 spike K202B bispecific	3.10	3449	+0.83	–0.02	+0.76	+0.25	+0.18
EMD-51351	70S + SaEF-G POST	3.20	5790	+0.76	+0.81	+0.70	—	—
EMD-19872	Integrator INTS5/8/15	3.20	1262	+0.37	—	+0.40	—	—
EMD-42603	Human p97/VCP + inhibitor	3.23	4012	+0.55	+0.45	+0.43	+0.04	+0.15
EMD-16091	Beta-galactosidase (5 ms time-resolved)	3.30	3824	+0.44	+0.36	+0.27	+0.20	+0.01
EMD-16119	GroEL-GroES ADP (turnover)	3.40	7780	+0.84	+0.09	+0.76	+0.35	+0.49
EMD-13308	GroEL-GroES ADP (tight)	3.43	6438	+0.46	+0.67	+0.61	+0.20	+0.18
EMD-11497	SARS-CoV-2 spike 3 RBD-down	3.50	2765	+0.76	+0.74	+0.65	+0.30	+0.46
EMD-64133	26S proteasome-Midnolin MB state	3.50	9871	–0.00	–0.01	+0.02	—	—
EMD-45604	beta2AR inactive (carazolol)	3.50	876	+0.24	+0.36	+0.37	+0.06	–0.02
EMD-41598	MsbA inward-facing	3.60	942	+0.35	+0.30	+0.28	+0.08	+0.25
EMD-74099	Human U2/BP spliceosome SF3A	3.61	1750	+0.53	+0.48	+0.49	—	—
EMD-11149	Bovine ATP synthase Fo	3.61	60	+0.11	+0.08	+0.03	+0.07	+0.11
EMD-23130	TRPV1 pH 6c	3.66	1808	+0.53	+0.59	+0.59	+0.19	+0.11
EMD-48311	Yeast V-ATPase Vo	3.70	2983	+0.70	+0.76	+0.45	+0.09	+0.03
EMD-23129	TRPV1 pH 6a	3.70	2058	+0.67	+0.67	+0.61	–0.09	+0.12
EMD-28487	Eag1 voltage sensor up	3.90	2959	+0.59	+0.59	+0.46	+0.26	+0.19
EMD-18839	Integrator cleavage module intermediate	3.90	1685	+0.50	—	+0.49	—	—
EMD-7130	Epithelial Na+ channel (ENaC)	3.90	1195	+0.08	+0.03	–0.16	+0.07	—

EMDB	Protein	Res (Å)	n (Cα)	Reliability	ResMap	RELION	BlocRes	MonoRes
EMD-45603	beta2AR apo	3.95	855	+0.15	+0.14	+0.10	+0.02	–0.04
EMD-50267	Integrator arm module state 1	4.00	1712	+0.80	—	+0.77	—	—
EMD-4941	ClpB WT-2A (ATPgammaS)	4.00	2706	+0.34	+0.32	+0.31	+0.10	+0.25
EMD-41042	S. enterica WbaP in SMALP	4.10	410	–0.00	—	+0.08	—	—
EMD-64103	26S proteasome-Midnolin MA state	4.32	10652	+0.29	+0.09	+0.48	—	—
EMD-28498	Eag1 voltage sensor down	5.40	3117	+0.01	–0.01	+0.00	+0.02	+0.03
EMD-4940	ClpB WT-1 (ATPgammaS)	6.20	4052	+0.02	–0.02	–0.01	–0.02	–0.03
EMD-9156	NMDA GluN1-GluN2A Extended-1	6.84	2285	+0.03	+0.00	+0.00	+0.01	—

Maps are ordered by global resolution. The five predictor columns provide the per-map Spearman ρ between the predictor and Q-score. Local-resolution correlations are sign-aligned (raw ρ multiplied by –1) so that a positive value always indicates worse local quality co-varies with lower Q-scores, making them directly comparable to the reliability score column. A shaded cell with an em dash indicates that the program produced no usable output for that map. The cohort medians were +0.53 for the reliability score (n = 55), +0.46 for RELION (n = 54), +0.39 for ResMap (n = 47), +0.14 for BlocRes (n = 33), and +0.14 for MonoRes (n = 28).

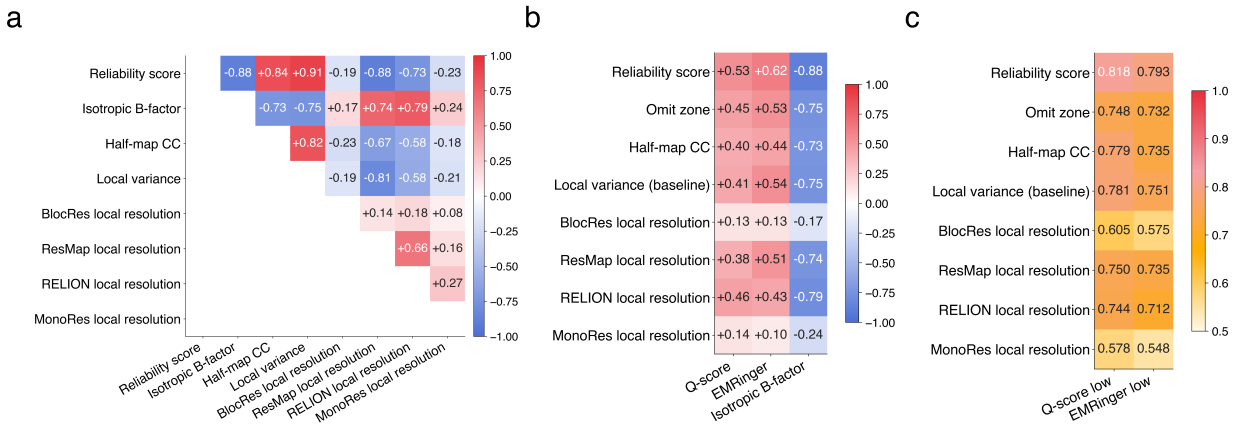

Figure 3.3: *Cohort cross-metric and validation summary. (A) Median within-map Spearman ρ between the map-only predictors (upper triangle; diagonal omitted). For each map, ρ was computed over in-mask residues with $\geq$30 finite pairs; the heatmap shows the cohort median across maps. Half-map CC is a real-space statistic: the Pearson correlation between the two independent half-maps within the same 5×5×5 voxel neighborhood used for the reliability score; higher values indicate locally reproducible density. Local resolution values are in raw Å (higher = worse). (B) Median per-map Spearman ρ between map-only predictors and post-model validators. Locres columns are sign-aligned (raw ρ multiplied by –1 so positive values mean worse local resolution co-varies with lower Q) so positive cells mean worse resolution co-varies with lower Q-score/EMRinger. (C) Median per-map ROC AUC for classifying low Q-score ($Q < 0.5$) and low EMRinger (below the in-map median) from each predictor score.*

The relationship between the reliability score and local resolution, the pre-model metric it most closely resembles, depends on the estimator used.[12] To determine whether the reliability score and local resolution captured the same atomic placement information, I compared their per-residue correlations with the Q-score and with each other across the cohort. The cross-metric heatmap

(fig. 3.3A) summarizes the within-map Spearman ρ between the map-only predictors at in-mask Cα positions. For each deposited structure, ρ was computed over residues with $\geq$30 finite pairs, and the cohort median was taken for each predictor pair. In panel A, the reliability score correlated strongly with the windowed half-map cross-correlation (CC; Pearson correlation of the two half-maps inside the same 5×5×5 neighborhood; +0.84) and local variance (variance of the half-map average in that neighborhood; +0.91), both computed over the same sliding window as the reliability score (Methods). These are real-space statistics that, like reliability scores, are sensitive to local density contrast. The reliability score was only weakly related to BlocRes (median ρ = –0.19, n = 33) and MonoRes (–0.23), but strongly related to RELION (–0.73, n = 54) and ResMap (–0.86, n = 47); all values are raw Å, so negative ρ means a higher reliability score co-varies with better resolution.

fig. 3.3B shows whether each predictor tracked the deposited model validation scores. Each cell is again a cohort median of per-map ρ; BlocRes, ResMap, RELION, and MonoRes local resolution are sign-aligned (multiplied by –1) so that positive ρ indicates worse Å resolution co-varies with lower Q-score or EMRinger. The reliability score was the strongest pre-model tracker of both validators (ρ = +0.53 vs. Q-score, n = 55; +0.62 vs. EMRinger, n = 53), but the estimators differed widely among themselves: RELION (+0.46 vs. Q-score, n = 54) and ResMap (+0.39, n = 47) tracked Q-score substantially better than BlocRes (+0.14, n = 33) or MonoRes (+0.14, n = 28). On the 31 maps for which BlocRes, ResMap, and RELION were all computed, the reliability score exceeded every estimator by paired Wilcoxon test (all $p < 0.001$), but the effect sizes were uneven: the median gain was +0.35 over BlocRes, +0.09 over RELION, and only +0.02 over ResMap. Therefore, the reliability score's advantage over the best-performing estimators is real but small. The margin over ResMap was statistically supported (n = 47 maps, paired Wilcoxon p = 0.0022, Cohen's d_z = 0.33; bootstrap CIs for both correlations in Table **??**).

fig. 3.3C shows the same separation in the ROC space, with a median per-map AUC for classifying $Q < 0.5$ of 0.82 for the continuous reliability score against 0.75 for continuous ResMap Å across every map in the analysis cohort returning a finite ResMap AUC (n = 53 of 55; EMD-11638 and EMD-52525 return none). The same ordering holds on the narrower paired panel, where both predictors return a value on all 31 maps: 0.82 against 0.71 (Table **??**). Restricting ResMap to the maps on which it produced a usable per-voxel volume narrows the margin further, to 0.82 for the reliability score (n = 54 of 55, all except EMD-11638, apoferritin at 1.22 Å, which has no $Q < 0.5$ residues and therefore no defined AUC) against 0.80 for continuous ResMap (n = 46); this 46-map subset and the 53-map finite-AUC set are not nested, so the two denominators should not be read as one containing the other. Within the 53-map finite-AUC set, the ordering is stable across cohort tiers (core cohort 0.82 versus 0.71, n = 32 and 31, respectively; expansion cohort 0.81 versus 0.76, n = 22), but the cohort-level advantage over ResMap is small.

The spread among the local resolution estimators is informative. All four programs answered the same nominal question — up to what spatial frequency do the half-maps agree at each location — yet they differed by a factor of three in how well that answer predicted the model quality. RELION and ResMap tracked the Q-score and EMRinger well, whereas BlocRes and MonoRes tracked them weakly. The programs differed in how they defined the local neighborhood over which agreement was assessed, whether that neighborhood was fixed or adaptive, and how the local signal was separated from noise; these implementation choices evidently mattered more for placement guidance than their shared Fourier-space framing would suggest. Several differences are plausible. BlocRes[13] computes a local FSC inside a moving cube whose size is a free user parameter; therefore, its estimate is averaged over a fixed spatial support that is coarse relative to a single residue. MonoRes[14] compares the amplitude of a monogenic signal against a noise distribution estimated outside the mask, which does not adapt to variations within compositionally heterogeneous particles. ResMap[11] instead asks,

voxel by voxel, whether a sinusoid of a given wavelength is detectable above local noise, whereas RELION[15] applies the same soft-masked post-processing used for global FSC. Notably, the split does not follow the Fourier-versus-statistical divide: RELION is FSC-based and ResMap is not, yet both track model quality well, while BlocRes and MonoRes track it weakly despite sitting on opposite sides of that divide. Therefore, the effective spatial support of the estimate and choice of noise reference appear to matter more than the underlying signal model. This is offered as an interpretation rather than a demonstrated mechanism because the cohort was not designed to isolate these factors.

The reliability score outperformed even the best-performing local-resolution estimator against both post-model validators, although by a small margin against ResMap and RELION. The reliability score was computed the same way for every map, requiring only the two unsharpened half-maps and returning a value for all 55 maps in the cohort. Given the large variability between local-resolution programs and the practical difficulty of running some of them to completion, the value of the reliability score lies less in measuring something orthogonal to local resolution and more in providing a single consistent, model-free quantity that is directly relevant to placement.

3.2 The Reliability Score Provides Actionable Placement Guidance

To test whether the reliability score provides useful guidance before a model exists, I evaluated it using three operational measures: how well the omit zone concentrates hard-to-place residues (enrichment), how well the continuous score ranks them (classification accuracy, ROC/AUC), and rank recovery. I classified residues as "hard-to-place" using a Q-score threshold of 0.5, the same cutoff used throughout this work, below which atom-level model–map agreement is generally considered poor.[16] The reliability score ranks the hard-to-place residues well. On the 31-map panel where BlocRes, ResMap, and RELION were all available, the median per-map AUC for classifying

low-Q residues (Q $<$ 0.5) was 0.82 for the continuous reliability score against 0.71 for continuous ResMap Å, with windowed half-map CC at 0.78; on the subset with usable ResMap coverage the two are closer, 0.82 for the reliability score (n = 54) against 0.80 for ResMap (n = 46) (Table **??**). In practical terms, an AUC of 0.82 means that given one low-confidence and one high-confidence residue drawn randomly from a map, the reliability score ranks the low-confidence residue as worse approximately 82% of the time, compared with approximately 71% for ResMap. Applying the same binary rule to both predictors instead — the omit tercile against a ResMap median split — gives 0.79 for the reliability score against 0.73 for ResMap on the same 31 maps. The continuous comparison is statistically supported: across those same 31 maps, the continuous reliability score's per-map AUC exceeds continuous ResMap under a paired Wilcoxon signed-rank test (median paired difference = +0.014, 95% CI [0.003, 0.025], one-sided p = 0.0070, Cohen's dz = 0.46, favoring the reliability score on 23 of 31 maps; bootstrap CIs for both AUCs are listed in Table **??**). The median paired difference is much smaller than the gap between the two cohort medians because it summarizes the per-map differences, most of which are small and positive, rather than the difference between the two medians. The effect was modest but consistent. A flag-to-flag comparison under a common median-split rule is reported separately under cross-validation (LOMO Cross-Validation Confirms Generalization), where the difference does not reach significance.

Across the 33-map placement panel, the omit zones recovered a cohort-median of 38% of all low-Q residues (Q $<$ 0.5), roughly twice the 19% cohort-median share of in-mask Cα positions that fall in the omit zone, which is the rate expected if omit-zone membership were independent of placement difficulty. The omit tercile is one-third of in-mask voxels but a smaller share of Cα positions, so the Cα figure rather than the voxel fraction is the correct null here. The residues they contain are also enriched for low Q: ~55% of omit-zone Cα have Q $<$ 0.5, compared with ~25% among all in-mask Cα, a ~2.2-fold enrichment. In the head-to-head comparison across three

operational metrics: enrichment, balanced classification accuracy, and rank recovery, the reliability score and omit zone performed comparably to windowed half-map CC and substantially better than BlocRes local resolution and local variance in all three criteria. However, BlocRes is the weakest of the four tested estimators, and the comparison with RELION and ResMap reported above is a more demanding test. For a builder triaging a newly deposited chain, this translates into a concrete workflow gain: focusing manual inspection on the roughly one-fifth of in-mask Cα positions that fall in the omit zone recovers approximately twice as many low-confidence residues per residue reviewed as an unguided read-through of the model would.

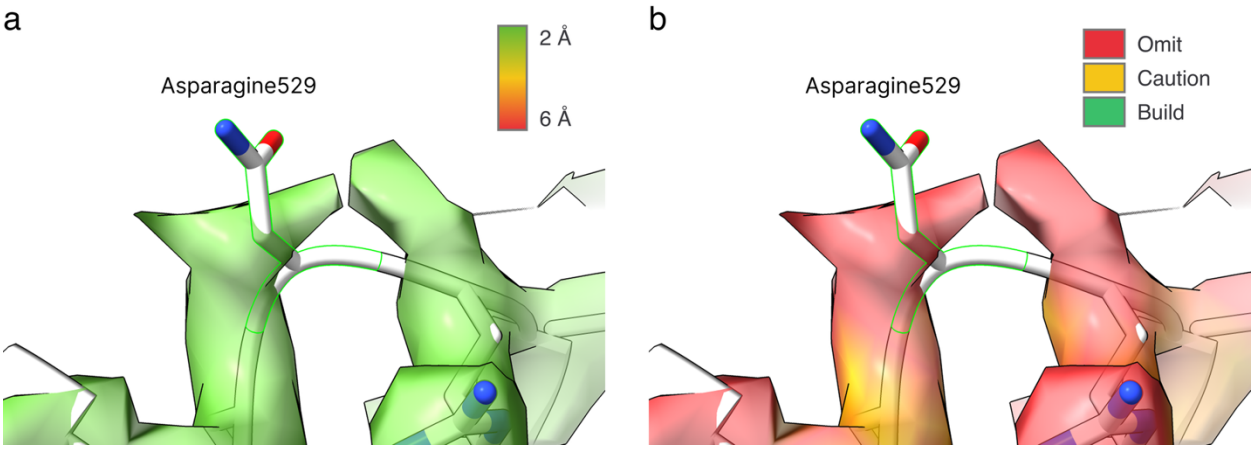

Figure 3.4: *ResMap local resolution and build zones at Asn529 (asparagine 529) of MsbA (EMD-41596). (A) ResMap local-resolution field (2.3 Å at the Cα site) on the deposited-model isosurface. (B) Build-zone map from map-only reliability (reliability score terciles within the contour mask: omit / caution / build). Auth/PDB residue numbers are used (PDB 8TSO). ResMap is essentially uniform across this loop, while Asn529 is classified omit with a Q-score of 0.16, despite local resolution at or better than the global estimate (2.68 Å).*

The utility of this approach can be observed directly on a deposited map other than the reference map: MsbA outward-facing (EMD-41596), a membrane transporter resolved at a global resolution of 2.68 Å (PDB 8TSO; fig. 3.4). In the close-up centered on Asn529, the ResMap local resolution is comparatively uniform: the same loop reports ~2.3 Å at neighboring positions and at Asn529 itself; therefore, a model builder consulting ResMap alone would have little basis for treating

this site cautiously. The reliability score, however, gives a different reading. The same view shows spatial variation in sharpness that is not apparent in the ResMap field, with Asn529 in the omit tercile (reliability score ≈ 0.17). The deposited Q-score at Asn529 was low (0.16), indicating a weak atom-level model–map agreement, as shown by the side chain not sitting within the visible density envelope. The half-map cross-correlation remained high (~0.93); therefore, the disagreement was not driven by an obvious half-map failure but by a local drop in geometric sharpness on the averaged map. The omit/caution/build assignments make this actionable by flagging specific residues for caution or omission in regions where the local resolution would not have prompted concern.

3.3 Independent Validation Against MEDIC Coordinate Corrections

To evaluate whether the reliability score predicts genuine modeling difficulties rather than merely reproducing existing validation metrics, I compared it with independently published coordinate corrections from MEDIC.[4] The MEDIC study identified residues requiring coordinate correction in deposited cryo-EM structures by comparing them against independent higher-resolution re-determinations and rebuilt models. I analyzed the original deposited maps and half-maps used in that study, computing cryoRIDGE without reference to the deposited coordinates or the MEDIC outputs. Subsequently, I asked whether residues corrected by MEDIC were preferentially assigned low reliability by the map-only metric.

Across the six analysis maps of the first validation block, MEDIC-corrected residues were enriched within the omit zone, with a median per-map odds ratio of 9.51 (Table 3.2). EMD-10366 was designated as a negative control in advance and was excluded from this median; it returned an odds ratio of 1.33 ($p = 0.47$), as expected. Using the continuous reliability score rather than discrete terciles, corrected residues were consistently ranked above uncorrected residues (median per-map AUC = 0.73), indicating that regions identified as difficult by cryoRIDGE before model

interpretation substantially overlapped with those requiring subsequent coordinate correction. A second block comparing deposited structures against MEDIC-rebuilt versions of the same entries showed a stronger effect: a median per-map odds ratio of 16.34 across six maps, with a median per-map AUC of 0.80. Accounting for clustering by map, the cluster-robust odds ratio for omit versus build was 3.09 (95% CI 1.56–6.13) in the first block and 5.74 (95% CI 1.93–17.09) in the second (Table 3.2).

3.4 Which Error Types the Reliability Score Detects

MEDIC labels each corrected residue by error type, which allows this comparison to be resolved beyond a single pooled enrichment figure. For every map, I recomputed the per-map AUC of the continuous reliability score separately against register shifts, secondary-structure errors, loop errors, and carbonyl misorientations, using the same in-mask residues and clean-residue controls as the pooled analysis (Table 3.3). Across the six analysis maps of the first block, the score separated loop errors best (median per-map AUC = 0.86, range 0.64–0.94) and secondary-structure errors next (0.82, 0.56–0.97), both above the pooled all-error value of 0.73. Register shifts were detected less reliably (0.73, 0.66–0.87), and carbonyl misorientations were weakest (0.63, 0.53–0.76).

The same ordering largely held in the rebuilt block, where loop errors again ranked highest (median AUC = 0.88 across six maps), followed by register shifts (0.82, n = 5), secondary-structure errors (0.77, n = 5), and carbonyl misorientations (0.74). Pooling both blocks (n = 12 analysis maps, the negative control excluded) yielded medians of 0.85 for loop errors, 0.82 for secondary-structure errors, 0.78 for register shifts, and 0.68 for carbonyl misorientations. On the pre-designated negative control (EMD-10366), the loop and carbonyl AUCs fell to 0.46 and 0.52, respectively, at or near the 0.5 no-discrimination line, as expected.

This breakdown is consistent with what the metric measures. Loop and secondary-structure errors arise where the backbone density envelope is weak or ambiguous, and a drop in local gradient magnitude precisely reflects the loss of geometric definition. Carbonyl misorientation, by contrast, is a sub-angstrom rotational error that can occur in density that remains locally sharp; therefore, a gradient-based score has little basis on which to flag it. Register shifts fall between these cases: the local envelope may be well defined even where the sequence assignment within it is incorrect. The practical implication is that the omit/caution/build zones are most informative for the error classes a builder is most likely to introduce, such as poorly defined loops and the ends of secondary-structure elements, and least informative for fine carbonyl and rotameric geometry, which remains the province of post-model validators.

Table 3.2: *Per-map MEDIC coordinate-correction enrichment within the omit zone. The MEDIC errors column counts residues meeting this work's operational criterion — Cα shift greater than 1 Å after alignment, joined in-mask — which is broader than the curated, error-typed total reported by MEDIC itself and is therefore not comparable to it.*

Block	EMDB	Deposited / corrected PDB	MEDIC errors	In-mask Cα	Fisher OR	Fisher p	AUC
1	EMD-23130	7L2J / 7LP9	698	1831	17.19	1.6×10^{-79}	0.84
1	EMD-7041	6B3Q / 7K1F	876	1731	3.38	8.7×10^{-23}	0.66
1	EMD-7062	6B70 / 7K1F	852	1887	3.17	1.5×10^{-12}	0.66
1	EMD-258	6HRB / 7BGY	229	1302	18.24	3.8×10^{-30}	0.79
1	EMD-10366	6T24 / 6W7V	224	351	1.33	0.47	0.53
1	EMD-30995	7E5Z / 7VW6	416	765	2.10	5.1×10^{-5}	0.64
1	EMD-10100	6S5T / 7MIR	134	795	15.64	4.9×10^{-18}	0.80
1	Headline (6 analysis maps)	—	3205	8311	9.51	—	0.73
2	EMD-9884	6JT1 / rebuilt	124	1094	10.23	1.5×10^{-18}	0.78
2	EMD-13947	7QFQ / rebuilt	406	958	70.92	2.5×10^{-44}	0.89
2	EMD-24933	7S9D / rebuilt	580	997	9.47	4.6×10^{-34}	0.75
2	EMD-3529	5MM4 / rebuilt	738	1106	2.29	2.8×10^{-7}	0.58
2	EMD-32050	7VOJ / rebuilt	50	782	∞	1.5×10^{-21}	0.93
2	EMD-8959	6E1O / rebuilt	202	1129	22.46	1.2×10^{-35}	0.83
2	Headline (6 maps)	—	2100	6066	16.34	—	0.80

Block 1 compares the deposited structures against independent re-determinations, whereas Block 2 compares the deposited structures against the MEDIC-rebuilt versions of the same entry. Odds ratios are per-map Fisher tests of MEDIC-corrected residues falling in the omit tercile versus elsewhere, and AUC uses the continuous reliability score to rank corrected against uncorrected residues. The headline rows report the median of the per-map odds ratios and the median per-map AUC across the analysis maps in each block; the MEDIC-error and in-mask Cα entries on those rows are sums across those maps, and no pooled odds ratio or Fisher p-value is reported. EMD-10366 was designated as a negative control in advance and is listed as a row but held out of the Block 1 headline medians, since a map chosen in advance to contain no correctable residues is not part of the population the enrichment estimate describes; including it lowered the Block 1 median odds ratio from 9.51 to 3.38 and the median AUC from 0.73 to 0.66. Map-level inference is given by the cluster-robust odds ratios below rather than by a residue-pooled Fisher's exact test. Cluster-robust odds ratios (omit versus build, accounting for clustering by map) were 3.09 (95% CI 1.56–6.13) for Block 1 and 5.74 (95% CI 1.93–17.09) for Block 2, respectively. EMD-32050 returned an infinite odds ratio because no corrected residue fell outside the omit zone.

Table 3.3: *Per-map AUC of the reliability score for detecting each MEDIC error class*

MEDIC error class	Block 1 median AUC	Block 2 median AUC	Pooled median AUC	Negative control (EMD-10366)
Loop	0.86	0.88	0.85	0.46
Secondary structure	0.82	0.77	0.82	0.86
Register shift	0.73	0.82	0.78	0.54
Carbonyl misorientation	0.63	0.74	0.68	0.52
All MEDIC errors	0.73	0.80	0.78	0.53

Values are medians of per-map AUCs for the continuous reliability score ranking of MEDIC-labelled errors of each class against clean in-mask residues. Block 1 comprises the six deposited-versus-corrected analysis maps (per-map ranges: loop 0.64–0.94; secondary structure 0.56–0.97; register shift 0.66–0.87; carbonyl 0.53–0.76); Block 2 comprises the six deposited-versus-rebuilt maps, for which register-shift and secondary-structure AUCs were computable on five. Thirteen maps returned usable labels overall: seven in Block 1, including the negative control, and six in Block 2. The pooled column combines the twelve analysis maps from both blocks. EMD-10366 was designated a negative control in advance and was excluded from the Block 1 and pooled medians.

3.5 LOMO Cross-Validation Confirms Generalization

Leave-one-map-out cross-validation was performed as a conservative check against overfitting, given that the reliability score thresholds are fixed terciles that require no map-specific training. The held-out AUC distribution for the reliability score (mean ≈ 0.77, median ≈ 0.82) remained well above chance across the cohort, and in the like-for-like continuous comparison, it outperformed local resolution. Ranking residues by the continuous reliability score and by local resolution under identical folds, the held-out median AUC is 0.82 for the continuous reliability score against 0.73 for a ResMap worse-than-in-map-median flag and 0.61 for BlocRes, with the reliability score higher on 23 of the 31 held-out maps (two-sided Wilcoxon $p = 0.019$). The corresponding one-sided continuous-versus-continuous comparison against ResMap is reported above. A separate head-to-head export with intact ResMap coverage gives the same ordering: across 32 held-out maps, the continuous reliability score retains a median balanced accuracy of 0.74 and a median AUC of 0.83, against 0.69 and 0.73 for ResMap worse-than-in-map-median flags. A second, stricter comparison substitutes the discrete omit tercile flag for the continuous score and a worse-than-in-map-median split for ResMap. Under that coarser rule, on the 27-map held-out subset with complete BlocRes,

ResMap, and MonoRes coverage — the 33-map core cohort less six maps failing quality control — ResMap was the stronger classifier (median AUC 0.81 versus 0.75 for the omit flag; median balanced accuracy 0.70 versus 0.64, one-sided p = 0.29), while held-out rank recovery was equivalent (median ρ = 0.46 versus 0.45) and BlocRes (0.19) and MonoRes (0.15) trailed both. The same pattern appears under a simple median split: scored by balanced accuracy over the same 31 folds — flagging a residue when the reliability score or ResMap falls below the training-fold median — the reliability-based flag (0.58) is only modestly higher than the ResMap-based flag (0.55), a difference that does not reach significance (one-sided p = 0.11, n = 31). These are not independent tests but alternative readings of a single cross-validation, and together they locate the effect: the reliability score separates hard-to-place residues well when its full range is used, and much of that separation is given up whenever it is reduced to a coarse label, whether a three-level build zone or a binary threshold level. The wide per-map distributions reflect genuine structural variability (i.e., some maps are simply more predictable than others); however, the central tendency is stable and consistent with the non-cross-validated results.

It is important to be precise about what these results show and do not show. The reliability score performed comparably to the windowed half-map CC on most operational metrics but was not substantially better. The incremental value of the reliability score after conditioning on variance, CC, and local resolution together is modest: a median held-out ΔR^2 of +0.037 for Q-score prediction under LOMO cross-validation. This is the median gain in held-out R^2 obtained by adding the reliability score to a model that already contains those three predictors, evaluated on each cohort map in turn while it is held out of the fit (see Methods, Leave-One-Map-Out Cross-Validation). This gain is small but consistent: within the 2.5–4 Å band, adding the reliability score to the variance + CC + local-resolution baseline improved the held-out prediction in 28 of the 34 maps in that band that carry a local-resolution baseline and a sufficient sample for testing (sign-test p = 2.0×10^{-4};

Table 3.4). Against the narrower CC + local-resolution baseline, the reliability score is also the strongest single addition: compared head-to-head as candidate additions to that baseline (n = 34 maps with a local-resolution baseline), the reliability score's own ΔR^2 (0.074) exceeds that of local variance (0.029) on 30 of 34 maps with a large effect size (Wilcoxon $p = 3.2 \times 10^{-6}$, Cohen's $d_z = 0.88$; Table 3.4). Therefore, the reliability score's value is not that it outperforms other half-map statistics as a sole predictor, but that it is the half-map statistic most directly aligned with regions where the density does not adequately constrain placement, and that it is the only one of these statistics exported directly as a per-voxel reliability score and a build-zone volume that the builder can load alongside the map during interpretation.

Table 3.4: *Incremental held-out variance explained (ΔR^2) for Q-score prediction under leave-one-map-out cross-validation*

Predictor added	Baseline	Resolution bin	Maps	Median ΔR^2	Mean ΔR^2	Maps improved	Sign-test p
Local variance	CC + local resolution	$\leq$ 2.5 Å	5	0.059	0.073	5/5	0.063
Local variance	CC + local resolution	2.5–4 Å	34	0.029	0.015	24/34	0.024
Reliability score	CC + local resolution	$\leq$ 2.5 Å	5	0.064	0.041	4/5	0.375
Reliability score	CC + local resolution	2.5–4 Å	34	0.074	0.064	29/34	3.9×10^{-5}
Reliability score	variance + CC + local resolution	$\leq$ 2.5 Å	5	0.020	–0.032	4/5	0.375
Reliability score	variance + CC + local resolution	2.5–4 Å	34	0.037	0.045	28/34	2.0×10^{-4}

ΔR^2 is the gain in held-out R^2 from adding the named predictor to the stated baseline, evaluated on each cohort map in turn while that map is held out of the fit. The 2.5–4 Å band is the atomic-building regime this thesis focuses on and is the only bin with enough maps to support a significance test; the $\leq$ 2.5 Å, 4–6 Å, and $>$ 6 Å bins contain 5, 3, and 2 maps, respectively. These counts are restricted to maps carrying a local-resolution baseline and therefore sum to 44 rather than to the full 55-map analysis cohort; the n = 42 reported for ρ(Q, reliability score) in the same band is the larger set of maps returning a finite correlation, which requires no local-resolution baseline. Compared head-to-head as candidate additions to the same CC + local-resolution baseline across the 34 maps in the 2.5–4 Å band, the reliability score's ΔR^2 exceeds local variance's on 30 of 34 maps (Wilcoxon p = 3.2×10^{-6}, Cohen's dz = 0.88).

3.6 Why the Reliability Score Correlates With B-factor

The strong negative correlation between the reliability score and deposited isotropic B-factors (cohort-median $\rho \approx -0.82$; per-map values in Table **??**) follows directly from the relationship between atomic displacement[28] and the local gradient structure of the density.

In crystallographic and cryo-EM model refinement, the contribution of atom i to the scattering density is controlled by the Debye–Waller factor, which describes how thermal motion and positional disorder smear the expected density peak. For an atom with isotropic B-factor B_i centered at position $\mathbf{r}_i$, the density contribution falls off following this function.

$$\rho\left(\mathbf{r}\right) \propto \exp\left(-\frac{4\pi^2}{B_i}\left|\mathbf{r}-\mathbf{r}_i\right|^2\right)$$

This is a Gaussian whose width is directly controlled by B_i. By taking the gradient of this expression, the following equation appears:

$$|\nabla\rho(\mathbf{r})| \propto \frac{8\pi^2}{B_i}\,|\mathbf{r}-\mathbf{r}_i|\exp\left(-\frac{4\pi^2}{B_i}\,|\mathbf{r}-\mathbf{r}_i|^2\right)$$

The peak gradient magnitude occurs at a radial distance from the atomic center that scales with the width of the Gaussian, and its value scales inversely with that width. In other words, a low B_i corresponds to a narrow, sharp Gaussian and therefore to a steep gradient around the peak, yielding a high reliability score. Conversely, a high B_i corresponds to a broad, flat Gaussian and therefore to a shallow gradient around the peak, yielding a low reliability score.

Therefore, reliability score and B-factor are expected to be negatively correlated, and the strength of the correlation reflects how faithfully the gradient structure of the map encodes the underlying atomic displacement field. The correlation is strong but not perfect ($\rho \approx -0.82$ instead of -1.0) because deposited B-factors in cryo-EM are not purely physical quantities. They strongly reflect refinement and model-specific choices made during structure determination. B-factors therefore have limited direct physical interpretability.[29] For this reason, B-factor comparisons are treated as a supplement and not a primary validation result, and the circularity introduced by comparing the reliability score (computed from the map) against B-factors (derived from a model refined into the same map) is noted explicitly.

Chapter 4

DISCUSSION

This work was motivated by the question of whether map-only statistics computed from deposited maps and half-maps can flag regions where downstream model validation will report poor placement before a model is created. Standard placement validators, such as Q-score and EMRinger, answer this question only after fitting and refinement;[16,20] local-resolution estimators, such as BlocRes, ResMap, and RELION, instead localize Fourier-domain reproducibility, which is related to but not identical to placement confidence.[10,11,13] To address this gap, I developed cryoRIDGE, which estimates a windowed squared-gradient sharpness field, termed the reliability score, on the half-map average and exports continuous reliability scores and build-zone volumes for interactive model building. Across a 55-map cohort, the reliability score tracks per-residue Q-score and EMRinger more strongly than BlocRes local resolution and, on cohort-level ROC for the operational threshold $Q < 0.5$, more strongly than ResMap local resolution when both are scored continuously, without using the deposited coordinates, B-factors, or external locres binaries. That continuous-score advantage does not survive reduction to a coarse label: under a three-level build zone against a ResMap median split, ResMap is the stronger classifier (see Results, LOMO Cross-Validation Confirms Generalization). The central finding is that map-only geometric sharpness is a usable pre-model predictor of placement outcomes in the 2.5–4 Å regime, where most building occurs. Current cryo-EM workflows rely on post hoc validation after an atomic model has been built.[8] This work demonstrates that meaningful placement confidence can be estimated directly from the map itself. The method provides a complementary pre-model measure that identifies regions likely to challenge model builders and offers an additional source of guidance during atomic interpretation.

The four local resolution estimators tested here do not behave interchangeably, and the spread between them is sufficiently large to change how their results should be interpreted. Against the Q-score, RELION (median $\rho = +0.46$) and ResMap (+0.39) track deposited model quality several times more closely than BlocRes (+0.14) or MonoRes (+0.14), and the same ordering holds against EMRinger and deposited B-factors. All four answer the same nominal question, but they differ in how the local neighborhood is defined, how masking and noise are handled, and what significance criterion is applied; those implementation choices evidently matter more for model-building guidance than the shared Fourier-space framing would suggest.[11,13,14] On the 31 maps where BlocRes, ResMap, and RELION were all available, the reliability score's median advantage is +0.35 over BlocRes, +0.09 over RELION, and +0.02 over ResMap, all significant by paired Wilcoxon test but very different in practical size. The estimators differ in how reliably they can be run at all, and ResMap produced no usable output on two cohort maps while yielding a finite Q correlation on only 47 of 55, against 54 for RELION; therefore, its coverage across the map types is less predictable. Taken together, the case for the reliability score rests less on measuring something that local resolution cannot and more on supplying one quantity, computed identically from the half-maps of every map, which is framed around placement from the outset.

4.1 What the Reliability Score Does & Does Not Offer

The incremental value of the reliability score after conditioning on variance, CC, and local resolution was modest (median held-out $\Delta R^2 = +0.037$ for Q-score prediction), although this gain proved statistically consistent rather than noise across the 34 maps of the 2.5–4 Å band (see Results: LOMO Cross-Validation Confirms Generalization). The reliability score did not substantially outperform the windowed half-map CC as a standalone predictor. The two predictors are comparable in terms of enrichment, classification accuracy, and rank recovery. This comparability is worth

stating plainly rather than minimizing because it sharpens the argument that follows rather than undermining it. The windowed half-map CC is a reproducibility proxy by construction, but the cohort data show that the map-only predictors do not partition along the reproducibility/placement line that the labels imply. They are partitioned by domain: the real-space local statistics — the reliability score, windowed CC (median $\rho = +0.84$), and local variance ($+0.91$) — cluster together and all track the Q-score, while the frequency-domain local-resolution estimates diverge among themselves, with BlocRes (-0.19) and MonoRes (-0.23) essentially decoupled from the reliability score and only weakly related to the Q-score ($+0.14$ each). Therefore, the windowed CC behaves like the real-space density-structure measure from which it is computed, not like the FSC-based estimators with which it is nominally categorized. Accordingly, the claim supported by this work is narrower and better evidenced than the statement that reproducibility and placement are separate axes: placement confidence is carried by the local real-space density structure, and windowed FSC estimates recover it only partially and for only some implementations. That two independently motivated real-space statistics, the gradient energy on the half-map average and the Pearson agreement between the two half-maps, converge on the same residues corroborates that the placement axis is a property of the density rather than an artifact of one particular formula.

Against that background, the contribution of the reliability score is specific rather than sweeping: it carries the larger incremental signal in a head-to-head test ($\Delta R^2 = 0.074$ versus 0.029 for local variance, favored in 30 of the 34 maps; Wilcoxon $p = 3.2 \times 10^{-6}$, Cohen's $d_z = 0.88$; Table 3.4), it is computed from the half-map average as a single scalar field rather than requiring the half-map pair as a correlation, and it is the only one of these statistics exported as a per-voxel volume and omit/caution/build zones that a builder can load alongside the map before any model exists. The roughly 30% of rank agreement not shared between the reliability score and CC (median $\rho = +0.84$) also marks where the sharper test lies: locating regions where the two

disagree and establishing which is correct against deposited placement quality would convert a comparability result into a discriminating one, and is the natural next step. A second derivative (Hessian) alternative was also screened: the minimum Hessian eigenvalue correlated with the Q-score at a comparable magnitude but with an inverted sign and no gain in the high-reliability score/low-Q regime, where a placement flag was most needed (Table **??**). Therefore, the windowed gradient energy was retained as the operational pre-model score.

The resolution ceiling deserves a more precise statement than a single cutoff implies. The association with the Q-score is strongest at or below 2.5 Å (n = 6, median $\rho \approx 0.71$), remains the primary reported association within the 2.5–4 Å model-building range (n = 42 maps with a finite ρ, median $\rho \approx 0.54$), weakens substantially but is not absent in the 4–6 Å range (n = 5, median $\rho \approx 0.29$, with wide heterogeneity), and is not supported above approximately 6 Å (n = 2, median $\rho \approx 0.02$). This graded falloff is not a limitation specific to the reliability score; it reflects the diminishing discriminative power of Q-scores themselves at lower resolution, where map features become too coarse for either metric to distinguish easy from difficult placement at the residue scale.[16,19] Therefore, the 2.5–4 Å band is the primary operating regime for this metric, not a hard boundary beyond which it abruptly ceases to function.

4.2 B-factor Correlation & Interpretation

The strong negative correlation between the reliability score and deposited isotropic B-factors[28] (cohort-median $\rho \approx -0.82$) follows from the Debye–Waller relationship between atomic displacement and gradient magnitude derived in the Results section (Why the Reliability Score Correlates With B-factor): a sharper, more constrained density around an atom implies both a lower B-factor and a higher reliability score; the two quantities are therefore expected to vary inversely. This correlation is presented as a supplementary consistency check rather than a primary validation

result for the two reasons that follow. First, B-factors absorb refinement- and model-specific choices beyond pure atomic disorder, limiting their direct physical interpretation.[29] Second, B-factors are derived from a model built into the map from which the reliability score is computed, which makes the relationship partly circular by design. Read this way, the B-factor/reliability-score coupling is best understood as confirming that the reliability score tracks genuine, physically meaningful density sharpness, rather than as independent evidence for the placement guidance claims made elsewhere in this thesis. These claims are based on the Q-score, which is derived from an independently motivated validation procedure and remains the primary comparison.

4.3 Limitations

This tool has several limitations. The build zones are fixed at terciles; therefore, approximately one-third of the voxels in every map are labeled omit, regardless of the map's overall quality; it is unlikely that one-third of every deposited map is genuinely unbuildable. The tercile split is deliberately assumption-free and requires no per-map calibration, but a user-configurable cut (i.e., flagging the lowest 10% of residues, or thresholding on an absolute reliability value rather than a within-map quantile) would better reflect how map quality varies between depositions and is a straightforward extension. The cohort of 55 maps, although diverse in resolution, size, and composition, was not large enough to draw strong conclusions about the performance of the metric within specific structural niches. LOMO cross-validation addresses overfitting concerns but cannot account for systematic biases in the cohort, such as the underrepresentation of very large assemblies and of nucleic-acid-containing structures, which enter the cohort only through the ribosomal entries discussed below; because no parameters are learned, cross-validation cannot correct for cohort composition. The 4 Å figure is best read as the edge of the regime rather than a failure point, and it does not limit the metric's practical reach, since most of the field's structurally challenging recent depo-

sitions, including megacomplexes, membrane scaffolds, and partial ribosomal assemblies, are solved within the 2.5–4 Å atomic-building regime. An expansion cohort assembled specifically around this kind of target (n = 22, median global resolution 2.96 Å) showed a median ρ(Q, reliability score) of 0.50, close to the core cohort's 0.54 and the full-cohort median of 0.53; megacomplex entries alone reached a median $\rho = 0.72$, including in situ human 80S ribosomes at 2.6–2.8 Å ($\rho \approx 0.81$–0.88), which also provide the cohort's principal nucleic-acid-containing test cases. Therefore, the metric continues to perform on exactly the kind of large, compositionally complex target it was designed for, on targets whose difficulty is compositional rather than resolution-limited. The 4 Å figure marks where the association weakens (median $\rho \approx 0.29$ across the five maps in the 4–6 Å band) rather than where it disappears; above approximately 6 Å it is not supported by the available data. Finally, all validations against Q-scores carry the implicit assumption that Q-scores are a reliable ground truth for placement confidence. This assumption is well supported in the literature[19] within the atomic-building regime, but it weakens near the 4 Å resolution boundary and beyond, where the discriminative power of the Q-score decreases at worse resolutions.

4.4 Future Directions

Several natural extensions of this work are apparent. The most immediately impactful open question, the connection between omit zones and post-hoc error detection, is addressed in this work: MEDIC-corrected residues are enriched in omit zones across two independent validation blocks (see Results, Independent Validation Against MEDIC Coordinate Corrections). This analysis does not establish that omit-zone residues are generally correctable. A residue may be flagged because the density genuinely does not constrain its placement (e.g., through conformational flexibility, thermal motion, or compositional heterogeneity in the particle population), in which case the deposited coordinates may already be the best model supported by the data, and rebuilding would not im-

prove them. The reliability score identifies residues whose placement is not determined by the map; separating those that are actually misplaced from those that are simply unresolvable requires information that the map alone does not carry. A prospective study that distinguishes the two by rebuilding flagged residues and recording which corrections measurably improve fit is the natural next step.

Chapter 5

CONCLUSION

A deposited map can report near-atomic global resolution while local placement quality remains heterogeneous, and the validators that quantify this heterogeneity require a fitted model that does not yet exist when the first modeling decisions are made. The central question was therefore whether a real-space readout, computed directly from the map and half-maps in voxel coordinates, could supply model-building guidance before any atomic coordinates are placed. This work shows that such a readout is feasible. cryoRIDGE computes a windowed squared-gradient sharpness field — the reliability score — on the half-map average: a purely real-space, pre-model quantity, and converts it to continuous reliability scores and omit/caution/build zones that track downstream placement confidence across 55 structures (median ρ(Q, reliability score) $\approx +0.53$; median per-map AUC ≈ 0.82 for Q < 0.5). The contribution is not a replacement for the Q-score or local resolution, but a complementary placement-confidence layer that answers the question: where, in real space, should a builder expect placement to be difficult? These results show that placement confidence is carried by the local real-space density structure, which windowed FSC estimates capture inconsistently across implementations, and demonstrate that a map-only, pre-model gradient metric provides a more direct route to placement guidance than local resolution alone.

The reliability score derived from the gradient magnitude field translated this finding into practice. Omit zones recover low-Q residues at roughly twice the rate expected by chance, and the reliability score achieves a median per-map AUC of 0.82 against Q < 0.5, compared to 0.71 for continuous ResMap local resolution on the 31-map paired panel and 0.80 on the subset with usable ResMap coverage. This advantage was maintained under leave-one-map-out cross-validation

when the continuous score was used, indicating that the ranking generalizes beyond the specific maps used for its characterization; however, it did not hold under the coarser omit-tercile label. Simultaneously, the reliability score has limitations: it performs comparably to the windowed half-map CC on most operational metrics, and its incremental predictive value after accounting for the variance, CC, and local resolution together is modest. The reliability score's contribution is not that it is the single best predictor available, but that it directly targets the placement axis in a way that the field's default tool, local resolution, does not.

The distinct question of whether a builder can confidently place an atom before any model exists requires its own instrument, and this work provides one: it is simple to compute, requiring only the half-maps already produced by any standard reconstruction pipeline, and is validated against the field's own most widely used placement-confidence reference, the Q-score.

REFERENCES

(1) Kühlbrandt, W. Biochemistry. The Resolution Revolution. *Science* **2014**, *343*, 1443–1444.

(2) Bai, X.-c.; McMullan, G.; Scheres, S. H. W. How Cryo-EM Is Revolutionizing Structural Biology. *Trends in Biochemical Sciences* **2015**, *40*, 49–57.

(3) Lamb, A. L.; Kappock, T. J.; Silvaggi, N. R. You Are Lost without a Map: Navigating the Sea of Protein Structures. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2015**, *1854*, 258–268.

(4) Reggiano, G.; Lugmayr, W.; Farrell, D.; Marlovits, T. C.; DiMaio, F. Residue-Level Error Detection in Cryoelectron Microscopy Models. *Structure* **2023**, *31*, 860–869.e4.

(5) Croll, T. I.; Diederichs, K.; Fischer, F.; Fyfe, C. D.; Gao, Y.; Horrell, S.; Joseph, A. P.; Kandler, L.; Kippes, O.; Kirsten, F., et al. Making the Invisible Enemy Visible. *Nature Structural & Molecular Biology* **2021**, *28*, 404–408.

(6) Rosenthal, P. B.; Henderson, R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology* **2003**, *333*, 721–745.

(7) Chojnowski, G.; Sobolev, E.; Heuser, P.; Lamzin, V. S. The Accuracy of Protein Models Automatically Built into Cryo-EM Maps with ARP/wARP. *Acta Crystallographica Section D Structural Biology* **2021**, *77*, 142–150.

(8) Lawson, C. L.; Kryshtafovych, A.; Adams, P. D.; Afonine, P. V.; Baker, M. L.; Barad, B. A.; Bond, P.; Burnley, T.; Cao, R.; Cheng, J., et al. Cryo-EM Model Validation Recommendations Based on Outcomes of the 2019 EMDataResource Challenge. *Nature Methods* **2021**, *18*, 156–164.

(9) Ramírez-Aportela, E.; Vilas, J. L.; Glukhova, A.; Melero, R.; Conesa, P.; Martínez, M.; Maluenda, D.; Mota, J.; Jiménez, A.; Vargas, J., et al. Automatic Local Resolution-Based Sharpening of Cryo-EM Maps. *Bioinformatics* **2020**, *36*, 765–772.

(10) Vilas, J. L.; Heymann, J. B.; Tagare, H. D.; Ramirez-Aportela, E.; Carazo, J. M.; Sorzano, C. O. S. Local Resolution Estimates of cryoEM Reconstructions. *Current opinion in structural biology* **2020**, *64*, 74–78.

(11) Kucukelbir, A.; Sigworth, F. J.; Tagare, H. D. Quantifying the Local Resolution of Cryo-EM Density Maps. *Nature Methods* **2014**, *11*, 63–65.

(12) Kartte, D.; Sachse, C. Robust Quality Assessment of Cryo-EM Maps, Tomograms and Micrographs by Statistics-Based Local Resolution Estimation, 2026.

(13) Cardone, G.; Heymann, J. B.; Steven, A. C. One Number Does Not Fit All: Mapping Local Variations in Resolution in Cryo-EM Reconstructions. *Journal of Structural Biology* **2013**, *184*, 226–236.

(14) Vilas, J. L.; Gómez-Blanco, J.; Conesa, P.; Melero, R.; Miguel De La Rosa-Trevín, J.; Otón, J.; Cuenca, J.; Marabini, R.; Carazo, J. M.; Vargas, J., et al. MonoRes: Automatic and Accurate Estimation of Local Resolution for Electron Microscopy Maps. *Structure* **2018**, *26*, 337–344.e4.

(15) Fernandez-Leiro, R.; Scheres, S. H. W. A Pipeline Approach to Single-Particle Processing in RELION. *Acta Crystallographica Section D: Structural Biology* **2017**, *73*, 496–502.

(16) Pintilie, G.; Zhang, K.; Su, Z.; Li, S.; Schmid, M. F.; Chiu, W. Measurement of Atom Resolvability in Cryo-EM Maps with Q-scores. *Nature Methods* **2020**, *17*, 328–334.

(17) Afonine, P. V.; Klaholz, B. P.; Moriarty, N. W.; Poon, B. K.; Sobolev, O. V.; Terwilliger, T. C.; Adams, P. D.; Urzhumtsev, A. New Tools for the Analysis and Validation of Cryo-EM Maps and Atomic Models. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 814–840.

(18) Pintilie, G. Diagnosing and Treating Issues in Cryo-EM Map-Derived Models. *Structure* **2023**, *31*, 759–761.

(19) Pintilie, G.; Shao, C.; Wang, Z.; Hudson, B. P.; Flatt, J. W.; Schmid, M. F.; Morris, K. L.; Burley, S.; Chiu, W. Q-Score as a Reliability Measure for Protein, Nucleic Acid and Small-Molecule Atomic Coordinate Models Derived from 3DEM Maps. *Acta Crystallographica Section D: Structural Biology* **2025**, *81*, 410–422.

(20) Barad, B. A.; Echols, N.; Wang, R. Y.-R.; Cheng, Y.; DiMaio, F.; Adams, P. D.; Fraser, J. S. EMRinger: Side Chain–Directed Model and Map Validation for 3D Cryo-Electron Microscopy. *Nature Methods* **2015**, *12*, 943–946.

(21) Rudin, L. I.; Osher, S.; Fatemi, E. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D: Nonlinear Phenomena* **1992**, *60*, 259–268.

(22) Spearman, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* **1904**, *15*, 72–101.

(23) Wu, T.; Yoshida, S.; Akiyama, Y.; Stock, M.; Ushio, T.; Kawasaki, Z. Preliminary Breakdown of Intracloud Lightning: Initiation Altitude, Propagation Speed, Pulse Train Characteristics, and Step Length Estimation. *Journal of Geophysical Research: Atmospheres* **2015**, *120*, 9071–9086.

(24) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognition Letters* **2006**, *27*, 861–874.

(25) Han, X.; Terashi, G.; Christoffer, C.; Chen, S.; Kihara, D. VESPER: Global and Local Cryo-EM Map Alignment Using Local Density Vectors. *Nature Communications* **2021**, *12*, 2090.

(26) Zeinert, R.; Zhou, F.; Franco, P.; Zöller, J.; Madni, Z. K.; Lessen, H.; Aravind, L.; Langer, J. D.; Sodt, A. J.; Storz, G., et al. P-Type ATPase Magnesium Transporter MgtA Acts as a Dimer. *Nature Structural & Molecular Biology* **2025**, *32*, 1633–1643.

(27) Ramírez-Aportela, E.; Maluenda, D.; Fonseca, Y. C.; Conesa, P.; Marabini, R.; Heymann, J. B.; Carazo, J. M.; Sorzano, C. O. S. FSC-Q: A CryoEM Map-to-Atomic Model Quality Validation Based on the Local Fourier Shell Correlation. *Nature Communications* **2021**, *12*, 42.

(28) Trueblood, K. N.; Bürgi, H. B.; Burzlaff, H.; Dunitz, J. D.; Gramaccioli, C. M.; Schulz, H. H.; Shmueli, U.; Abrahams, S. C. Atomic Displacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A Foundations of Crystallography* **1996**, *52*, 770–781.

(29) Wlodawer, A.; Li, M.; Dauter, Z. High-Resolution Cryo-EM Maps and Models: A Crystallographer's Perspective. *Structure* **2017**, *25*, 1589–1597.e1.

APPENDIX. Supplementary Material

A.1 Cohort and Map Details

Table A.1 lists the full analysis cohort of 55 cryo-EM maps used in this thesis, including the EMDB accession number, protein identity, global resolution, and whether a deposited PDB model and Q-score were available. The maps were sorted by global resolution. Structural class assignments followed the categories analyzed in the Results section and were made by manually inspecting the deposited structures and the associated publications.

Table A.1: *Full cohort manifest*

EMD ID	Protein	Global Resolution (Å)	Has PDB / Q-score
EMD-11638	Apoferritin	1.22	Yes / Yes
EMD-61596	HSV-2 gB (+Fab16F9)	2.18	Yes / Yes
EMD-50549	Aerolysin WT in SMALP	2.2	Yes / Yes
EMD-47552	YnaI mechanosensitive channel (DDPC nanodisc)	2.3	Yes / Yes
EMD-24120	70S pre-translocation (E. coli)	2.33	Yes / Yes
EMD-51569	Human muscle nAChR in nanodisc	2.48	Yes / Yes
EMD-52525	Complex III (mitochondrial)	2.52	Yes / Yes
EMD-37504	GLP-1R + Gs (ligand-free)	2.54	Yes / Yes
EMD-51902	Human myoferlin in nanodisc	2.56	Yes / Yes
EMD-71338	Human 80S hibernating class4	2.57	Yes / Yes
EMD-48923	MgtA (E2.Mg.BeF3)	2.59	Yes / Yes
EMD-71435	Human 80S + EBP1 in situ	2.67	Yes / Yes
EMD-41596	MsbA outward-facing	2.68	Yes / Yes
EMD-29275	Human pre-60S state L1 composite	2.7	Yes / Yes
EMD-49450	MgtA (E2P+E1)	2.73	Yes / Yes
EMD-6287	T20S proteasome	2.8	Yes / Yes
EMD-51377	Human SLC45A4 in detergent	2.8	Yes / Yes
EMD-62841	Thermophile spliceosome ILS	2.8	Yes / Yes
EMD-71331	Human 80S P-E in situ	2.8	Yes / Yes
EMD-25418	70S post-translocation (E. coli)	2.9	Yes / Yes
EMD-29262	Human pre-60S state G composite	2.9	Yes / Yes
EMD-41756	LRRK2 CTD (+GZD824)	2.9	Yes / Yes
EMD-39332	Human 20S proteasome assembly intermediate	2.91	Yes / Yes
EMD-44471	Human nucleosome 3.0 A	3.0	Yes / Yes
EMD-53474	Human NTCP in nanodisc	3.0	Yes / Yes
EMD-53483	Rat SLCO2A1 in detergent-lipid micelle	3.0	Yes / Yes
EMD-54253	70S + EF-G hybrid H1	3.0	Yes / Yes
EMD-48534	MgtA (E2P.Mg)	3.07	Yes / Yes
EMD-33734	SARS-CoV-2 spike K202B bispecific	3.1	Yes / Yes
EMD-19872	Integrator INTS5/8/15	3.2	Yes / Yes
EMD-51351	70S + SaEF-G POST	3.2	Yes / Yes
EMD-42603	Human p97/VCP + inhibitor	3.23	Yes / Yes
EMD-16091	Beta-galactosidase (5ms time-resolved)	3.3	Yes / Yes
EMD-16119	GroEL-GroES ADP (turnover)	3.4	Yes / Yes
EMD-13308	GroEL-GroES ADP (tight)	3.43	Yes / Yes

EMD ID	Protein	Global Resolution (Å)	Has PDB / Q-score
EMD-11497	SARS-CoV-2 spike 3 RBD-down	3.5	Yes / Yes
EMD-45604	beta2AR inactive (carazolol)	3.5	Yes / Yes
EMD-64133	26S proteasome-Midnolin MB state	3.5	Yes / Yes
EMD-41598	MsbA inward-facing	3.6	Yes / Yes
EMD-11149	Bovine ATP synthase Fo	3.61	Yes / Yes
EMD-74099	Human U2/BP spliceosome SF3A	3.61	Yes / Yes
EMD-23130	TRPV1 pH 6c	3.66	Yes / Yes
EMD-23129	TRPV1 pH 6a	3.7	Yes / Yes
EMD-48311	Yeast V-ATPase Vo	3.7	Yes / Yes
EMD-7130	Epithelial Na+ channel (ENaC)	3.9	Yes / Yes
EMD-18839	Integrator cleavage module intermediate	3.9	Yes / Yes
EMD-28487	Eag1 voltage sensor up	3.9	Yes / Yes
EMD-45603	beta2AR apo	3.95	Yes / Yes
EMD-4941	ClpB WT-2A (ATPgammaS)	4.0	Yes / Yes
EMD-50267	Integrator arm module state 1	4.0	Yes / No
EMD-41042	S. enterica WbaP in SMALP	4.1	Yes / Yes
EMD-64103	26S proteasome-Midnolin MA state	4.32	Yes / Yes
EMD-28498	Eag1 voltage sensor down	5.4	Yes / Yes
EMD-4940	ClpB WT-1 (ATPgammaS)	6.2	Yes / Yes
EMD-9156	NMDA GluN1-GluN2A Extended-1	6.84	Yes / Yes

VITA

Sarthak Mohanty earned his Bachelor of Science in Biochemistry from The University of Texas at San Antonio in August 2026. He began research in an organic chemistry group on campus, after he learned small-molecule X-ray crystallography with Dr. Hadi Arman and served as a teaching assistant for Dr. Arman.

He continued his research career in Dr. Maria Falzone's laboratory, working with Dr. Falzone and Dr. Rahul Mojidra on the structure and function of phospholipase C epsilon (PLCε), a signaling protein frequently dysregulated in cancers. That work used cryogenic electron microscopy and fluorescence spectroscopy to study how phospholipases regulate cellular signaling pathways and how those mechanisms might eventually be targeted therapeutically. He was co-advised by Dr. Falzone and Dr. Arman. After graduation, he is pursuing an MD–PhD physician-scientist program.