Scale-dependent hierarchical resolution: applications to atomic resolution and model validation in cryoEM

Korak Kumar Ray[1], Colin D. Kinz-Thompson[2,*]

[1]*Single Molecule Imaging Group, MRC-London Institute of Medical Sciences, London W12 0NN, UK.*

[2]*Department of Chemistry, Rutgers University-Newark, Newark, NJ 07102.*

**Correspondence email: colin.kinzthompson@rutgers.edu*

Significance Statement

Information about biomolecular structure is very useful to researchers investigating the mechanistic basis of biomolecular function. To enable researchers to assess the reliability of biomolecular structural information that is derived from experimental data, a new theoretical definition of atomic resolution has been developed. This definition forms the basis of a machine learning-based tool called HARP that can be used to measure the local atomic resolution of cryoEM maps, as well as the quality of the biomolecular structural models derived from such cryoEM maps. To demonstrate the power and utility of these conceptually novel approaches, this work demonstrates how they may be used to investigate the scientific and social factors that have historically affected the quality of cryoEM experiments.

Abstract

The recent cryoEM resolution revolution has had a tremendous impact on our ability to investigate biomolecular structure and function. However, outstanding questions about the reliability of using a cryoEM-derived molecular model for interpreting experiments and building further hypotheses limit its full impact. Significant amounts of research have been focused on developing metrics to assess cryoEM model quality, yet no consensus exists. This is in part because the meaning of cryoEM model quality is not well defined. In this

work, we formalize cryoEM model quality in terms of whether a cryoEM map is better described by a model with localized atomic coordinates or by a lower-resolution model that lacks atomic-level information. This approach emerges from a novel, quantitative definition of image resolution based upon the hierarchical structure of biomolecules, which enables computational selection of the length scale to which a biomolecule is resolved based upon the available evidence embedded in the experimental data. In the context of cryoEM, we develop a machine learning-based implementation of this framework, called hierarchical atomic resolution perception (HARP), for assessing local atomic resolution in a cryoEM map and thus evaluating cryoEM model quality in a theoretically and statistically well-defined manner. Finally, using HARP, we perform a meta-analysis of the cryoEM-derived structures in the Protein Data Bank (PDB) to assess the state of atomic resolution in the field and quantify factors that affect it.

Introduction

Cryogenic electron microscopy (cryoEM) has greatly improved our ability to obtain high quality structural models of biomolecules (1). In a typical cryoEM study, images of a biomolecule embedded in vitreous ice are acquired with a transmission electron microscope. Those images are then combined in a single-particle analysis (SPA) to generate a three-dimensional 'cryoEM map' that contains information about the electrostatic potential of the biomolecule (2, 3). Because the electrostatic potential depends on the atomic coordinates of the biomolecule, researchers can use a cryoEM map to infer those atomic coordinates and create a molecular model of the biomolecule being imaged (4). Recent technological and computational advances in cryoEM methodology have significantly increased the quality of the cryoEM maps generated from SPA, which has consequently increased the accuracy and precision with which atomic coordinates are inferred from cryoEM maps (5). This improvement in cryoEM map quality is generally described as an increase in the 'resolution' of the cryoEM maps—higher quality cryoEM maps contain more information about smaller structural features and are thus considered to be of higher resolution. In this light, the rapid increase in the ability of SPA cryoEM to yield mechanistically useful information about biomolecular structure has been hailed as a "resolution revolution" (6).

On the other hand, the cryoEM resolution revolution has been coupled with debate on the exact relationship between the meaning of 'resolution' and the ability to identify biomolecular structural information in a cryoEM map (7). A corollary to this debate is that we lack an exact understanding of how the quality of a cryoEM map is related to the quality of the inferred structural model. Many metrics that assess, *e.g.*, local map resolution (8–13) or map-to-model quality (14–17), have been developed to provide insight into these questions (18). Despite that work, it still remains unclear how grounded the assignment of atomic coordinates for a biomolecular structural model is in the local experimental evidence present in a cryoEM map. The answers to these questions are important not just for the structural biologist assessing the quality of their work, but also for the researchers who use biomolecular structural information as the basis for developing molecular hypotheses and mechanistic models in their own studies (*e.g.*, as

highlighted in Ref. (19)). In this work, we have answered these questions by developing a theoretical framework to quantify atomic resolution in structural biology experiments, and then applying that framework to quantify the quality of SPA cryoEM experiments and the reliability of the resulting biomolecular structural models.

The major complication addressed in this work is that the meaning of 'atomic resolution' is itself ambiguous; even amongst structural biologists, there is not a consensus on its exact definition (20, 21). In part, this is because the spatial resolution of a cryoEM map (*i.e.*, the specific lengths scales covered by the data) and its structural information content (*i.e.*, the specific molecular features captured in the data) are two distinct concepts that both affect biomolecular structure determination. In the related field of optical imaging, the difference between these two concepts is highlighted by the two distinct approaches to defining resolution: the Abbe 'limit' and the Rayleigh 'resolution criteria' (22). In the Abbe approach, resolution corresponds to the maximal spatial frequency of the object being imaged (*e.g.*, a biomolecule) that is successfully transferred into the image. This approach forms the basis of the most common formulation of resolution in cryoEM, where the reported resolution of a cryoEM map corresponds to the spatial frequency at which the Fourier Shell Correlation (FSC) calculated between two independent reconstructions from separate halves of a single dataset (23) crosses a statistically defined threshold (24). For clarity, we refer to this formulation of resolution as the FSC resolution. In contrast, in the Rayleigh approach, the resolution of an image is defined as the extent to which two closely spaced objects may be differentiated from each other in the image (22). In any image, objects appear distorted due to effects that are captured by the point spread function (PSF) (25), and those distortions can cause objects to appear significantly broadened and even to overlap. A Rayleigh-like resolution criterion defines a cutoff distance between two objects below which their apparent overlap in an image is considered so significant that they cannot be distinguished from one another, and above which they are considered to be resolved.

The Abbe definition of resolution, which deals with length scales, and the Rayleigh definition, which focuses on specific objects, are clearly related (22). Intuitively, for an imaging system that captures higher spatial frequencies, the PSF is likely to be narrower,

and thus objects in the image should be more easily differentiated. Yet, these two approaches are fundamentally different in the questions they seek to answer. In the Abbe definition, the resolution of an image is formulated in terms of a conditional limit—all imaged features with spatial frequencies below the limit are 'resolved' and those above are not—yet no consideration is given to the specific identities of the features themselves. In the Rayleigh definition, the objects themselves are the focus of a binary question of being resolved, yet the definition is limited to only that specific feature; it does not give information on other features at different length scales. Both of these approaches are valid as definitions for resolution, but the utility of each one depends on the specific question being asked.

In the context of these two definitions of resolution, the general debate over the meaning of 'atomic resolution' in structural biology may be easily rationalized. The question of atomic resolution is most appropriately asked in the context of the Rayleigh definition, since it is a binary question of whether specific structural features (*i.e.*, atoms) may confidently be differentiated in the experimental data. In many sub-fields of structural biology, however, the standard definition of resolution used by researchers follows the Abbe definition (*e.g.*, gold-standard FSC for cryoEM (23), or nominal resolution (d_{min}) in X-ray crystallography (26)). Since biomolecules consist of a hierarchy of localized structural features that span different length scales (27), the Abbe resolution has proved particularly useful in this context as it specifies a global length-scale limit to which those biomolecular structural features can be resolved. Nonetheless, this mismatch between a question about specific features (*i.e.*, atoms) that is best asked in the Rayleigh framework, and an insightful but tangential answer provided by the Abbe framework is why, in our opinion, the definition of atomic resolution for cryoEM has been so contentious. In an ideal scenario, the definition of resolution would provide insight into both the resolvability of atoms and also the other higher-order structural features present in a biomolecule to incorporate the benefits of both frameworks. To the best of our knowledge, no such definition has been formulated.

In this work, we have developed a new definition of atomic resolution for imaged biomolecules that leverages the advantages of both the Abbe and Rayleigh approaches.

This definition requires the construction of several length scale-dependent models of the imaged biomolecule following the hierarchical description of biomolecular structure (*i.e.*, atomic, primary, secondary, tertiary, quaternary structure (27)) (Fig. 1). Individually, each model is used to describe the resolvable structural features that are embedded within the experimental data at that length scale by utilizing a separate Rayleigh resolution criterion for each hierarchical level. A probabilistic inference framework is then used to relate how well these models match the latent structural information present in the experimental data (28). For a dataset that contains spatial information on a particular structural feature, a 'higher-level' model with less spatial information will not account for the observed complexity of the data, while a 'lower-level' model with more spatial information will contain more complexity than the data contains. The optimal model is the one whose complexity best matches the spatial information present in the data. Crucially, since the ability to resolve biomolecular structural features on one length scale (*e.g.*, the atomic level) necessitates the ability to resolve biomolecular structural features on all larger length scales (*e.g.*, the primary- and secondary-structure levels), selecting the most optimal model from this hierarchy enforces an Abbe-like conditional dependence on length scale. Thus, our framework combines the benefits of both the Abbe and Rayleigh approaches into a scale-dependent, hierarchical probability measure of resolution that quantifies biomolecular structural features across multiple length scales. While the focus of this work is primarily an application of this framework to cryoEM maps, it is worth noting that our definition of a probability measure for resolution operates in a relatively technique-agnostic manner.

Using this scale-dependent hierarchical resolution framework, we have defined atomic resolution for SPA cryoEM experiments as the condition when atomic-level features provide a better description of the latent structural information embedded in a cryoEM map than do higher level features (Fig. 1). This approach enabled us to define a probability measure for atomic resolution, P, and we have implemented a machine learning-based approach to calculate P for local regions of a cryoEM map, which we call hierarchical atomic resolution perception (HARP). In addition to locally quantifying

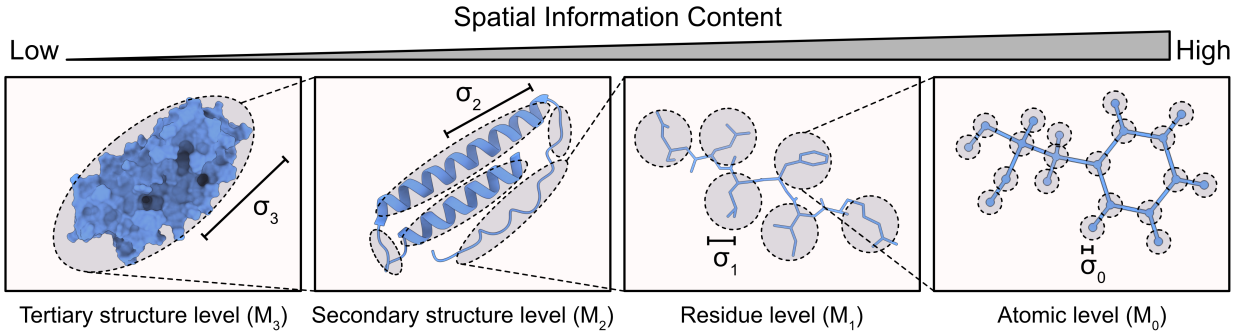

Figure 1. A hierarchy of structural models that define cryoEM maps at different levels of biomolecular structure over a range of length scales. *Schematic showing the structure of a biomolecule (a monomer of apoferritin, PDB ID: 7A4M) with four levels of biomolecular structural features spanning a range of length scales. These levels may be described by corresponding models M_is with characteristic image-profile widths σ_is, where lower values of i denotes models with increasing spatial information on the structure of the biomolecule.*

whether atomic resolution has been achieved, HARP determines whether the structural model associated with a cryoEM map is justified given the evidence in the experimental data. Therefore, HARP also acts as an approach to map-to-model validation. With those capabilities in mind, we used HARP to quantify the distribution of local atomic resolution for each of the cryoEM-derived structures and maps deposited in the protein databank (PDB) (29) and electron microscopy databank (EMDB) (30). Using a statistical model to interpret these distributions, we extended the results of HARP to perform a meta-analysis of the state of cryoEM structural biology (31), and explored several of the intrinsic scientific and social factors that affect biomolecular structure quality. These insights from an extensive evaluation of the entire field not only provide a practical, applied validation of HARP, but also highlight the power and flexibility of the concept of scale-dependent, hierarchical resolution.

Results and Discussion

The profile of isolated atoms in cryoEM maps

To quantitatively determine how well a cryoEM map captures biomolecular structural information, we first sought to understand how individual atoms are represented in a cryoEM map. When using a microscope, the PSF (or its Fourier transform, the contrast transfer function (CTF)) determines how an imaged point-like object is captured in the

resulting two-dimensional micrograph (25). Given the resolving capabilities of a transmission electron microscope, however, atoms do not appear as point-like objects in two-dimensional cryoEM micrographs. Instead, they have a broad profile that results from the image formation process of an electron beam interacting with those atoms in the vitreous ice-embedded biomolecule (32–34). Using this knowledge, in this section we elucidate how atoms appear in the three-dimensional cryoEM maps that are reconstructed from such two-dimensional micrographs in SPA cryoEM.

Under the projection assumption, the weak phase object approximation, and the isolated atom superposition approximation (32–34), atoms in an 'ideal' (*i.e.*, perfectly detected and CTF-corrected) electron micrograph would appear as two-dimensional Gaussians that represent the square of the electrostatic interaction potential between the atoms and the incident electron wave used for imaging (see *Supporting Information*). In the absence of any experimental sources of heterogeneity, a SPA using many such ideal micrographs and the projection-slice theorem (35–38) would yield a three-dimensional cryoEM map in which the 'profile' of an atom is represented by an isotropic three-dimensional Gaussians with a width of $\sigma_{theory} \approx 0.056\,\text{Å}$—a number that is largely independent of the specific element being imaged (at least for neutral atoms of the elements that are typically found in biomolecular structures, *i.e.*, H, C, N, O, P, S) (Fig. S1). This theoretical profile width is consistent with the image formation process of a transmission electron microscope where the incident planar wave of electrons is primarily scattered by the electron-shielded electrostatic field of the atomic nuclei (2); however, from a practical point-of-view, this width is surprisingly small and it is immediately obvious that this is not achieved in real experimental situations.

Indeed, the profiles of atoms in an experimentally derived cryoEM map appear to be broadened further than in the ideal, theoretical case. One major cause is that a cryoEM map contains information from thousands of individual, vitrified molecules. While these molecules are assumed to be perfectly static in the vitrified ice, in reality they undergo conformational dynamics both before and after vitrification (39–41). The reconstructed cryoEM map is therefore broadened by ensemble averaging of the conformational heterogeneity between different molecules in the SPA reconstruction, and temporal

averaging of the molecular motions of the individual molecules during imaging. Regardless of the source, the effect of this averaging on the cryoEM map may be described, at first approximation, by a broadening of the profile of each atom by a distortion factor (DF). As a result, atoms in the reconstructed cryoEM map can be represented by a broader isotropic Gaussian with a width, σ_0, given by $\sigma_0^2 = \sigma_{theory}^2 + \sigma_{DF}^2$, where σ_{DF} represents the amount of DF-induced broadening. While the term DF is similar to the atomic displacement parameter or B-factor (42), it is meant here to encompass a broader range of effects, such as errors in the SPA reconstruction due to misestimation of particle poses in micrographs. The smallest possible value of σ_{DF} is achieved when the broadening of atomic profiles in the cryoEM map occurs only due to thermal motions of the atoms in the vitrified ice (*i.e.*, with no errors in reconstruction or conformational heterogeneity). To approximate this situation, we calculated the DF for the best-case scenario that arises for a well-behaved, plunge-frozen crystal of metmyoglobin at 80 K with an overall B-factor of 5 Å^2 (43). In this case, we found that $\sigma_0 \approx 0.24\,\text{Å}$ (Fig. S1). Thus, the expected profile width of an atom in a cryoEM map under the most ideal experimental conditions for a biomolecule is almost an order of magnitude larger than the ideal profile width (*i.e.,* $0.056\,\text{Å}$), and is primarily determined by the distortions in the experimental sample (as described by σ_{DF}) instead of the distortions due to imaging (as described by σ_{theory}).

Atomic profile widths in experimental cryoEM maps

Due to the other sources of DF-based broadening that occur in real cryoEM experiments, the value of σ_0 should be even larger than our best-case scenario estimate. To understand exactly how broad the profiles of atoms in an experimental SPA-derived cryoEM map are, we calculated the optimal global value of σ_0 for several cryoEM maps. Using atomic coordinates from the corresponding structural models for these cryoEM maps, we constructed models of the cryoEM maps (M_0) by centering individual, isotropic three-dimensional Gaussian densities with the same characteristic global value of σ_0 at

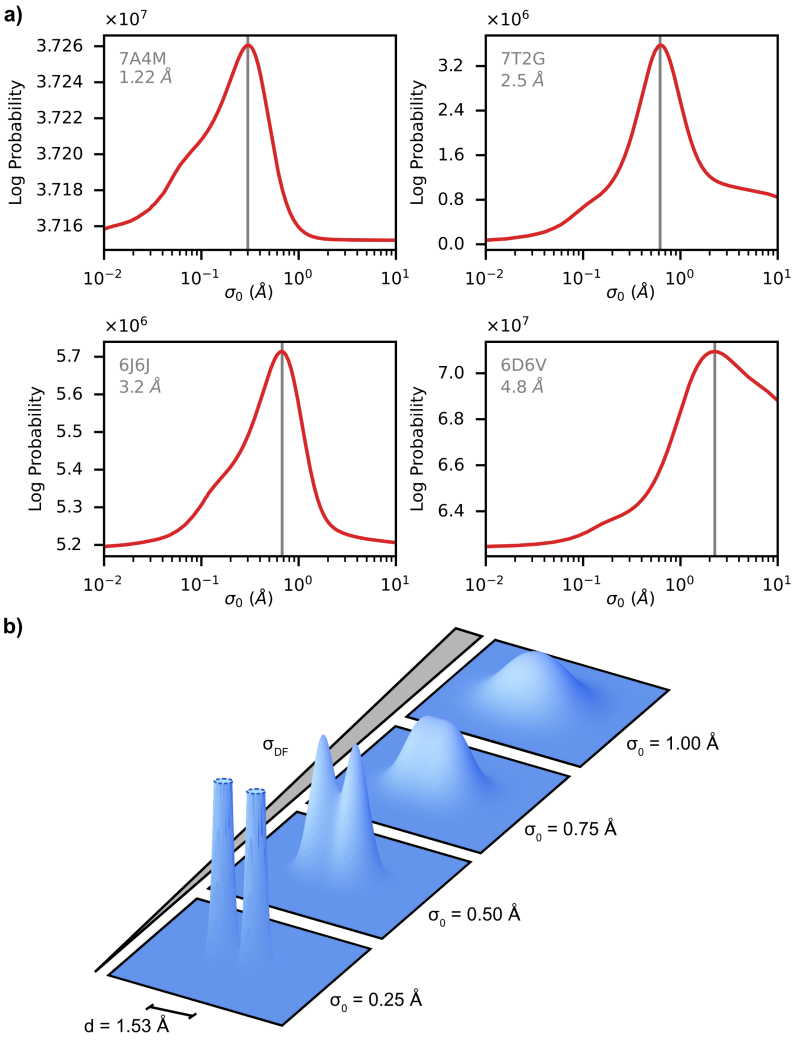

Figure 2. Isotropic image-profile widths for atoms in cryoEM maps. a) *Probability distributions for the global, Gaussian image-profile widths (σ_0) for atoms in four cryoEM maps at different FSC resolutions (the PDB ID for the structural models are shown for each). The peak for these distributions is shown with the grey vertical line and corresponds to the most probable values of σ_0 for each map.* ***b)*** *Illustration showing the densities for two two-dimensional Gaussians separated by a certain distance (in this case, the length of a C-C single bond) with different σ_0s. As σ_{DF} for these Gaussians is increased, the two Gaussians slowly merge untill they cannot be distinguished, thereby signifying a loss of resolution by the Rayleigh definition.*

each atomic position. Subsequently, we calculated the probability, $P(Y \mid M_0, \sigma_0)$, that the M_0 model with a global atomic profile width σ_0 explains the latent structural information in the experimental cryoEM map, Y (see *Supporting Information*). These probabilities are calculated in a way that is independent of the scale, background, or noise by using a previously developed mathematical framework for analyzing the latent structural information in high dimensional datasets (28), thereby removing the dependence on

parameters that vary greatly between different cryoEM maps due to the SPA reconstruction process (see *Supporting Information* for further details). Using this approach, we investigated four previously published cryoEM maps that span a range of reported FSC resolutions from 1.22 Å to 4.8 Å, and calculated $P(Y \mid M_0, \sigma_0)$ for a range of different σ_0 values to identify the most probable σ_0 for each cryoEM map (Fig. 2a) (44–47). In these four cases, the most probable σ_0 trends with the FSC resolution of the cryoEM map. Notably, for the best (*i.e.*, lowest value) FSC resolution cryoEM map (apoferritin at a reported FSC resolution of 1.22 Å, PDB ID: 7A4M (44)), we found that $\sigma_0 = 0.31$ Å, only ~30% higher than our estimate for the lowest experimentally achievable σ_0 (*i.e.*, 0.24 Å). This shows that by using well-behaved samples along with state-of-the-art equipment and reconstruction algorithms, it is possible to achieve ideal imaging conditions in a cryoEM SPA experiment.

While this performance is expected for the best FSC resolution cryoEM maps, it was surprising to us that the cryoEM maps with a higher value of FSC resoution (*e.g.*, with a reported FSC resolution of 4.8 Å) also had well-defined peaks in the plots of $P(Y \mid M_0, \sigma_0)$ *vs.* σ_0 (Fig. 2a). This meant that the atoms in these cryoEM maps can still be defined as relatively narrow Gaussians under these conditions. By obtaining a definite value of σ_0, we had effectively determined the PSF for these cryoEM maps, and thus seemed to have 'super-resolved' the atomic coordinates within them. Therefore, this calculation appears to have obtained atomic-level information from cryoEM maps with FSC resolutions so poor that they conventionally are not believed to contain such information. The seeming contradiction is resolved by considering the hypothetical images of two atoms separated by a fixed distance. As σ_{DF} increases, the profiles of the two atoms begin to overlap to the point where the atoms are no longer distinguishable in the image, which causes a loss of resolution by the Rayleigh definition (Fig. 2b). Even in this situation, however, the profile widths (σ_0) can still be accurately estimated if the atomic coordinates are already known. When estimating the global σ_0 values for the different cryoEM maps in Fig. 2a, the calculation of $P(Y \mid M_0, \sigma_0)$ assumes the atomic coordinates deposited into the PDB are 'true', and thus we were able to estimate σ_0 conditioned upon knowledge of the 'true' atomic coordinates. Yet the deposited atomic

coordinates were inferred from cryoEM maps, which may or may not have contained sufficient atomic-level latent information to enable accurate inference of those atomic coordinates. By including presumed prior knowledge about atomic coordinates in the calculation of $P(Y \mid M_0, \sigma_0)$, we could always find an optimal global σ_0 value and thus access atomic-level information, regardless of whether such information was present in the cryoEM map in the first place.

While it is clear that using the M_0 model to obtain the optimal global σ_0 value does allow us to extract some atomic-level information from a cryoEM map (Fig. 2a), that information is only reliable when it comes from cryoEM maps where the atomic coordinates can be specified with a high degree of certainty. Understanding the certainty with which atomic coordinates can be specified from the evidence present in a cryoEM map requires a comparison with a quantitative description of what a cryoEM map looks like when atomic-level information missing (see below and Fig. 1). Fortunately, biomolecular structural features exist in a hierarchy across several different length scales (27) and atomic coordinates are not needed to describe, *e.g.*, the location of a residue or a secondary-structure element.

A hierarchy of models to describe cryoEM maps at different length scales

In order to quantify how cryoEM maps appear both with and without atomic-level structural information, we defined a hierarchy of models to represent the spatial information present in a cryoEM map at different 'levels' of biomolecular structure (*i.e.*, different length scales) (Fig. 1). These hierarchical models are referred to as M_n, where M_0 is the model for atomic-level information, M_1 is the model for residue-level, and higher M_n's are for information at larger length scales, such as the secondary- and tertiary-structure levels. While 'intermediate' levels may be created through various coarse-graining schemes (48), we have not explored these in this work.

Similar to M_0, we defined a residue-level model M_1 for the case where the individual atoms comprising the biomolecule imaged in a cryoEM map are not resolved, and the cryoEM map is best explained with spatial information encoded at the residue level (Fig. 1). To account for the absence of atomic location information in M_1, we

represent each residue by collapsing the atoms of that residue into the residue center-of-mass and placing an isotropic, three-dimensional Gaussian of width σ_1 at that center. Similar to our treatment of σ_0 for M_0, we sought to estimate the value of σ_1 that would be experimentally observed under ideal imaging conditions. In the previous section, we found that a cryoEM map of apoferritin with a very low value of FSC resolution (PDB ID: 7A4M (44)) nearly achieved our theoretical estimate for the minimum experimental value σ_0, and thus represented some of the most ideal cryoEM imaging conditions to date. We therefore used this cryoEM map and the corresponding M_1 structural model to estimate a lower-bound on σ_1. By modeling all the residues in this structure with a global value of σ_1, we were able to calculate the probability, $P(Y \mid M_1, \sigma_1)$ that the M_1 model with a global residue profile width σ_1 explains the latent structural information in the cryoEM map, Y (see *Supporting Information*). By varying the value of σ_1, we found that the most probable value for this cryoEM map was $\sigma_1 = 0.75$ Å (Fig. S2). We expect this value of σ_1 is close to the lower bound for all experimentally observed σ_1s, which will otherwise be larger than 0.75 Å due to distortions from the imaging and reconstruction process (*e.g.*, molecular heterogeneity).

While this work primarily focuses on a comparison between M_0 and M_1 (see below), we note that the higher levels of the hierarchy of models can be defined using similar approaches. For example, M_2 represents the situation where secondary structure elements (*e.g.*, α-helices in proteins, or B-form double helices in nucleic acids) can be identified in a cryoEM map, but the individual residues within a secondary structure element cannot be differentiated (Fig. 1). CryoEM maps that are best explained by M_2 thus encode spatial information at this secondary-structure level. However, secondary structural elements exhibit significant anisotropy, and thus M_2 requires a significantly more complicated distribution than an isotropic, three-dimensional Gaussian distribution for the image profile of these secondary structure elements. Similarly, M_3 represents the situation where an entire biomolecule or the individual subunits of a multimeric complex can be identified in a cryoEM map, but the individual secondary-structure elements cannot be differentiated (Fig. 1). Higher order M_n, or models containing fractional definitions of structural elements, may be defined for more complex biomolecular

structures.

When properly specified and arranged, the hierarchical nature of these models means that if the experimental data is resolved at one level of the hierarchy, then it is also resolved at all higher levels. In this sense, for 'high-resolution' techniques like cryoEM that routinely report on atomic positions, exact definitions of the highest-level models are not always necessary because they will follow from the lower-level models (*e.g.*, residue-level) that are effectively always resolved in practice. While rigorous anisotropic definitions of those higher-level models are beyond the scope of this work, they would be very useful for 'low-resolution' structural techniques, such as small-angle X-ray scattering (49) or video-rate atomic force microscopy (50).

Defining the resolution limits of the hierarchy of models

Having defined a hierarchy of models to describe cryoEM maps at different length scales, we used this idea to develop a new definition of resolution and quantify the amount of biomolecular structural information contained within a cryoEM map. Specifically, we sought to calculate whether a cryoEM map had achieved atomic resolution by determining whether M_0 explains the latent structural information within a cryoEM map better than all of the other models in the hierarchy. However, a fair comparison between the levels in the hierarchy of models requires that we assess the ability of each level to describe the cryoEM map only across the length-scale regimes that each can be considered to be resolved. As seen in the case of increasing σ_0 for M_0 (see above) (Fig. 2B), defining these length-scale regimes is equivalent to creating an independent Rayleigh-like resolution criterion for each of the M_n levels of the hierarchy of models. The result is a set of resolution criteria that describe when a structural feature on a particular level can be considered 'resolved.' Because these Rayleigh-like resolution criteria define the length scales over which their respective M_n are applicable, they thus enforce the structural hierarchy of the models.

For levels of the hierarchy of models that have isotropic structural features, these resolution criteria are effectively defined by a maximum image profile width (*i.e.*, σ_n^{max} for the model M_n). Each σ_n^{max} ensures that the spatial length scale of the M_n level does not

expand past the point where its structural features transform into the features of a higher-level model (Figs. 1 and 2). For example, σ_0^{max} represents the cutoff below which M_0 contains atomic-level spatial information; and above which M_0 has grown to the point where it no longer contains the atomic-level spatial information it is meant to describe, but instead contains only residue-level and higher spatial information. Thus, by enforcing the σ_n^{max} cutoffs, this hierarchy of models with Rayleigh-like resolution criteria creates a natural separation between the different length scales of biomolecular structure. Importantly, in this framework, a cryoEM map is then resolved along a hierarchical set of the M_n (given by $H = \{M_n\}$). Determining which level of H best captures the spatial information present in a particular cryoEM map then determines the best resolved structural features, and thus quantifies the resolution of the cryoEM map.

For our hierarchy of models, we propose that the Rayleigh-like resolution criterion for two structural objects described by three-dimensional isotropic Gaussians corresponds to the cutoff point where a shell encompassing 50% of the Gaussian density of one object intersects the center of the density for the other object (Fig. 2b). Thus, two objects separated by a certain distance are resolved when the 50% density shell for one does not go past the center of the density for the other. This definition sets the maximum image-profile width (σ_n^{max}) for which two such objects separated by a given distance may still be considered resolved. According to this definition, the relative value of the three-dimensional density at the midpoint between two objects at this Rayleigh-like resolution criterion cutoff point is 0.744, which is very close to the generalized Rayleigh criterion cutoff value of 0.735 for point objects in two-dimensional optical microscopy (22).

For the isotropic Gaussian-distributed atoms in M_0, we have chosen a reference distance for the 50% overlap to be the length of a carbon-carbon (C-C) single bond (1.53 Å in saturated hydrocarbons) (51), which we expect to be the majority of the covalent bonds in a biomolecular structure. For two atoms separated by this distance, the 50% overlap occurs when $\sigma = 0.999$ Å, leading us to set a cutoff $\sigma_0^{max} = 1.0$ Å for all atomic profiles in M_0. Surprisingly, this value of σ_0^{max} is larger than the best-case scenario value of σ_1 that we determined earlier ($\sigma_1 = 0.75$ Å). This suggests the residue-level model of a biomolecule may serve as a relevant description of biomolecular structure even under

conditions where atoms are resolved, and thus that the detection sensitivity and/or signal-to-noise ratio of the cryoEM map, rather than a frequency cutoff, determines which is a better description under these conditions.

Similarly, we proceeded to use our Rayleigh-like resolution criterion to define a cutoff σ_1^{max} for M_1. Unfortunately, using the same approach we took for M_0 of using the C-C bond as a reference distance is more complicated in this case, as there are several unique circumstances under which two nearby residues can be separated. For instance, the typical distance between nucleic-acid residues stacked in a B-form helix is 3.4 Å, the typical C_α-C_α distance between consecutive amino-acid residues in a protein is 3.8 Å (52), and the typical distances between the two strands in a β-sheet or α-helix are 4.7 Å and 5.4 Å, respectively (27). Fortunately, the fact that all these values are close to each other means any of them would be appropriate for defining σ_1^{max}, and that the optimal choice depends on the context in which we are employing M_1. A key strength of our general hierarchy of models is the flexibility that enables researchers to customize their model definitions and constraints for different experimental situations. To define a σ_1^{max} for our work here, we calculated the distance to the closest residue for each residue in all of the cryoEM-derived biomolecular structures in the PDB with reported FSC resolutions of less than 8 Å (see *Methods and Materials* for further details). On average, this center-of-mass to center-of-mass distance was 4.31 Å (Fig. S3). Taking this as the reference distance for M_1, we see that the 50% overlap condition between two three-dimensional isotropic Gaussians separated by this distance occurs for a Gaussian width of $\sigma = 2.80$ Å. Thus, for M_1, we have used a cutoff $\sigma_1^{max} = 2.80$ Å. Notably, the ease with which we defined the cutoff width and thus the resolution criteria for M_0 and M_1 comes from the fact that both use image models that are three-dimensional isotropic Gaussians. Anisotropic image models, such as those associated with high-order M_ns, will require more complex resolution criteria.

The framework for hierarchical atomic resolution perception (HARP) in cryoEM

Having defined the resolution limits for σ_0 and σ_1 that characterize the Rayleigh-like resolution criteria for M_0 and M_1, we used Bayes' rule (53, 54) to calculate a probability

measure of atomic resolution, $P(M_0 \mid Y,H)$, which is the probability that a particular experimental cryoEM map, Y, provides the most evidence for containing the atomic-level spatial information described by M_0 given our hierarchy of models, H (see *Supporting Information* for details). Notably, while other variations of H can be developed and used, our results here are conditionally dependent upon our specific definition of $H = \{M_0, M_1\}$. Because this approach is based upon a machine learning-based quantification of the latent structural information within the cryoEM map, we call this calculation hierarchical atomic resolution perception (HARP). In a HARP calculation, $P(M_0 \mid Y,H)$ measures the balance of atomic- *versus* residue-level resolution; when $P(M_0 \mid Y,H) \approx 1$, the cryoEM map has the most evidence to support M_0, and when $P(M_0 \mid Y,H) \approx 0$, the cryoEM map has the most evidence to support the other levels of the hierarchy (in this case just M_1). Therefore, $P(M_0 \mid Y,H)$ is a quantitative measure of atomic resolution for a high-resolution cryoEM map.

Atomic resolution is not expected to be constant across different regions of a cryoEM map for a variety of reasons. For instance, since SPA cryoEM maps are generated by reconstructing a three-dimensional map from a large number of individual, two-dimensional micrographs (35–38, 55), poor angular coverage of the molecules imaged in those micrographs limits isotropic resolution in the reconstruction (4). Similarly, the ensemble averaging of the conformational heterogeneity can lead to local regions of the cryoEM map containing different amounts of spatial information (56). With this in mind, we noted that the HARP comparison of just M_0 and M_1 resulted in our particular H being granular down to the residue level, and therefore neighboring residues would contribute independently to the calculation (28). Thus, local variations in atomic resolution could be captured using this approach.

Therefore, we calculated $P \equiv P(M_0 \mid Y,H)$ for the local region around each individual residue imaged within a cryoEM map (see *Supporting Information*). By quantifying whether each particular residue of a biomolecule is imaged at atomic resolution, the resulting set of Ps yields the distribution of spatial information present across the entire cryoEM map. Furthermore, the arithmetic mean of this set, $\bar{P}$, describes the state of atomic resolution for the whole biomolecule, and can be interpreted as the

extent to which any given residue in the cryoEM map of the biomolecule exhibited atomic resolution. For example, a value of $\bar{P} = 0.1$ can be interpreted as 10% of the residues in the cryoEM map encoding spatial information at the atomic level. Therefore, $\bar{P} = 0.1$ corresponds to a 'low resolution' cryoEM map relative to one with a value of $\bar{P} = 0.85$, which can be interpreted as 85% of the residues encoding atomic-level information into a relatively 'high resolution' cryoEM map.

P accurately captures atomic resolution for SPA cryoEM

The ability to quantify the level of atomic resolution achieved by each residue in a biomolecular structure provides the opportunity to analyze the factors that influence the distribution of P within each structure. Thus, instead of using HARP to just calculate an average $\bar{P}$ for each structure, we quantified the distributions of P across sets of structures using statistical modeling (Fig. S4). In this statistical model, each residue in a structure can either succeed at achieving atomic resolution (*i.e.*, M_0 is the better description of the residue in the cryoEM map) or fail (*i.e.*, M_1 is the better description). The entire HARP calculation for each individual structure can then be considered as a collection of Bernoulli trials, one for each residue, where the probability of successfully achieving atomic resolution was unknown and could subsequently be inferred (53, 57) (see *Supporting Information*). We mathematically described the inferred probability of success for these Bernoulli trials with a beta distribution, where the parameter α denotes the amount of success (*i.e.,* the effectiveness of M_0), and β denotes the amount of failure (*i.e.,* the effectiveness of M_1) (53, 57). Because the values of α and β for each structure were unknown, we quantified the distributions of atomic resolution within a group of structures (*e.g.*, all available structures with reported FSC resolution between 3.0 Å to 3.1 Å) by inferring the distributions of α and β within that group (see *Supporting Information*). As discussed below, quantifying these distributions within judiciously chosen groups of structures enabled this statistical modeling to provide insight into the underlying factors that govern atomic resolution for cryoEM-derived structures.

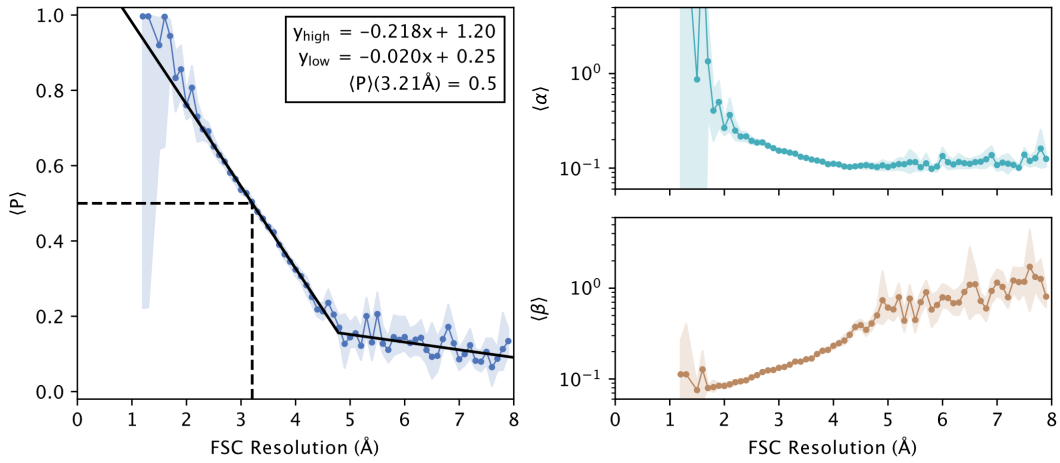

Figure 3. Dependence of $\langle P \rangle$ on the reported FSC resolution. *(left) The value of $\langle P \rangle$ for subsets of cryoEM maps at a reported FSC resolution (blue) along with a predictive linear model (black) which shows the phases of low atomic resolution (FSC resolution > 5.0 Å) and intermediate atomic resolution (FSC resolution < 5.0 Å). $\langle P \rangle$ saturates below ~1.5 Å, where full atomic resolution is achieved. Linear equations for the high- and low-FSC resolution phases are shown in the inset along with the FSC resolution corresponding to a $\langle P \rangle$ of 0.5. (right) Inferred values of $\langle \alpha \rangle$ (top; teal) and $\langle \beta \rangle$ (bottom; beige). The shaded regions in all plots represent the 95% highest posterior density interval (HPDI) for these distributions.*

Using HARP and this statistical model, we analyzed the distributions of atomic resolution for all pairs of cryoEM structures and cryoEM maps deposited in the PDB (29) and EMDB (30) that had a reported FSC resolution less than 8 Å (a total of 12,470 structures; see *Methods and Materials*). The first step of this statistical modelling process was to run HARP on each of these structures. These HARP calculations took roughly one minute *per* structure and about half a day in total to complete all when run in parallel on a 20-core desktop computer. We subsequently used these calculations to investigate the relationship between atomic resolution as calculated by HARP and the reported FSC resolution of the cryoEM maps. We analyzed sets comprising all of the cryoEM maps with a particular FSC resolution, and estimated the average value of P, $\langle P \rangle$, for these sets of cryoEM maps as a function of FSC resolution (Fig. 3; left). We also calculated the corresponding distributions of α and β for these sets as a function of FSC resolution (Fig. 3; right). Roughly, the average value of α, $\langle \alpha \rangle$, captures how 'atomic' the cryoEM map appears, whereas the average value of β, $\langle \beta \rangle$, captures how 'residue'-like or 'blob'-like the cryoEM map appears. Notably, a cryoEM map being atomic-like or residue-like are not mutually exclusive conditions. As expected, $\langle P \rangle$ increased monotonically with

decreasing FSC resolution values, and this validates the interpretation of P as a measure of the atomic resolution of cryoEM maps. Interestingly, the correlation between $\langle P \rangle$ and FSC resolution showed three distinct phases.

For the first phase, which occurs at FSC resolution values greater than ~5 Å, $\langle P \rangle$ is relatively independent of the reported resolution (Fig. 3). Maps at this FSC resolution and beyond encode little spatial resolution at the atomic level, and therefore better FSC resolution does not necessarily lead to a significant increase in atomic resolution. The non-zero value of $\langle P \rangle \approx 0.1$ in this regime is a consequence of the experimental maps being beyond the point where even the residue-level M_1 is effectively resolved. Thus, our particular choice of H, without M_2 or higher level models, is only exact for relatively high-resolution structures where the FSC resolution value is less than ~5 Å—a range that aligns well with modern cryoEM methodologies (1). Having performed this analysis, however, the equal *a priori* model prior probability distributions used for M_0 and M_1 in the HARP calculation can be updated for future analyses; rather than assuming that atomic resolution is equally as probable as residue-level resolution for this range of FSC resolution, an updated prior probability signifying M_0 is less probable than M_1 at these resolutions would reduce $\langle P \rangle \approx 0$ in this regime. While simply repeating the current round of HARP calculations with such an *ex post facto*-prior reassignment is not statistically sound (53, 57), this result can guide prior probability choices for future calculations.

The behavior of $\langle P \rangle$ in this first phase should be interpreted as neither M_0 nor M_1 providing a good description of the spatial information content in a cryoEM map, rather than as ~10% of the residues being resolved at the atomic level. This is clear from the corresponding behaviors of $\langle \alpha \rangle$ and $\langle \beta \rangle$ in this regime. The value of $\langle \alpha \rangle$ is low and stays relatively constant, showing the lack of atomic level information at these resolutions. Similarly, the value of $\langle \beta \rangle$ stays high, but also relatively constant, showing that the residue-level spatial information also stays constant at FSC resolutions values greater than ~5 Å. Higher-level models, such as M_2 or above, are required to more accurately describe the latent structural information in this phase.

The second phase is an intermediate regime between ~2 Å and ~5 Å that exhibits a sharp linear relationship between $\langle P \rangle$ and FSC resolution (Fig. 3). In this regime, cryoEM

maps with better FSC resolution begin to incorporate increasing amounts of atomic-level spatial information, as evidenced by the corresponding increase in $\langle\alpha\rangle$. This increase in atomic-level information is coupled with a corresponding decrease in non-atomic residue-level information, which is reflected in the decreasing trend in $\langle\beta\rangle$. Together these trends for $\langle\alpha\rangle$ and $\langle\beta\rangle$ combine to yield a sharp linear increase in $\langle P\rangle$ in this regime. In particular, at an FSC resolution of 3.21 Å, $\langle P\rangle = 0.5$; this is the point where half of the residues imaged in an average cryoEM map will exhibit atomic resolution.

Finally, the third phase appears at reported FSC resolution values less than ~2 Å where $\langle P\rangle$ shows a sharper dependence on FSC resolution and eventually saturates at reported resolutions of less than ~1.5 Å (Fig. 3). This saturation signifies that cryoEM maps below this FSC resolution have uniformly achieved atomic resolution to the point where atomic coordinates can be very accurately specified in the molecular models built from these cryoEM maps. Notably, the rapid increase and eventual saturation of $\langle P\rangle$ in this phase are mostly driven by a drastic increase in $\langle\alpha\rangle$ below ~2 Å, which indicates that an abundance of atomic information only begins being embedded into the map below ~2 Å. We note that the large uncertainty in this region is simply because there are very few structures that have achieved this level of FSC resolution. Interestingly, the behavior of $\langle\beta\rangle$ is not perfectly anti-correlated to that of $\langle\alpha\rangle$ in this regime. Unlike the change in the dependence of $\langle\alpha\rangle$ with FSC resolutions below ~2 Å, the dependence of $\langle\beta\rangle$ on FSC resolution stays relatively unchanged until it plateaus at ~1.5 Å where $\langle P\rangle$ saturates. This suggests that the ability to fully capture atomic resolution details in a cryoEM map that occurs when FSC resolution values drops below ~2Å is unrelated to the cryoEM map becoming less 'blob-like', but instead occurs through a distinctly different method of incorporating atomic level spatial information. One possible cause for this effect is that the types of cryoEM maps that have achieved such low FSC resolution values are those of biomolecular samples that exhibit low molecular heterogeneity. Consequently, the latent structural information in such cryoEM maps would be better described by M_0 than those for heterogeneous molecules, because the current formulation of M_0 used here does not explicitly account for biomolecular heterogeneity. Another possible cause is that molecular modeling software becomes significantly more capable in this regime.

Altogether, our results here indicate that P is a rich, informative probability measure that fully captures atomic resolution in cryoEM maps by interrogating the latent structure of the data using a scale-dependent definition of resolution. However, they also highlight that only a small minority of structures achieve this a high level of atomic resolution, raising the question of what are the underlying factors that affect atomic resolution in cryoEM.

A steep rise in P driven by technological and computational advances

In addition to determining how P relates to FSC resolution, we also investigated how it varied with experimental circumstances. By sorting cryoEM-derived structures into sets defined by the year of their deposition in the PDB, we analyzed the distributions of P for different years using the statistical model described in the previous section. This analysis shows that there has been a meteoric rise in $\langle P \rangle$ over the last 15 years (Fig. 4a). Given the rapid technological and computational advances in cryoEM during this time period that gave rise to the 'resolution revolution' (6), this result follows our intuition and further validates our approach to measuring atomic resolution. However, our analysis provides further insight into the factors affecting the steep rise in $\langle P \rangle$. We see that this rise cannot be attributed to an increase in $\langle \alpha \rangle$—which remains relatively unchanged over the years—but instead, is explained by a steady decrease in $\langle \beta \rangle$ (Fig. 4a; right). Given our interpretations of $\langle \alpha \rangle$ and $\langle \beta \rangle$ as the amounts of atom-like and 'blob-like' quality in the cryoEM map, respectively (see above), this result suggests that the increase in atomic resolution is driven by advances in imaging and map reconstruction that have led to sharper cryoEM maps. Consequently, those sharper cryoEM maps decreased the ability of M_1 to explain the spatial information present in a cryoEM map. This is in contrast to any significant advances in molecular modeling or refinement driving the increase in $\langle P \rangle$, which would enable better extraction of the atomic level spatial information and thus yield increases in $\langle \alpha \rangle$—the ability of M_0 to explain the spatial information present in a cryoEM map.

We further investigated these trends in technological advances by analyzing the effects of electron microscopy equipment and cryoEM-related software. However, we

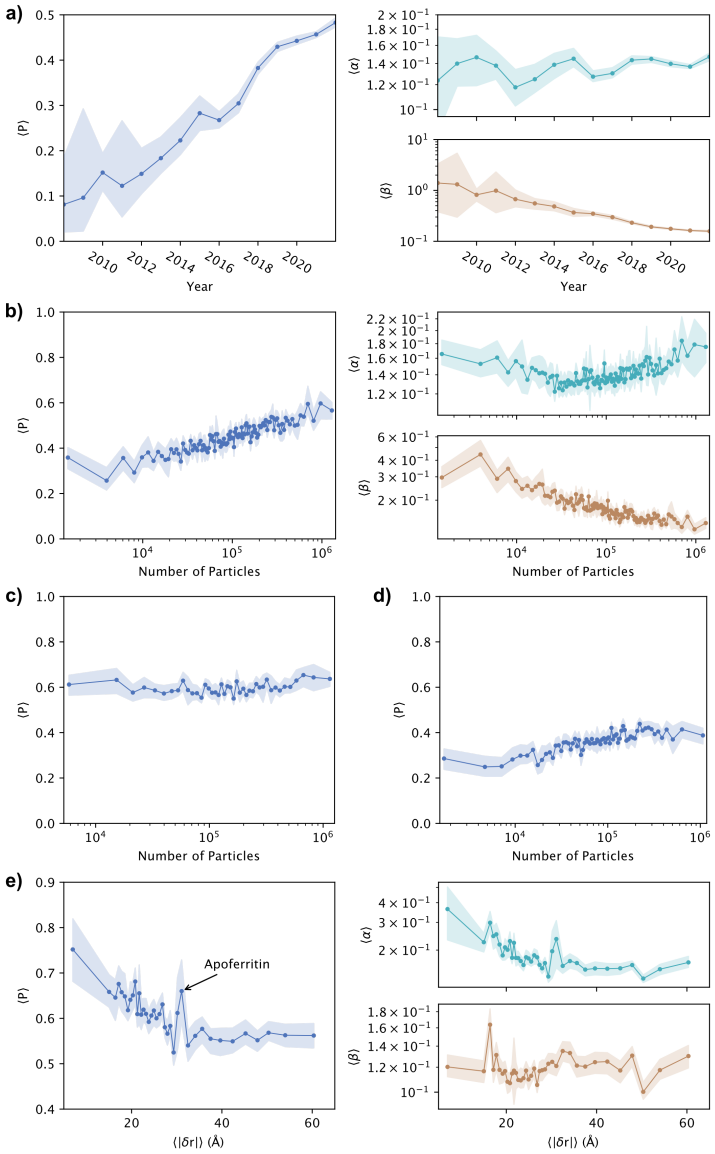

Figure 4. Dependence of ⟨P⟩ on experimental and sample conditions. *Plots of ⟨P⟩ (left; blue) and, wherever applicable, the corresponding ⟨α⟩ (top right; teal) and ⟨β⟩ (bottom right; beige) shown over a range of* ***a)*** *years,* ***b)*** *varying numbers of particles used for the 3D reconstruction of the corresponding map for maps with FSC resolution values less than 8.0 Å,* ***c)*** *varying numbers of particles for reconstruction for maps with FSC resolution values less than 3.2 Å,* ***d)*** *varying numbers of particles for reconstruction for maps with FSC resolution values greater than 3.2 Å, and* ***e)*** *average residue center-of-mass to the molecular center-of-mass distances. The shaded regions in all plots represent the 95% HPDI for these distributions.*

note that our ability to interpret these analyses are limited by confounding variables. For instance, direct electron detectors (58) and Bayesian reconstruction software (59) are relatively recent developments that have both led to significant increases in the quality of cryoEM maps and models. However, their usage is correlated with other factors that are

not uniformly distributed across deposited cryoEM maps and structures (*e.g.*, new technologies are often first tested on biologically well-behaved standard samples; some equipment may be upgraded piecemeal, while others are part of replacements for the entire facility and analysis pipeline; *etc.*). As such, these results should be taken as a descriptive analysis of these technologies as they were historically used, rather than a prescriptive guideline for which technology to adopt to improve atomic resolution in future experiments. To account for the non-uniform widespread uptake of these advanced technologies, we limited our analysis to structures deposited from 2018 and on. Analysis of this subset of structures shows that $\langle P \rangle$ varies significantly with the type of electron detecting camera used (Fig. S5). As expected, structures determined using newer camera models (*e.g.*, Gatan K3, FEI Falcon IV, Direct Electron DE-64) largely achieved greater atomic resolution than those using earlier camera models (*e.g.*, Gatan K2 Base, FEI Falcon II, Direct Electron DE-20). These increases in quality are best explained by a decrease in $\langle \beta \rangle$ (Fig. S5), which alines with our conclusions about the factors affecting $\langle P \rangle$ over the past 15 years.

We performed similar analyses for the software used to generate cryoEM maps. Unfortunately, in addition to the confounding variables mentioned above, these results were futher complicated by the fact that this information is not uniformly well-documented in the metadata of the mmCIF files (60) deposited in the PDB (29). For example, in a randomized subset of these structures that we manually inspected, we found many with incomplete annotations and sometimes misannotations of software used for specific steps. While our analyses of the use of these software as reported in the PDB are thus neither exhaustive nor completely accurate, clear trends can nonetheless be observed. For instance, later versions of a software package used for 3D reconstruction perform better than earlier version (e.g., RELION 4 (61) *vs.* RELION 1 (59)) (Fig. S6). Overall, these results show the clear impact that advances in technology and computation have had on atomic resolution in cryoEM, and how they have led to the 'resolution revolution'.

Extra-scientific factors that correlate with increases in P.

While technological and computational advances have clearly improved the amount of

atomic-level spatial information present in cryoEM maps, we also identified variations in atomic resolution and structural model quality that correlated with extra-scientific considerations. For instance, a relation between sociological factors and P becomes apparent when considering the dependence of $\langle P \rangle$ on the month that each structure was deposited in the PDB. Across multiple years, we found seasonal variations in $\langle P \rangle$ that seemed to cyclically align with academic semesters, with peaks in May, September, and December, and also a significant drop between December and January (Fig. S7). While it Is tempting to ascribe these correlations to teaching burdens, funding cycles, or even weather and humidity, we note that these results are only correlative and not causative.

Similar correlations were also observed for the specific journal that a cryoEM study was published in (Fig. S8). It seems that structures ultimately published in specialist journals relating to structural biology (*e.g., Structure* or *Nature Structure and Molecular Biology*) tend to have higher $\langle P \rangle$ than those in generalist journals (*e.g.*, *Nature, Science*, *Cell*), rather than tracking with impact factor (62). Notably, cryoEM studies appearing in the journals *Structure* and *Proceedings of the National Academy of Science* were characterized by significantly higher $\langle P \rangle$ in comparison to other journals, a difference that is attributed to a relatively high $\langle \alpha \rangle$ for these studies, rather than a change in $\langle \beta \rangle$ which remained around the same value as for other journals. One possible explanation is that researchers publishing in these journals spend more time perfecting their structural models, which leads to an increase in α, before they publish. While these results are clearly not a prescriptive guideline for increasing atomic resolution and cryoEM map quality, they demonstrate how our analysis of cryoEM-associated structural biology here is a description of the historical record. As such, the correlations between P and extra-scientific factors are should be carefully investigated for the study of cryoEM as a practice, and could be the subject of an in-depth sociological investigation that is beyond the scope of this study.

The dependence of P on experimental procedures and sample characteristics

Our statistical modelling further revealed how experimental procedures and sample characteristics also affect the atomic resolution of cryoEM maps. On considering all

cryoEM maps with a reported FSC resolution value of less than 8 Å, an increasing trend was observed between $\langle P \rangle$ and the number of individual micrographs (*i.e.*, particles) used for the three-dimensional reconstruction (Fig. 4b). Interestingly this trend was primarily driven by the 'lower-resolution' cryoEM maps, *i.e.*, ones containing less atomic-level spatial information. We divided the set of cryoEM maps into two subsets with FSC resolution values greater than or less than 3.2 Å (*i.e.*, with $\langle P \rangle$ less than or greater than 0.5, respectively), and repeated this analysis on each set (Fig. 4c-d). We saw that the trend between $\langle P \rangle$ and the number of particles was maintained for the lower-resolution set of maps, whereas, for the higher-resolution set, $\langle P \rangle$ stayed at a nearly constant value (of ~0.6), regardless of the number of particles used in their reconstruction (Fig. 4c). Intuitively, increasing the number of particles is expected to improve the precision of a reconstruction, *e.g.*, by lowering the uncertainty in posterior probability distribution of the cryoEM map coefficients in Fourier space (38), and by increasing the total amount of spatial information incorporated in a specific map (63, 64). In contrast, our results show that this only holds for low-resolution maps (Fig. 4d). In the case of high-resolution cryoEM maps, the amount of spatial information appears to not be limited by particle numbers, but rather by other factors, such as the degree of compositionally and conformationally homogeneity of the biomolecular samples. This limit suggests that newer cryoEM reconstruction and modeling algorithms which can address such molecular heterogeneity (*e.g.*, more powerful classification algorithms) will form an important part of the next cryoEM breakthroughs (65–67).

We subsequently examined several other experimental and sample characteristics, and investigated their respective roles in achieving atomic resolution in cryoEM experiments. We saw that $\langle P \rangle$ had very little dependence on the electron dose (Fig. S9) or the accelerating voltage of the electron microscope (Fig. S10), but that lower humidity in the sample chamber during the plunge-freezing vitrification process produced slightly better results than higher humidity (Fig. S11).

Finally, we elucidated the effects of molecular size, both in terms of reported formula weight and spatial size, on P. Because of constraints due to computational tractability on the sizes of the different sets we could analyze, we only used 'high-

resolution' cryoEM maps (*i.e.*, with FSC resolution values less than 3.2 Å where $\langle P \rangle$ is greater than 0.5) for this statistical modeling. Intuitively, because higher contrast micrographs enable better particle picking and pose/orientation estimation, we expected images of larger molecules to lead to higher-quality three-dimensional reconstructions (63). However, in this high-resolution regime, only a slight decrease in $\langle P \rangle$ with increasing formula weight was observed (Fig. S12). In contrast, we saw a significant decrease in $\langle P \rangle$ with increasing radial length of the biomolecule—measured as the average distance from the center-of-mass of the residues of the biomolecule to the molecular center-of-mass (Fig. 4e). This decrease plateaus after the radial length of the molecule becomes greater than ~30 Å (*n.b.*, the spike at 30 Å is caused by the many high-resolution structures of apoferritin, which is often used to benchmark new technologies (44, 68, 69)). This trend is almost entirely due to decreases in $\langle \alpha \rangle$ with increasing molecular size (Fig. 4e). Notably, it is not due to the cryoEM maps of the smaller biomolecules being sharper and less "blob-like", which would be expected to show corresponding changes in $\langle \beta \rangle$. One possible explanation for this trend is that atomic-level information can be more readily extracted from cryoEM maps for smaller biomolecules, because it is relatively easier to successfully model their atomic structures into the corresponding cryoEM maps compared to modeling large megadalton-sized complexes.

Extent of atomic resolution for individual residues depends on chemical identity

The previous sections show how global factors affect atomic resolution for an entire biomolecule imaged in an SPA cryoEM experiment. We next sought to investigate how local factors affect the atomic resolution of specific residues. From the set of high-resolution maps where the FSC resolution value is less than 3.2 Å (*i.e.*, where $\langle P \rangle$ is greater than 0.5), we separated residues by their chemical identity and then used our statistical model to infer the distributions of the average P for each type of residue, $\langle P \rangle_{res}$. Although each $\langle P \rangle_{res}$ is the average over residues of a certain identity in a biomolecule,

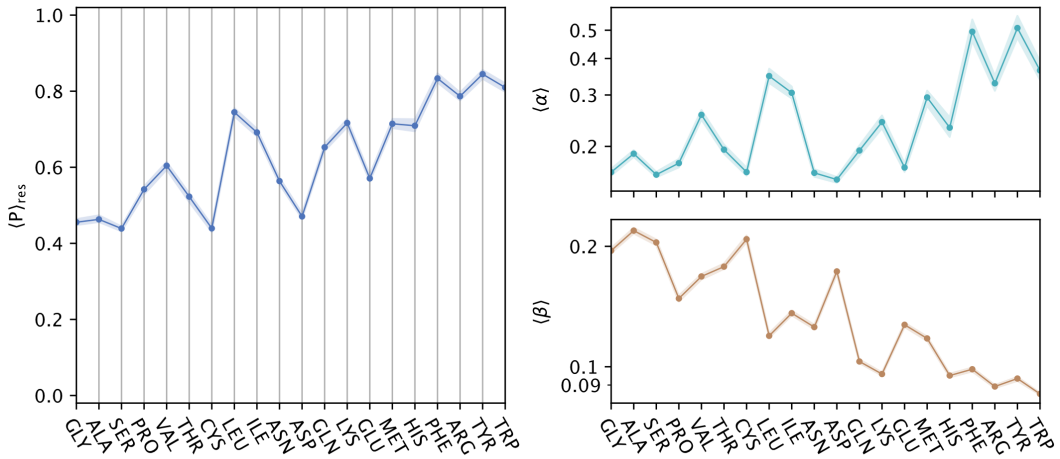

Figure 5. Dependence of $\langle P \rangle_{res}$ on identity of amino-acid residue. *Plot of $\langle P \rangle_{res}$ (left; blue) and the corresponding $\langle \alpha \rangle$ (top right; teal) and $\langle \beta \rangle$ (bottom right; beige) shown for the twenty canonical amino-acid residues, arranged by increasing formula weights of the residues. The shaded regions in all plots represent the 95% HPDI for these distributions.*

the values of $\langle P \rangle_{res}$ can still be interpreted similarly to those of $\langle P \rangle$ in which all residues are simultaneously considered (see above).

For amino acids, the values of $\langle P \rangle_{res}$ across these structures showed a clear positive correlation with both the atomic mass (Fig. 5) and radius of gyration (Fig. S13) of the residue. Interestingly, this trend of increasing $\langle P \rangle_{res}$ correlates both with increasing $\langle \alpha \rangle$ and decreasing $\langle \beta \rangle$ for these amino acids, which suggests that the increased atomic-level information for the larger amino-acid residues is due both to increased ease in modelling the residues in the cryoEM density (making M_0 a better description) and because these larger residues appear less like isotropic blobs (making M_1 a worse description).

Cysteine, aspartic acid, and glutamic acid were notable outliers to this trend. While the $\langle \beta \rangle$ for these residues followed the same general trend as for other residues, it was the corresponding $\langle \alpha \rangle$ that are notably lower than similarly sized residues. Interestingly, these amino acids have been shown to be degraded by damage from electron radiation during the imaging process (3), which could explain why the atomic model for the undamaged residues performed worse for them. However, we also note that outside of a protein, aspartate and glutamate are expected to be negatively charged, and our atomic-level M_0 did not account for electron scattering from charged atoms (70) due to

uncertainties in assigning charge states. Negatively charged atoms could result in different apparent atomic profiles in the cryoEM map (3), which would cause M_0 to underperform for these residues, however, we also note that positively charged residues appear to be relatively unaffected. Regardless, the fact that $\langle P \rangle_{res}$ was perturbed by electron damage and/or charge states highlights the sensitivity of HARP, and demonstrates that it truly captures information at the atomic level for amino acids (Figs. 5, S13), and nucleic acids (Figs. S14-S15).

While beyond the scope of this work, we note that the approach of using HARP for *in silico* comparisons between different atomic models is very promising for assessing subtle atomic changes in high-resolution cryoEM maps, such as those due to post-transcriptional or post-translational modifications (Fig. S16), or radiation damage (Fig. S17). Altogether, the results of our statistical model demonstrate the full power of HARP in not just calculating local atomic resolution and map-to-model quality, but also in elucidating how different factors affect atomic resolution and in detecting subtle changes in molecular details caused by local environmental or experimental conditions.

Conclusion

In recent times, the significant increase in the quality of cryoEM maps has led to an improvement in the ability to extract accurate atomic level information from them by creating structural models. These cryoEM-derived models have consequently become of great interest not just to structural biologists and biophysicists, but to the broader fields of biology and chemistry, where researchers routinely use such models for the design and interpretation of their experiments (19) and, more recently, for the training of sophisticated machine learning algorithms for structure prediction (71, 72). The question of the extent and quality of atomic information that is embedded in these maps is therefore of great importance for the wider scientific community in general. While experts in structural biology might be able to use FSC resolution and other metrics to assess the quality and amount of spatial information present in a particular map, it has not been immediately obvious how non-experts who make use of cryoEM studies can determine the extent of atomic resolution achieved in order to judge the corresponding structural model.

Hierarchical Atomic Resolution Perception (HARP) provides that understanding, and we hope it will fulfil the needs of the field in assessing cryoEM-derived structural models.

The basis of HARP is our formulation of a novel definition of resolution that combines the advantages of the Abbe and Rayleigh approaches by using a set of hierarchical models to explain the cryoEM map at different length scales. This definition explicitly formulates a fundamental relationship between the resolution of a cryoEM map and the ability to extract spatial information from it. Comparisons between the specific models of the hierarchy are used to calculate the resolution of specific structural features in the cryoEM map. In particular, for the case of atomic resolution, we used comparisons between M_0 and M_1 to calculate the probability measure, P, which quantifies the local atomic resolution for individual residues in the cryoEM map. Operating in this probabilistic framework enabled the development of a statistical model to estimate the average P for a group of biomolecules, $\langle P \rangle$, and consequently assess its dependence on a range of experimental and sample conditions to show how cryoEM map quality relates to various practices in the field. Finally, the granularity of this approach allowed us to probe atomic resolution at more than just the biomolecular level by investigating the subtle changes at the chemical level *via* $\langle P \rangle_{res}$.

To enable researchers calculate P and $\bar{P}$ by themselves, we have developed open-source software that implements HARP, and that can be used for any pair of cryoEM map and structural model. HARP will enable researchers to evaluate their structural models of interest, report the average atomic resolution across the structure, and identify regions of the structure that have been modeled with high and low confidence from the cryoEM map. While we have focused in this work on atomic resolution, we note that this approach can be expanded to deal with lower-resolution imaging techniques. Finally, while this study largely deals with the evaluative aspect of our framework, which is of interest to the broader scientific community, the calculations can be easily reversed to be used as a cost function that can be used to refine atomic coordinates into experimental maps or search for atomic-scale features, such as ligands and water molecules—a modality which we hope will prove useful in building more accurate structural models of biomolecules.

Materials and Methods

PDB Analysis. The set of models used for our meta-analysis of the PDB was identified by using the RCSB PDB API (73, 74) to search for depositions that satisfied the parameters specified in Table 1. Briefly, these were chosen to select deposited EM models that were released before January 1, 2023, with reported FSC resolution vaues less than 8 Å. Deposited structures without valid EMDB entries were manually excluded. Altogether, PDBx/mmcif (60) files for 12,470 entries fitting these criteria were downloaded using FTP access from the wwPDB (29). EMDB IDs for the corresponding cryoEM maps were identified from the PDBX/mmcif metadata from the line starting with “emd-”, and the cryoEM maps were downloaded in MRC format (75) from the EMDB (30) using FTP access from the wwPDB (29) for all identified entries but 70JF. MRC files were parsed using the Python library *mrcfile* (76). All HARP code is open-source and was written in Python, using NumPy (77) and Numba (78), or C, and is publicly available through a Git repository at https://github.com/bayes-shape-calc/HARP. Open-source Python code used to perform the analyses of the PDB done in this work is available at https://github.com/bayes-shape-calc/HARP_paper; data were stored in HDF5 format (79), and are available on Zenodo at https://zenodo.org/records/10011336. Plots were generated using Matplotlib (80).

Table 1. RCSB PDB API search parameters.

Attribute	Operator	Value
exptl.method	exact_match	ELECTRON MICROSCOPY
em_3d_reconstruction.resolution	less_or_equal	8.0
rcsb_accession_info.has_released_experimental_data	exact_match	Y
rcsb_accession_info.initial_release_date	greater	1900-01-01
rcsb_accession_info.initial_release_date	less	2023-01-01

HARP Calculations. Calculations of P were performed using the HARP code (see above and *Supporting Information*). Briefly, models were generated by analytically integrating 3D Gaussian densities corresponding to non-interacting atoms across the voxels of the grid specified in the MRC map file associated with a particular structural model. Each atom had a varied width and an element-specific weight (Table S1); the ‘super-atoms’

used in residue-level M_1 models had a varied width and had a weight that was the sum of the weights of the atoms composing that residue. For computational speed, contributions to the total density were only calculated for voxels within 5σ of each atom. The map around a local residue was taken as the ± 8.0 Å cubic grid centered around the Cartesian mean location of all atoms in a residue (*i.e.*, the center-of-mass) (*i.e.,* 1.6 nm sides). Element-specific weights were taken relative to carbon (0.05, 1.0, 1.0, 1.0, 2.0, 2.0 for H, C, N, O, P, S, respectively). Unspecified atoms were given a default weight of 1.0 (*e.g.*, a non-standard atom in a post-translational modification). Only chains corresponding to polymeric entities as listed in the PDBx/mmcif metadata were used (*i.e.*, no ions, water, or small molecules were included). Model evidences were calculated using Eqn. 2.5 from Ref. (28) (*i.e.*, the expression for $m > 0$, $b \in \mathbb{R}$, $\tau > 0$). For M_0, models were calculated at 10 log10-spaced points between $0.25 \text{ Å} \leq \sigma_0 \leq 1.0 \text{ Å}$. For M_1, models were calculated at 20 log10-spaced points between $0.25 \text{ Å} \leq \sigma_1 \leq 2.8 \text{ Å}$. Model priors for these various σ models were log-uniform distributions over the respective range (*i.e.,* the maximum entropy choice for an unknown magnitude), integrated between the midpoints of neighboring σs. Bayes' rule was then used to calculate P as the probability that the M_0 models best explained the observed map.

Statistical Model of Distributions of *P*. Using HARP to process all the PDB entries identified above takes less than 24 hours using a desktop computer with an i9-10900 CPU with 20 threads, and 64 GB DDR4 RAM. Results were compiled and stored on disk in an HDF5 file (79). Statistical modeling of these results was performed using Bayesian Inference (see *Supporting Information*; Fig. S4*)*. Briefly, this was done by identifying sets of structures with a common value or within a range of values for a feature in the PDBx/mmCIF metadata. The set of P for each structure in this set was modeled with a Beta distribution, $Beta(\{P\}_k \mid \alpha_k, \beta_k)$ for the *k*th structure in the set, and the distributions of α_k and β_k were modeled using a log-normal distribution (*e.g.,* $\mathcal{N}(\ln \alpha_k \mid \mu_\alpha, \tau_\alpha^{-1})$ for all α_k, where $\mathcal{N}(x \mid \mu, \sigma^2)$ denotes a Gaussian distribution for x with mean μ and variance σ^2). Plots of $\langle\alpha\rangle$ and $\langle\beta\rangle$ for the entire set are of the form $\langle\alpha\rangle \equiv e^{\mu_\alpha}$ and $\langle\beta\rangle \equiv e^{\mu_\beta}$. The average P within a subset of structures was calculated as $\langle P\rangle \equiv \langle\alpha\rangle/(\langle\alpha\rangle + \langle\beta\rangle)$. Bayesian

inference of $\theta = \left\{\{\alpha\}, \{\beta\}, \mu_\alpha, \mu_\beta, \tau_\alpha, \tau_\beta\right\}$ for a subset of structures was performed using the Laplace approximation (57). The Newton-Raphson method was used to locate the maximum of the posterior and the Hessian was calculated analytically. The prior probability distributions for the μs were uniform and for the τs were log-uniform. The number of structures in any set was kept to less than ~3,000, which is the point where the inference process for the ~6,000 variables associated with such a set became prohibitively slow. The PDBx/mmCIF metadata entries for *_em_3d_reconstruction.resolution*, *pdbx_database_status.recvd_initial_deposition_date* and *_em_3d_reconstruction.num_particles* were used for Fig. 3, Fig. 4a, and Fig. 4b-d respectively. For Fig. 5, only structures with reported FSC resolutions less than 3.2 Å were used.

Author Contributions

C.K. and K.R. performed research, analyzed data, and wrote the paper. C.K. designed research and contributed new reagents/analytic tools.

Author Declaration

The authors declare no competing interest.

Acknowledgments

The authors thank Ruben Gonzalez, John Hunt, Ann McDermott, Eric Greene, Angelo Cacciuto, Riley Gentry, and Erik Hartwick for helpful discussions. This work was supported by a grant to C.K. from the National Science Foundation (NSF) (CHE 2137630), and to Rutgers University-Newark from the NSF (OAC 2117429).

References

1. J. Frank, Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nat. Protoc.* **12**, 209–212 (2017).

2. R. M. Glaeser, How Good Can Single-Particle Cryo-EM Become? What Remains Before It Approaches Its Physical Limits? *Annu Rev Biophys* **48**, 45–61 (2019).

3. M. A. Marques, M. D. Purdy, M. Yeager, CryoEM maps are full of potential. *Curr. Opin. Struct. Biol.* **58**, 214–223 (2019).

4. J. Frank, *Three-dimensional electron microscopy of macromolecular assemblies* (Academic Press, 1996).

5. X. Bai, G. McMullan, S. H. W. Scheres, How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015).

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

7. C. L. Lawson, *et al.*, Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nat. Methods* **18**, 156–164 (2021).

8. G. Cardone, J. B. Heymann, A. C. Steven, One number does not fit all: Mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).

9. A. Kucukelbir, F. J. Sigworth, H. D. Tagare, Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

10. J. L. Vilas, *et al.*, MonoRes: Automatic and Accurate Estimation of Local Resolution for Electron Microscopy Maps. *Structure* **26**, 337-344.e4 (2018).

11. J. B. Heymann, Guidelines for using Bsoft for high resolution reconstruction and validation of biomolecular structures from electron micrographs. *Protein Sci.* **27**, 159–171 (2018).

12. S. Aiyer, C. Zhang, P. R. Baldwin, D. Lyumkis, “Evaluating Local and Directional Resolution of Cryo-EMCryo-electron microscopy (Cryo-EM) Density Maps” in *CryoEM: Methods and Protocols*, Methods in Molecular Biology., T. Gonen, B. L. Nannenga, Eds. (Springer US, 2021), pp. 161–187.

13. M. Dai, Z. Dong, K. Xu, Q. C. Zhang, CryoRes: Local Resolution Estimation of Cryo-EM Density Maps by Deep Learning. *J. Mol. Biol.* **435**, 168059 (2023).

14. B. A. Barad, *et al.*, EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).

15. A. P. Joseph, I. Lagerstedt, A. Patwardhan, M. Topf, M. Winn, Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.* **199**, 12–26 (2017).

16. G. Pintilie, *et al.*, Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat. Methods* **17**, 328–334 (2020).

17. E. Ramírez-Aportela, *et al.*, FSC-Q: a CryoEM map-to-atomic model quality validation based on the local Fourier shell correlation. *Nat. Commun.* **12**, 42 (2021).

18. Z. Wang, A. Patwardhan, G. J. Kleywegt, Validation analysis of EMDB entries. *Acta Crystallogr. Sect. Struct. Biol.* **78**, 542–552 (2022).

19. J. H. Van Drie, L. Tong, Cryo-EM as a powerful tool for drug discovery. *Bioorg. Med. Chem. Lett.* **30**, 127524 (2020).

20. A. Wlodawer, Z. Dauter, `Atomic resolution': a badly abused term in structural biology. *Acta Crystallogr. Sect. Struct. Biol.* **73**, 379–380 (2017).

21. W. Chiu, *et al.*, Responses to `Atomic resolution': a badly abused term in structural biology. *Acta Cryst D* **73**, 381–383 (2017).

22. C. J. R. Sheppard, Resolution and super-resolution. *Microsc. Res. Tech.* **80**, 590–598 (2017).

23. S. H. W. Scheres, S. Chen, Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).

24. P. B. Rosenthal, R. Henderson, Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

25. W. B. Wetherell, "CHAPTER 6 - The Calculation of Image Quality" in *Applied Optics and Optical Engineering*, R. R. Shannon, J. C. Wyant, Eds. (Elsevier, 1980), pp. 171–315.

26. M. S. Weiss, Global indicators of X-ray data quality. *J. Appl. Crystallogr.* **34**, 130–135 (2001).

27. C. R. Cantor, P. R. Schimmel, *Biophysical Chemistry: Part I: The Conformation of Biological Macromolecules*, 1st edition (W. H. Freeman, 1980).

28. K. K. Ray, A. R. Verma, R. L. Gonzalez, C. D. Kinz-Thompson, Inferring the shape of data: a probabilistic framework for analysing experiments in the natural sciences. *Proc. R. Soc. Math. Phys. Eng. Sci.* **478**, 20220177 (2022).

29. wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).

30. C. L. Lawson, *et al.*, EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.* **44**, D396–D403 (2016).

31. P. Neumann, A. Dickmanns, R. Ficner, Validating Resolution Revolution. *Structure* **26**, 785-795.e4 (2018).

32. H. Rullgård, L.-G. Öfverstedt, S. Masich, B. Daneholt, O. Öktem, Simulation of transmission electron microscope images of biological specimens: SIMULATION OF TEM IMAGES OF BIOLOGICAL SPECIMENS. *J. Microsc.* **243**, 234–256 (2011).

33. M. Vulović, *et al.*, Image formation modeling in cryo-electron microscopy. *J. Struct. Biol.* **183**, 19–32 (2013).

34. M. Vulović, L. M. Voortman, L. J. van Vliet, B. Rieger, When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-EM. *Ultramicroscopy* **136**, 61–66 (2014).

35. R. A. Crowther, L. A. Amos, J. T. Finch, D. J. De Rosier, A. Klug, Three Dimensional Reconstructions of Spherical Viruses by Fourier Synthesis from Electron Micrographs. *Nature* **226**, 421–425 (1970).

36. P. Penczek, M. Radermacher, J. Frank, Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* **40**, 33–53 (1992).

37. C. O. S. Sorzano, L. G. de la Fraga, R. Clackdoyle, J. M. Carazo, Normalizing projection images: a study of image normalizing procedures for single particle three-dimensional electron microscopy. *Ultramicroscopy* **101**, 129–138 (2004).

38. S. H. W. Scheres, A Bayesian View on Cryo-EM Structure Determination. *J. Mol. Biol.* **415**, 406–418 (2012).

39. T. Nakane, D. Kimanius, E. Lindahl, S. H. Scheres, Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife* **7**, e36861 (2018).

40. L. V. Bock, H. Grubmüller, Effects of cryo-EM cooling on structural ensembles. *Nat. Commun.* **13**, 1709 (2022).

41. B. Toader, F. J. Sigworth, R. R. Lederman, Methods for Cryo-EM Single Particle Reconstruction of Macromolecules Having Continuous Heterogeneity. *J. Mol. Biol.* **435**, 168020 (2023).

42. O. Carugo, Atomic displacement parameters in structural biology. *Amino Acids* **50**, 775–786 (2018).

43. H. Hartmann, *et al.*, Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. *Proc. Natl. Acad. Sci.* **79**, 4967–4971 (1982).

44. T. Nakane, *et al.*, Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).

45. Q. Qu, *et al.*, Insights into distinct signaling profiles of the μOR activated by diverse agonists. *Nat. Chem. Biol.* **19**, 423–430 (2023).

46. J. Jiang, *et al.*, Structure of Telomerase with Telomeric DNA. *Cell* **173**, 1179-1190.e13 (2018).

47. X. Fan, *et al.*, Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. *Nat. Commun.* **10**, 2386 (2019).

48. J. Jin, A. J. Pak, A. E. P. Durumeric, T. D. Loose, G. A. Voth, Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory Comput.* **18**, 5759–5791 (2022).

49. M. Kornreich, R. Avinery, R. Beck, Modern X-ray scattering studies of complex biological systems. *Curr. Opin. Biotechnol.* **24**, 716–723 (2013).

50. G. R. Heath, S. Scheuring, High-speed AFM height spectroscopy reveals μs-dynamics of unlabeled biomolecules. *Nat. Commun.* **9**, 4983 (2018).

51. D. R. Lide, “STRUCTURE OF FREE MOLECULES IN THE GAS PHASE” in *CRC Handbook of Chemistry and Physics*, 104th Edition, (CRC Press/Taylor & Francis, 2023).

52. S. Chakraborty, R. Venkatramani, B. J. Rao, B. Asgeirsson, A. M. Dandekar, “Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms” (F1000Research, 2013) https:/doi.org/10.12688/f1000research.2-211.v3 (March 28, 2023).

53. E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).

54. C. D. Kinz-Thompson, K. K. Ray, R. L. Gonzalez, Bayesian Inference: The Comprehensive Approach to Analyzing Single-Molecule Experiments. *Annu. Rev. Biophys.* **50**, 191–208 (2021).

55. J. Frank, Averaging of low exposure electron micrographs of non-periodic objects. *Ultramicroscopy* **1**, 159–162 (1975).

56. S. H. W. Scheres, "Processing of Structurally Heterogeneous Cryo-EM Data in RELION" in *Methods in Enzymology*, (Elsevier, 2016), pp. 125–157.

57. C. M. Bishop, *Pattern recognition and machine learning* (Springer, 2006).

58. S. Wu, J.-P. Armache, Y. Cheng, Single-particle cryo-EM data acquisition by using direct electron detection camera. *Microscopy* **65**, 35–41 (2016).

59. S. H. W. Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).

60. J. D. Westbrook, *et al.*, PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology. *J. Mol. Biol.* **434**, 167599 (2022).

61. D. Kimanius, L. Dong, G. Sharov, T. Nakane, S. H. W. Scheres, New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **478**, 4169–4185 (2021).

62. A. Fersht, The most influential journals: Impact Factor and Eigenfactor. *Proc. Natl. Acad. Sci.* **106**, 6883–6884 (2009).

63. R. Henderson, The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).

64. H. Y. Liao, J. Frank, Definition and estimation of resolution in single-particle reconstructions. *Struct. Lond. Engl. 1993* **18**, 768–775 (2010).

65. P. Cossio, G. Hummer, Bayesian analysis of individual electron microscopy images: towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.* **184**, 427–437 (2013).

66. P. Cossio, *et al.*, BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Comput. Phys. Commun.* **210**, 163–171 (2017).

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

68. K. Zhang, G. D. Pintilie, S. Li, M. F. Schmid, W. Chiu, Resolving individual atoms of protein complex by cryo-electron microscopy. *Cell Res.* **30**, 1136–1139 (2020).

69. K. M. Yip, N. Fischer, E. Paknia, A. Chari, H. Stark, Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).

70. L.-M. Peng, Electron Scattering Factors of Ions and their Parameterization. *Acta Crystallogr. A* **54**, 481–485 (1998).

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

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

73. Y. Rose, *et al.*, RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *J. Mol. Biol.* **433**, 166704 (2021).

74. S. Bittrich, *et al.*, RCSB Protein Data Bank: Efficient Searching and Simultaneous Access to One Million Computed Structure Models Alongside the PDB Structures Enabled by Architectural Advances. *J. Mol. Biol.* **435**, 167994 (2023).

75. A. Cheng, *et al.*, MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography. *J. Struct. Biol.* **192**, 146–150 (2015).

76. T. Burnley, C. M. Palmer, M. Winn, Recent developments in the CCP-EM software suite. *Acta Crystallogr. Sect. Struct. Biol.* **73**, 469–477 (2017).

77. C. R. Harris, *et al.*, Array programming with NumPy. *Nature* **585**, 357–362 (2020).

78. S. K. Lam, A. Pitrou, S. Seibert, Numba: a LLVM-based Python JIT compiler in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15., (Association for Computing Machinery, 2015), pp. 1–6.

79. The HDF Group, Hierarchical Data Format, version 5 (1997).

80. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Supplemental Information

A scale-dependent hierarchy of molecular resolution: applications to atomic resolution and cryoEM model validation

Korak Kumar Ray[1], and Colin D. Kinz-Thompson[2,*]

[1] *Single Molecule Imaging Group, MRC-London Institute of Medical Sciences, London W12 0NN, UK.*
[2] *Department of Chemistry, Rutgers University-Newark, Newark, NJ 07102 USA.*
[*] *Correspondence: colin.kinzthompson@rutgers.edu.*

1 Nomenclature and definitions

1.1 Definitions for molecular structures

In the sections below, we use the following terms and definitions to describe molecular structures:

D : A particular database of molecular structures, *e.g.*, D_{PDB} refers to the PDB,
S : A molecular structure in any given D, composed of a collection of residues,
R : A residue in a given S, composed of a collection of atoms,
A : A particular atom in a given R.

Additionally, we have defined operators (denoted with a 'hat') that provide information about the how these items of molecular structure relate to one another. For example, for the i^{th} atom of the j^{th} residue of the k^{th} structure in database D, we can write

$$\begin{aligned} \hat{R}(A_{ijk}) &= R_{jk}, \\ \hat{S}(A_{ijk}) &= S_k, \\ \hat{S}(R_{jk}) &= S_k, \end{aligned}$$

where the indexing is explicitly noted as a subscript for clarity. Similarly, we use operators to obtain other properties of interest, such as

$\hat{T}(A_{ijk}) = t$ where t is the residue type of $\hat{R}(A_{ijk})$ (*e.g.*, Ala, Asn, Asp, ..., A, C, ..., dA, dC, ...),
$\hat{Q}(A_{ijk}) = q$ where q is the atom type of A_{ijk} in $\hat{R}(A_{ijk})$ (*e.g.*, Cα, Cβ, ...),
$\hat{E}(A_{ijk}) = e$ where e is the atomic element of A_{ijk} (*e.g.*, H, N, C, O, ...),
$\hat{r}(A_{ijk}) = \vec{r}$ which are the instantaneous Cartesian coordinates of A_{ijk},
$\hat{r}(R_{jk}) = \vec{r}$ which are the instantaneous Cartesian coordinates of the center-of-mass of R_{jk}.

In the last case, center-of-mass is calculated as the centroid $\hat{r}(R_{jk}) = \overline{\hat{r}(A_{ijk})} \equiv \frac{1}{I_j}\sum_{i=1}^{I_j}\hat{r}(A_{ijk})$, where the 'bar' denotes the average over all I_j atoms in $\hat{R}(A_{ijk})$. Notably, given some probability distribution for a property, we can calculate its expectation value (*i.e.*, $\langle x \rangle \equiv \int dx\, x\, p(x)$). For example,

$$\vec{\mu}_{ijk} \equiv \langle \hat{r}(A_{ijk}) \rangle = \int d\vec{r}_{ijk}\, \vec{r}_{ijk}\, p(\vec{r}_{ijk})$$

and therefore the expected center-of-mass location of a residue is $\langle \hat{r}(R_{jk}) \rangle = \overline{\vec{\mu}_{ijk}}$.

1.2 Definitions for cryoEM

To clarify the discussion of cryogenic electron microscopy (cryoEM) experiments, we have defined

$$\widehat{\mathcal{M}}(S_k) \quad = \quad \mathcal{M}, \quad \text{which yields the experimental technique used to generate } S_k.$$

Using this operator, the subset of D_{PDB} that contains just those structures obtained using cryoEM is

$$D_{cryo} \quad = \quad \{S_k \mid S_k \in D_{PDB},\, \widehat{\mathcal{M}}(S_k) = \text{cryoEM}\}$$

For each $S_k \in D_{cryo}$,

$$\begin{aligned}
\hat{Y}(S_k) &= Y_k(\vec{v}), \quad \text{which is the associated density map over grid-points } \vec{v} \text{ in the associated grid} \\
\hat{G}(S_k) &= G_k, \quad \text{which is the associated grid composed of all the grid-points } \{\vec{v}_l, \ldots\}
\end{aligned}$$

In nearly all cryoEM experiments, G_k is rectilinear and so G_k is completely defined by

$$\begin{aligned}
\hat{o}(G_k) &= (o_x, o_y, o_z), \text{ which are the Cartesian coordinates of the origin of } G_k \\
\hat{n}(G_k) &= (n_x, n_y, n_z), \text{ which are the number of grid-points in each direction} \\
\hat{\Delta}(G_k) &= (\Delta_x, \Delta_y, \Delta_z), \text{ which are the spacing between grid-points in each direction.}
\end{aligned}$$

Based on such a grid G_k, the density map mentioned above is just the set of intensity values given by

$$Y_k(\vec{v}) = \{y(\vec{v}_l), ...\} \text{ for } \vec{v}_l \in G_k$$

where the $y(\vec{v}_l)$ denotes the intensity value of the voxel at $\vec{v}_l$.

In this work, we further consider a subsection of $\hat{Y}(S_k)$ that is a local map centered around a particular residue. This is defined as the collection of voxels within a cube with side length 2ϵ that is centered at the center of mass of a specific residue R_{jk}. Specifically, the local map at R_{jk} is given by

$$\hat{\mathcal{Y}}(R_{jk}) = \mathcal{Y}_{jk}(\vec{v}) = \{y(\vec{v}_l) \mid (|(\vec{v}_l - \langle \hat{r}(R_{jk}) \rangle) \cdot \vec{u}_r| < \epsilon)\ \forall\, \vec{u}_r \in \{\vec{x}, \vec{y}, \vec{z}\}\} \text{ for } \vec{v}_l \in \hat{G}(\hat{S}(R_{jk})),$$

where $\{\vec{x}, \vec{y}, \vec{z}\}$ are the unit vectors of G_k, and ϵ is a cutoff distance.

To create models of experimental density maps, we have used isotropic Gaussian distributions to represent the basic structural elements of both M_0 and M_1 (see below). To parameterize these distributions, we have defined

$$\begin{aligned}
\hat{\sigma}_0^2(A_{ijk}) &= \vec{\sigma}_0^2, \quad \text{which is the 3D Gaussian variance of the image profile of } A_{ijk} \text{ in a density map} \\
\hat{\sigma}_1^2(R_{jk}) &= \vec{\sigma}_1^2, \quad \text{which is the 3D Gaussian variance of the image profile of } R_{jk} \text{ in a density map} \\
\hat{w}(A_{ijk}) &= w, \quad \text{which is the weight for an atom's image in a density map} \\
\hat{\sigma}_{DF}^2(A_{ijk}) &= \vec{\sigma}_{DF}^2, \quad \text{which is the distortion factor for an atom's image profile in a density map,}
\end{aligned}$$

where we use the same width for all atoms within a model and, additionally, we use element-specific weights such that $\hat{w}(A_{ijk}) \equiv w_e$ for all atoms where $\hat{E}(A_{ijk}) = e$.

2 Approach to modeling density maps in a cryoEM experiment

This section follows the approach given by Rullgard *et al.* [1] and Vulovic *et al.* [2]. We point the reader to these papers for more background details.

2.1 The latent structural information present in an ideal 2D cryoEM micrograph of a molecule

Using the projection assumption (PA) and an analogy to an 'optical potential' to describe the interaction of an incident electron wave with a sample in a transmission electron microscope (TEM), the intensity of the electron wave exiting a sample in the z direction is given by

$$\begin{aligned} I(x,y)_{\text{exit}} &= |\psi_{\text{exit}}|^2 * PSF(x,y) \\ &= |\exp\left(i\sigma V_z'\right)\exp\left(-\sigma V_z''\right)|^2 * PSF(x,y), \end{aligned}$$

where $V = V' + iV''$ is the optical potential analog with the imaginary absorptive component V'' arising mostly due to inelastic scatting, PSF is the point-spread function, and $\sigma \equiv \frac{2\pi m_e |e| \lambda}{h^2}$. In the weak phase object approximation (WPOA) (*i.e.*, $i\sigma V_z' \ll 1$), this simplifies to

$$I(x,y)_{\text{exit}} \approx \left(\left(1 + \sigma^2 V_z'^2(x,y)\right)\exp\left(-2\sigma V_z''(x,y)\right)\right) * PSF(x,y).$$

Under ideal imaging conditions, the point-spread function, (*i.e.*, the inverse Fourier transform of the contrast transfer function) can be perfectly corrected for such that, effectively, $PSF(x,y) \approx 1$. In this ideal scenario, the image will be

$$I(x,y) \approx I_0\left(1 + \sigma^2 V_z'^2(x,y)\right),$$

where I_0 is a proportionality constant depending on the specifics of imaging (*e.g.*, exposure time, absorptive losses, *etc.*). In the Born-Oppenheimer approximation, the optical potential analog component for the elastic scattering caused by the interaction of the incident electron wave with the atoms comprising a molecule is equal to the Coulombic interaction potential of the atoms, which is

$$V' \approx V^{int}(r) = \frac{1}{4\pi\epsilon}\left[\int \frac{\rho_e(y)}{|r-y|}dy - \sum_j \frac{eZ_j}{|r-\mathcal{R}_i|}\right],$$

where r is the position of the electron, $\mathcal{R}_i$ is the position of the i^{th} nucleus, ρ_e is the electron density function, Z is the nuclear charge. In the isolated atom superposition approximation (IASA), where we ignore the $\sim 5\%$ contribution to the potential from bonding [2], the Coulombic interaction potential simplifies to

$$V^{int}(r) \approx \sum_i V^{int}_{i,atom}(r).$$

The terms describing the contribution from individual atoms are given by

$$V^{int}_{atom}(r) = \frac{16\pi\hbar^2}{m_e e}\int d^3\xi f_z^{(e)}(\xi)\exp\left(4\pi i\xi r\right),$$

where $f_z^{(e)}(\xi)$ is the electron scattering factor (*n.b.*, the additional factor of two in the exponent arises from the definition of the scattering angle). The $f_z^{(e)}(\xi)$ for different atoms have analytical approximations, many of which are weighted sums of Gaussians (*n.b.*, these are typically constructed by converting X-ray diffraction data into electron scattering using the Mott-Bethe formula; see Ref. [3]). Thus, these approximations are generally of the form

$$f_z^{(e)}(\xi) = \sum_{l=1}^{N} a_l \exp\left(b_l \xi^2\right),$$

where the a_l and b_l are fitted parameters typically specified for each $\hat{E}(A_{ijk})$. In this work, we have used an approximation by Peng with $N = 5$ [3]. Using approximations of this form for the scattering factors, performing the inverse Fourier transform yields

$$V^{int}(\vec{r}) = \frac{2\pi\hbar^2}{m_e e} \sum_{i=1}^{N_{atoms}} \sum_{l=1}^{N} a_{li} \mathcal{N}\left(\vec{r} \mid \vec{r}_i, \frac{b_{li}}{8\pi^2}\right),$$

where $\vec{r}_i$ is the position of the i^{th} atom, and $\mathcal{N}(\vec{r} \mid \vec{\mu}, \sigma^2)$ denotes a 3D, isotropic Gaussian distribution with mean $\vec{\mu}$ and variance σ^2. Because atoms are mobile and/or their positions are uncertain, we approximate this condition by modeling the probability of finding an atom at a particular location using an isotropic uncertainty factor that is distributed according to a Gaussian distribution as

$$p(\hat{r}(A_{ijk})|\theta) = \mathcal{N}(\hat{r}(A_{ijk}) \mid \langle \hat{r}(A_{ijk}) \rangle, \hat{\sigma}^2_{DF}(A_{ijk}))$$
$$p(\vec{r}_i|\vec{\mu}_i, \sigma^2_{DF,i}) = \mathcal{N}(\vec{r}_i \mid \vec{\mu}_i, \sigma^2_{DF,i}),$$

where θ represents the collection of all conditional dependencies, and σ^2_{DF} is a distortion factor that could, *e.g.*, play the role of a Debye-Waller temperature factor. Marginalizing out the $\vec{r}_i$ from $V^{int}(\vec{r})$ by integrating using the uncertainty in each atomic position yields

$$V^{int}(\vec{r}) = \frac{2\pi\hbar^2}{m_e e} \sum_{i=1}^{N_{atoms}} \sum_{l=1}^{N} a_{li}\, \mathcal{N}\left(\vec{r} \mid \vec{\mu}_i, \frac{b_{li}}{8\pi^2} + \sigma^2_{DF,i}\right).$$

Finally, the projection of this Coulombic interaction potential along the z-axis of a TEM is

$$V_z^{int}(x,y) = \int V^{int}(\vec{r})\, dz = \frac{2\pi\hbar^2}{m_e e} \sum_{i=1}^{N_{atoms}} \sum_{l=1}^{N} a_{li}\, \mathcal{N}\left(x \mid \mu_{x,i}, \frac{b_{li}}{8\pi^2} + \sigma^2_{DF,i}\right) \mathcal{N}\left(y \mid \mu_{y,i}, \frac{b_{li}}{8\pi^2} + \sigma^2_{DF,i}\right),$$

where the $\mathcal{N}(x|\mu, \sigma^2)$ are 1D Gaussian distributions with mean μ and variance σ^2. Thus, the real component of the optical potential analog used in image formation is just a weighted sum of Gaussians.

Returning to the TEM image formation process, we note that the TEM micrographs typically used during single-particle analysis (SPA) cryoEM are processed by the analysis software to yield a shifted and scaled image. Thus, the form of an ideal, 2D cryoEM micrograph that has been processed for SPA cryoEM can be obtained by taking the exit intensity image, $I_{exit}(x,y)$, and rearranging to give an image of the form,

$$\frac{I_{exit}(x,y) - I_0}{4\pi^2\lambda^2 I_0} = \left[\sum_{i=i}^{N_{atoms}} \sum_{l=1}^{N} a_{li}\, \mathcal{N}(x \mid \mu_{x,i}, \sigma^2_{0,li}) \mathcal{N}(y \mid \mu_{y,i}, \sigma^2_{0,li})\right]^2,$$

where $\sigma^2_{0,li} \equiv \frac{b_{li}}{8\pi^2} + \sigma^2_{DF,i}$. Because the summation on the right hand side is squared, the cross-terms with the largest a_{li} are the only terms that make significant contributions, and this enables the approximation that

$$\frac{I_{exit}(x,y) - I_0}{4\pi^2\lambda^2 I_0} \approx \sum_{i=1}^{N_{atoms}} \tilde{a}^2_{li} \mathcal{N}(x \mid \mu_{x,i}, 2\tilde{\sigma}^2_{0,li}) \mathcal{N}(y \mid \mu_{y,i}, 2\tilde{\sigma}^2_{0,li}),$$

where the 'tilde' denotes the parameter with the index l that corresponds to the largest value of a_{li} for each i^{th} atom (*i.e.*, $\tilde{x}_l = x_l$ for $l = \text{argmax}_l\, a_{li}$). To assess the effectiveness of this approximation, we fit the full expression using $N = 5$ Gaussian distributions to a single Gaussian and found that the maximum residual is only $\sim 2\%$ (Fig. S1). Thus, a TEM micrograph of a molecule is well approximated as a superposition of a single 2D Gaussian for each atom, weighted by the an element-specific weight (*i.e.*, $\tilde{a}^2_{li} = \tilde{a}^2_{\hat{E}(A_{ijk})}$). Due to differences in absorptive losses based on the particular choice of imaging conditions, we note that the weights of different elements may differ from experiment to experiment. Additionally, since ions have different scattering factors than neutral atoms [4], the ionization state of individual atoms would yield different weights in the image relative to the neutral atom.

2.2 Modeling voxels in a 3D cryoEM density map

During the SPA process of reconstructing a cryoEM density map, the two-dimensional TEM micrographs are typically normalized prior to the reconstruction [5, 6]. Since our approximation of the form of such a micrograph is already normalized (see above), it is clear a perfectly executed three-dimensional real-space or Fourier-space reconstruction using such images [7] will yield a weighted superposition of three-dimensional isotropic Gaussians centered at the mean location of each atom. Therefore, under ideal imaging conditions, we define the expected three-dimensional density map, ρ, of a structure S_k as

$$\mathbb{E}[\rho(\vec{r}|S_k,\theta)] = \sum_{R_{jk}\in S_k}\sum_{A_{ijk}\in R_{jk}} \tilde{a}^2_{ijk}\mathcal{N}(\vec{r}\mid\vec{\mu}_{ijk},\tilde{\sigma}^2_{0,ijk}),$$

where $\tilde{a}^2_{ijk} = \tilde{a}^2_{\hat{E}(A_{ijk})}$, $\vec{\mu}_{ijk} = \langle\hat{r}(A_{ijk})\rangle$, $\tilde{\sigma}^2_{0,ijk} = 2\left(\frac{\tilde{b}_{\hat{E}(A_{ijk})}}{8\pi^2} + \hat{\sigma}^2_{DF}(A_{ijk})\right)$, and θ represents the collection of parameters used to define the image. This expected form of $\mathbb{E}[\rho(\vec{r}|S_k,\theta)]$ is contrasted with the experimentally observed $\hat{Y}(S_k)$, which contains noise and imperfections. From preliminary investigations, it was apparent that different S_k had different optimal values for each $\tilde{a}_e$ (Table S1). For simplicity, in the following work we have used a single set of element-specific values for all S_k such that $\tilde{a}^2_{A_{ijk}} = w_{\hat{E}(A_{ijk})} \equiv w_e$. We also note that our approach requires that the w_e are relative weights to avoid the addition of an extra degree of freedom. Therefore, we chose all weights to be relative to carbon (*i.e.*, $w_C = 1$), and adopted the weights in Table S1.

Table S1: Element-specific weights for the template of a cryoEM map.

Element	Theory (Fig. S1)	7A4M ($\sigma_0^{\text{opt}} = 0.29$ Å)	6Z6U ($\sigma_0^{\text{opt}} = 0.32$ Å)	8B0X ($\sigma_0^{\text{opt}} = 0.48$ Å)	This work
H	0.046	0.0743	0.098	N.A.	0.1
C	1	1	1	1	1.0
N	1.185	1.065	1.151	1.142	1.0
O	1.322	1.001	1.062	0.865, phosphates 0.858, other	1.0
P	3.729	N.A.	N.A.	2.008	2.0
S	4.220	1.718	2.420	1.127	2.0

In experimental situations, however, the reconstructed cryoEM density map of a structure S_k is not continuous, but is instead discretized onto a grid, $\hat{G}(S_k)$. Thus, for any comparisons, the expected density map must also be discretized onto a grid, which yields

$$\mathbb{E}[\rho(\vec{v}|S_k,\theta)] = \iiint_{\vec{v}-.5\vec{\Delta}}^{\vec{v}+.5\vec{\Delta}} \mathbb{E}[\rho(\vec{r}|S_k,\theta)]d\vec{r}$$

$$= \frac{1}{8}\sum_{R_{jk}\in S_k}\sum_{A_{ijk}\in R_{jk}} w_{\hat{E}(A_{ijk})}\prod_{d\in[x,y,z]}\left[\text{erf}\left(\frac{((\vec{v}+.5\vec{\Delta}_k)-\hat{\mu}(A_{ijk}))\cdot\vec{u}_d}{\sqrt{2\tilde{\sigma}^2_{0,ijk}}}\right) - \text{erf}\left(\frac{((\vec{v}-.5\vec{\Delta}_k)-\hat{\mu}(A_{ijk}))\cdot\vec{u}_d}{\sqrt{2\tilde{\sigma}^2_{0,ijk}}}\right)\right]$$

where $\hat{\Delta}(G_k) = \vec{\Delta}_k$ is the spacing of the grid. It is worth noting that $\vec{v}$ can be equivalently expressed in Cartesian coordinates (as above) or using indices (such as l), which are interconverted, for example in the x-dimension, using the equations,

$$l = (x - o_x)/\Delta_x, \text{ and}$$
$$x = l \times \Delta_x + o_x,$$

where o_x is the x-coordinate for $\hat{o}(G_k)$, the origin of G_k.

This discretized, expected density is the basis of our 'template' (see Ref. [8]). Given the experimental density map around a particular residue $\hat{\mathcal{Y}}(R_{jk}) = \mathcal{Y}_{jk}$ where the distance cutoff $\epsilon = 8$ Å gives a cubic sub-grid and sub-density larger than the size of the R_{jk}, we construct a template, X_{jk}, for this region of a density map as

$$\hat{X}(R_{jk}, \theta) \equiv X_{jk} = \mathbb{E}[\rho(\vec{v}|R_{jk}, \theta)] \text{ for } \vec{v} \in \hat{G}(\hat{\mathcal{Y}}(R_{jk})),$$

where the operator $\hat{X}$ provides the output template X_{jk}, and θ again serves as a reminder that the template depends upon particular choices of parameters such as those that control the Gaussian profile widths, $\tilde{\sigma}_{0,ijk}$, which modulate the spatial information present in X_{jk}. In this work, we use a single value of the profile width (*i.e.*, $\tilde{\sigma}_{0,ijk} \equiv \sigma_0$) for all A_{ijk} that contribute to X_{jk}.

In the Bayesian inference-based shape-analysis framework [8], such templates are used to calculate the marginal likelihood probability of observing $\mathcal{Y}_{jk}$, regardless of any distortions to scale, offset, or noise, as

$$P(\mathcal{Y}_{jk} \mid X_{jk}, M_0, \sigma_0) = \iiint dm_{jk}\, db_{jk}\, d\tau_{jk} \left[\prod_{\vec{v} \in G_k} \mathcal{N}(\mathcal{Y}_{jk}(\vec{v}) \mid m_{jk} \cdot X_{jk}(\vec{v}) + b_{jk}, \tau_{jk}^{-1}) \right] \cdot p(m_{jk}, b_{jk}, \tau_{jk}),$$

where M_0 denotes that an atomic-level model from the hierarchy of structural representation is used here, and m, b, and τ are scale, offset, and noise parameters [8]. In particular, since density is positive, we have analytically performed this integration for the case where $m > 0$, $b \in \mathbb{R}$, and $\tau > 0$. Thus,

$$P(\mathcal{Y}_{jk} \mid X_{jk}, M_0, \sigma_0) = \frac{\Gamma(\frac{N-2}{2}) N^{-\frac{N}{2}} V_X^{-\frac{1}{2}}}{2\Delta m_{jk} \Delta b_{jk} \Delta \ln \tau_{jk}} \left[\pi V_{\mathcal{Y}} (1 - r^2)\right]^{-\frac{N-2}{2}} \left[1 + \frac{r}{|r|} I_{r^2}\left(\frac{1}{2}, \frac{N-2}{2}\right)\right],$$

where $\Gamma(a)$ is the gamma function, N is the number of voxels ($\vec{v}$) in X_{jk} and $\mathcal{Y}_{jk}$, $V_X = \langle X_{jk}^2 \rangle - \langle X_{jk} \rangle^2$ and $V_{\mathcal{Y}} = \langle \mathcal{Y}_{jk}^2 \rangle - \langle \mathcal{Y}_{jk} \rangle^2$ are the variances of X_{jk} and $\mathcal{Y}_{jk}$ respectively, $r = \frac{\langle X_{jk} \mathcal{Y}_{jk} \rangle - \langle X_{jk} \rangle \langle \mathcal{Y}_{jk} \rangle}{V_X V_{\mathcal{Y}}}$ is the correlation coefficient between X_{jk} and $\mathcal{Y}_{jk}$, and $I_\nu(\alpha, \beta)$ is the regularized incomplete beta function. Additionally, Δm_{jk}, Δb_{jk}, and $\Delta \ln \tau_{jk}$ are defined by $\Delta f(x) = f(x_{\text{max}}) - f(x_{\text{min}})$, and arise from the uniform, uniform, and log-uniform distributions used as prior probability distributions for m_{jk}, b_{jk}, and τ_{jk}, respectively. In this work, we have used the following values for these terms

$$\Delta m_{jk} = \Delta b_{jk} = 2 \times 10^5,$$
$$\Delta \ln \tau_{jk} = 2 \times \ln(10^3),$$

however, we also note that these values do not actually affect anything because these terms cancel in the model selection calculation performed that is HARP (see below).

3 Hierarchical atomic resolution perception (HARP) calculations

The basis of HARP is a Bayesian model selection calculation amongst the hierarchy of models that we use to describe biomolecular structure at different length scales. In the particular calculation performed here, we have used a reduced hierarchy of $H = \{M_0, M_1\}$. The model for M_0 is described above. The model M_1 is conceptually similar manner to M_0. However, because residues are the smallest structural element on this level, M_1 uses a coarse-grained 'super-atom' for each residue R_{jk} in S_k. These super-atoms are located at the expected center-of-mass of each R_{jk} (*i.e.*, $\langle \hat{r}(R_{jk}) \rangle = \vec{\mu}_{ijk}$), and each super-atom has a weight of $w_{jk} = \sum_i w_{ijk}$, which is the sum of the weights of all the A_{ijk} in R_{jk}.

The Bayesian model selection calculation is then for the probability P that M_0 is the better description amongst the hierarchy, and is specifically calculated as

$$P \equiv P(M_0|\mathcal{Y}_{jk}, X_{jk}, H) = \frac{P(\mathcal{Y}_{jk}|X_{jk}, M_0)P(M_0)}{P(\mathcal{Y}_{jk}|X_{jk}, M_0)P(M_0) + P(\mathcal{Y}_{jk}|X_{jk}, M_1)P(M_1)}$$

where $P(M_n)$ are the model prior probabilities, which are taken as

$$P(M_0) = P(M_1) = \frac{1}{2},$$

and $P(\mathcal{Y}_{jk}|X_{jk}, M_n)$ are the marginal likelihoods obtained by marginalizing σ_n from $P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_n)$ using

$$P(\mathcal{Y}_{jk}|X_{jk}, M_n) = \int d\sigma_n \, P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_n) \, P(\sigma_n|M_n).$$

The prior probability distributions used for $P(\sigma_n|M_n)$ in this marginalization calculation were taken to be a log-uniform distribution (*i.e.*, the maximum entropy distribution for an unknown magnitude), which is

$$P(\sigma_n|M_n) = \frac{\sigma_n^{-1}}{\ln(\sigma_{n,\text{max}}) - \ln(\sigma_{n,\text{min}})}$$

where the maximum and minimum values of σ_n are defined by the Rayleigh-like resolution criterion for each level. In our implementation of HARP, these upper and lower-bounds are

$$\begin{aligned} M_n &: \sigma_{n,\text{min}} \leq \sigma_n < \sigma_{n,\text{max}} \\ M_0 &: 0.25 \text{ Å} \leq \sigma_0 < 1.0 \text{ Å} \\ M_1 &: 0.25 \text{ Å} \leq \sigma_1 < 2.8 \text{ Å}. \end{aligned}$$

As described in the main text, the lower-bounds were determined from the Debye-Waller factor of a cryo-cooled metmyoglobin crystal structure [9] (Fig. S1), and the upper-bounds correspond to the resolution criterion cutoff that were derived by considering the length of a C-C single bond for M_0 and the empirical closest residue distribution for M_1 (Fig. S3).

The marginalization integral was calculated in several windows, which were specific regions spanning $\sigma_{n,\text{min}}$ and $\sigma_{n,\text{max}}$ where the i$^{\text{th}}$ region is in the range $[\sigma_{n,i}, \sigma_{n,i+1})$, such that

$$P(\mathcal{Y}_{jk}|X_{jk}, M_n) = \sum_{i=0}^{N-1} \int_{\sigma_{n,i}}^{\sigma_{n,i+1}} d\sigma_n \, P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_n) \, P(\sigma_n|M_n),$$

and where $\sigma_{n,0} = \sigma_{n,\text{min}}$ and $\sigma_{n,N} = \sigma_{n,\text{max}}$. For computational tractability, the contribution of each region to $P(\mathcal{Y}_{jk}|X_{jk}, M_n)$ was approximated by expanding the likelihood function $P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_n)$ in each region as a Taylor series at the midpoint $\sigma_{n,m_i} \equiv (\sigma_{n,i} + \sigma_{n,i+1})/2$ and truncating it after the initial term to yield

$$\begin{aligned} P(\mathcal{Y}_{jk}|X_{jk}, M_n) &= \sum_{i=0}^{N-1} \left(P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_{n,m_i}) \frac{\ln \sigma_{n,i+1} - \ln \sigma_{n,i}}{\ln(\sigma_{n,\text{max}}) - \ln(\sigma_{n,\text{min}})} + \mathcal{O}(\sigma_{n,i+1} - \sigma_{n,i}) \right) \\ &\approx \sum_{i=0}^{N-1} P(\mathcal{Y}_{jk}|X_{jk}, M_n, \sigma_{n,m_i}) \frac{\ln \sigma_{n,i+1} - \ln \sigma_{n,i}}{\ln(\sigma_{n,\text{max}}) - \ln(\sigma_{n,\text{min}})}. \end{aligned}$$

This approach is conceptually the same as a middle Riemann sum, except that it includes the full integrated contribution of the prior probability distribution. Effectively this approach assumes that the value of the likelihood within each small region is constant and approximately the value at the midpoint. The error in the approximation within each region scales with the size of the region. Practically, this approach means that the integration calculation can be carried out by evaluating the likelihood at regularly spaced points of σ_n that serve as the midpoints σ_{n,m_i}, and then determining the corresponding region boundaries $\sigma_{n,i}$ for the integration. The midpoints were chosen (10 for M_0, and 20 for M_1) along a $\log_{10}$-spaced range bounded by $\sigma_{n,\text{min}}$ and $\sigma_{n,\text{max}}$.

Finally, because the value of P obtained in a HARP calculation can be computed for any residue R_{jk} in a particular S_k, we index it as P_{jk}. When P_{jk} is calculated for all the residues in a particular S_k, this process yields the set of probabilities $\{P\}_k$. The sample average of this set, $\bar{P}_k$, provides a useful statistic to quantify the atomic resolution of a given S_k.

4 Statistical model of biomolecular structure quality

Sequentially running HARP on each S_k in a group of structures (*e.g.*, the d^{th} subset of structures within D_{cryo}, D_d) yields a set $\{P\}_k$ for each of the S_k. It can be useful to understand how these $\{P\}_k$ are distributed across such a D_d. For instance, it might be useful to compare how atomic resolution is differently distributed between two different subsets that are distinguished by their Fourier shell correlation (FSC) resolution values. However, large differences in both the molecular and experimental details between the S_k in a D_ds complicate a direct comparison of the $\{P\}_k$. On the other hand, use of a statistic such as the sample average $\bar{P}_k$ for comparison of the S_k in D_d will yield an incomplete, limited picture of the variation within a D_d. For a complete comparison, we have developed a statistical model to quantify the dispersion within the sets of $\{P\}_k$ for the different S_k within a D_d (Fig. S4).

In this statistical model, the P_{jk} in $\{P\}_k$ for a specific structure S_k are assumed to be independent and identically distributed (i.i.d) according to the same beta distribution. Thus,

$$P_{jk} \sim \text{Beta}(P_{jk}|\alpha_k, \beta_k) \qquad \forall\, P_{jk} \in \{P\}_k,$$

or, in other words, the probability density function (PDF) for each of the P_{jk} in $\{P\}_k$ for a specific S_k is

$$p(P_{jk}|\alpha_k, \beta_k) = \frac{P_{jk}^{\alpha_k - 1}(1 - P_{jk})^{\beta_k - 1}}{B(\alpha_k, \beta_k)},$$

where α_k and β_k are the hyperparameters of the beta distribution for a specific S_k and $B(\alpha, \beta)$ is the beta function. For the S_k within some D_d, we assume that all of the α_k and β_k are related to each other, and specifically that they are distributed according to a common, overarching distribution. Since, the α_k and β_k are necessarily positive as they are the parameters of a beta distribution, and the magnitudes of these α_ks or β_ks are not known *a priori*, we use a log-normal distribution to describe the distributions of the α_ks and β_ks, and thus the dispersion within D_d. This choice is based on the principle of parsimony, because a log-normal distribution utilizes only two parameters and its support renders it appropriate to represent an unknown magnitudes. Therefore, we assume that each of the α_ks in a particular D_d are i.i.d. according to the PDF

$$p(\alpha_k|\mu_{\alpha,d}, \tau_{\alpha,d}) = \frac{1}{\alpha_k}\sqrt{\frac{\tau_{\alpha,d}}{2\pi}}\exp\left[-\frac{\tau_{\alpha,d}}{2}(\ln\alpha_k - \mu_{\alpha,d})^2\right],$$

where $\mu_{\alpha,d}$ and $\tau_{\alpha,d}$ are the location and scale hyperparameters, respectively, of the log-normal distribution. The β_ks are similarly i.i.d. according to a log-normal distribution, but with hyperparameters $\mu_{\beta,d}$ and $\tau_{\beta,d}$. The final conditional dependencies between the parameters of this statistical model are shown in the directed acyclic graph (DAG) shown in Fig. S4.

The ultimate aim of our statistical modelling is to infer the hyperparameters of the log-normal PDFs as a way to quantitatively describe the dispersion within the sets $\{\alpha_k\}$ and $\{\beta_k\}$ for a particular D_d. To do this, we determine the most probable $\theta = \{\{\alpha_k\}, \{\beta_k\}, \mu_{\alpha,d}, \mu_{\beta,d}, \tau_{\alpha,d}, \tau_{\beta,d}\}$ that describe the HARP results for all of the residues within all of the S_k in a D_d. This is achieved within the Bayesian framework of probability by using Bayes' rule to calculate the posterior probability distribution $P(\theta|D_d)$ as

$$P(\theta|D_d) = \frac{P(D_d|\theta)\cdot P(\theta)}{P(D_d)}.$$

For this statistical model (Fig. S4), the likelihood $P(D_d|\theta)$ is

$$P(D_d|\theta) \equiv \mathcal{L} = \prod_{k=1}^{K_d}\left[\prod_{j=1}^{J_k} p(P_{jk}|\alpha_k, \beta_k)\right] p(\alpha_k|\mu_{\alpha,d}, \tau_{\alpha,d})\, p(\beta_k|\mu_{\beta,d}, \tau_{\beta,d}),$$

where J_k is the number of residues in S_k, K_d is the number of structures in D_d, and $p(P_{jk}|\alpha_k, \beta_k)$, $p(\alpha_k|\mu_{\alpha,d}, \tau_{\alpha,d})$, and $p(\beta_k|\mu_{\beta,d}, \tau_{\beta,d})$ are defined above. For the prior probability distribution, $P(\theta)$, we have assumed no prior knowledge about $\mu_{\alpha,d}$ or $\mu_{\beta,d}$, and only assume for $\tau_{\alpha,d}$ or $\tau_{\beta,d}$ that they of an unknown magnitudes and must be positive. Therefore, we use the corresponding maximum entropy principle-derived prior probability distributions that encodes such information [8]. Assuming that these parameters are independent of each other, the total prior probability distribution is based upon uniform distributions for $\mu_{\alpha,d}$ and $\mu_{\beta,d}$, and log-uniform distributions for $\tau_{\alpha,d}$ and $\tau_{\beta,d}$, which altogether is

$$P(\theta) = \frac{\tau_{\alpha,d}^{-1}\tau_{\beta,d}^{-1}}{\Delta\mu_{\alpha,d}\Delta\mu_{\beta,d}\Delta\ln\tau_{\alpha,d}\Delta\ln\tau_{\beta,d}},$$

where $\Delta f(x)$ is defined above. While we must set the minimum and maximum values for each of these parameters in order to completely encode our prior knowledge of the problem, practically, because we search for the *maximum a posteriori* (MAP) point to solve the inference problem (see below), the particulars of this choice do not change the inference procedure. Finally, we do not know an analytical expression for the evidence, $P(D_d)$, so we cannot determine the complete analytical expression for the posterior, $P(\theta|D_d)$, so instead we approximate the posterior using the Laplace approximation.

4.1 The Laplace approximation of the statistical model

The Laplace approximation allows us to estimate an unknown posterior distribution around the MAP point. Following the description given by Bishop [10], we can expand the logarithm of any function, $\ln f(z)$, at a local maximum, $z = z_0$, to the second order, yielding

$$\ln f(z) = \ln f(z_0) - \frac{1}{2}(z - z_0)^T A(z - z_0) + \cdots,$$

where $A = -\nabla\nabla \ln f(z)|_{z=z_0}$ is a matrix called the Hessian. Note that the first order term has disappeared, because this expression is an expansion at the local maximum, z_0. Dropping terms higher than second order and exponentiating gives

$$f(z) \approx f(z_0)\ \exp\left[-\frac{1}{2}(z - z_0)^T A(z - z_0)\right].$$

When $f(z)$ is a PDF, this approximating function can be normalized to yield an approximate PDF by recognizing that it matches the functional form of a Gaussian distribution, giving

$$f(z) \approx \mathcal{N}(z|z_0, A^{-1}).$$

Thus, the Laplace approximation can be used to approximate a posterior PDF as a Gaussian centered at the MAP point of the posterior (z_0) with a variance equal to the inverse of the Hessian of $-\ln f(z)$ at that point.

In order to use the Laplace approximation for our model shown in Fig. S4, we first locate the MAP point of the posterior by finding the θ that gives the zero of the derivative (F) of the log-joint probability, $\ln \mathcal{J}(D_d, \theta) \equiv \ln \mathcal{L}(D_d|\theta) + \ln P(\theta)$, because $\mathcal{J}(D_d, \theta)$ is proportional to the posterior. Specifically, we used the Newton-Raphson method to iteratively locate the MAP point where $F(\theta) = 0$, by starting with an initial guess $\theta_{t=0}$, and updating each iteration according to

$$\theta_{t+1} = \theta_t - J^{-1}(\theta_t)F(\theta_t)$$

where J is the Jacobian of the function F. The following sections describe how F, J, and J^{-1} are calculated from $\ln \mathcal{J}(D_d, \theta)$.

4.2 Calculating F using the first derivatives of $\ln \mathcal{J}(D_d, \theta)$

As we describe above, $\ln \mathcal{J}$ is a function of $\theta = \{\{\alpha_k\}, \{\beta_k\}, \mu_{\alpha d}, \mu_{\beta d}, \tau_{\alpha d}, \tau_{\beta d}\}$. The derivative F is therefore a vector of size $2K_d + 4$, where K_d is the number of S_k in D_d (*n.b.*, because each S_k has a its own α_k and β_k). To simplify notation in this and the following subsections, we drop the index d such that $\mu_{\alpha,d} \equiv \mu_\alpha$, *etc.*, and just use $\mathcal{J}$ to represent $\mathcal{J}(D_d, \theta)$.

The partial derivatives of $\ln \mathcal{J}$ with respect to (w.r.t) each individual α_k and β_k is given by

$$\frac{\partial \ln \mathcal{J}}{\partial \alpha_k} = \left(\sum_{j=1}^{J_k} \ln P_{jk} \right) - J_k \psi(\alpha_k) + J_k \psi(\alpha_k + \beta_k) - \alpha_k^{-1}(1 - \mu_\alpha \tau_\alpha + \tau_\alpha \ln \alpha_k),$$

and

$$\frac{\partial \ln \mathcal{J}}{\partial \beta_k} = \left(\sum_{j=1}^{J_k} \ln(1 - P_{jk}) \right) - J_k \psi(\beta_k) + J_k \psi(\alpha_k + \beta_k) - \beta_k^{-1}(1 - \mu_\beta \tau_\beta + \tau_\beta \ln \beta_k),$$

where J_k is the number of residues R_{jk}s in the corresponding structure S_k, and $\psi(a)$ is the digamma function. The partial derivatives w.r.t. μ_α and τ_α are given by

$$\frac{\partial \ln \mathcal{J}}{\partial \mu_\alpha} = \tau_\alpha \left(\sum_{k=1}^{K_d} \ln \alpha_k \right) - K_d \tau_\alpha \mu_\alpha,$$

and

$$\frac{\partial \ln \mathcal{J}}{\partial \tau_\alpha} = \frac{K_d - 2}{2} \tau_\alpha^{-1} - \frac{1}{2} \left(\sum_{k=1}^{K_d} (\ln \alpha_k)^2 \right) + \mu_\alpha \left(\sum_{k=1}^{K_d} \ln \alpha_k \right) - \frac{K_d}{2} \mu_\alpha^2.$$

Finally, the expressions for the partial derivatives w.r.t. μ_β and τ_β are identical to the ones above w.r.t. μ_α and τ_α respectively.

4.3 Calculating J using the second derivatives of $\ln \mathcal{J}(D_d, \theta)$

The Jacobian, $\mathcal{J}(D_d, \theta)$, of F is a square matrix of size $\mathcal{K} \times \mathcal{K}$, where $\mathcal{K} = 2K_d + 4$, which is composed of second derivatives of $\ln \mathcal{J}$. In this section, we provide the expressions for these second derivatives. The partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \alpha_k}$ are given by

$$\frac{\partial^2 \ln \mathcal{J}}{\partial \alpha_{k'} \partial \alpha_k} = \begin{cases} -J_k \psi'(\alpha_k) + J_k \psi'(\alpha_k + \beta_k) + \alpha_k^{-2}(1 + \mu_\alpha \tau_\alpha) - \tau_\alpha \alpha_k^{-2} + \tau_\alpha \alpha_k^{-2} \ln \alpha_k, & \text{if } k = k', \\ 0, & \text{if } k \neq k', \end{cases}$$

and

$$\frac{\partial^2 \ln \mathcal{J}}{\partial \beta_{k'} \partial \alpha_k} = \begin{cases} J_k \psi'(\alpha_k + \beta_k), & \text{if } k = k', \\ 0, & \text{if } k \neq k', \end{cases}$$

where $\psi'(a)$ is the trigamma function. Furthermore,

$$\frac{\partial^2 \ln \mathcal{J}}{\partial \mu_\alpha \partial \alpha_k} = \tau_\alpha \alpha_k^{-1},$$

and

$$\frac{\partial^2 \ln \mathcal{J}}{\partial \tau_\alpha \partial \alpha_k} = \alpha_k^{-1}(\mu_\alpha - \ln \alpha_k).$$

Finally,

$$\frac{\partial^2 \ln \mathcal{J}}{\partial \mu_\beta \partial \alpha_k} = \frac{\partial^2 \ln \mathcal{J}}{\partial \tau_\beta \partial \alpha_k} = 0.$$

The expressions for the partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \beta_k}$ are equivalent to the corresponding partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \alpha_k}$ described above.

The partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \mu_\alpha}$ are

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{J}}{\partial \mu_\alpha^2} &= -K_d \tau_\alpha, \\
\frac{\partial^2 \ln \mathcal{J}}{\partial \alpha_k \partial \mu_\alpha} &= \tau_\alpha \alpha_k^{-1}, \\
\frac{\partial^2 \ln \mathcal{J}}{\partial \tau_\alpha \partial \mu_\alpha} &= \left(\sum_{k=1}^{K_d} \ln \alpha_k \right) - K_d \mu_\alpha.
\end{aligned}$$

All other partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \mu_\alpha}$ are zero, and are therefore omitted here. The partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \mu_\beta}$ have equivalent expressions to the ones above.

Similarly, for $\frac{\partial \ln \mathcal{J}}{\partial \tau_\alpha}$, we have

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{J}}{\partial \tau_\alpha^2} &= -\frac{K_d - 2}{2} \tau_\alpha^{-2} \\
\frac{\partial^2 \ln \mathcal{J}}{\partial \alpha_k \partial \tau_\alpha} &= \alpha_k^{-1}(\mu_\alpha - \ln \alpha_k) \\
\frac{d^2 \ln \mathcal{J}}{d \mu_\alpha d \tau_\alpha} &= \left(\sum_{k=1}^{K_d} \ln \alpha_k \right) - K_d \mu_\alpha
\end{aligned}$$

All other partial derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \tau_\alpha}$ are zero, and are therefore omitted here. The derivatives of $\frac{\partial \ln \mathcal{J}}{\partial \tau_\beta}$ have equivalent expressions to the ones above.

4.4 Inverting the Jacobian to calculate J^{-1}

As mentioned above, J is a $\mathcal{K} \times \mathcal{K}$ matrix, so inverting can be non-trivial when $\mathcal{K}$ is very large as it is for most D_d. However, when we order $\theta = \{\alpha_1, \ldots, \alpha_{K_d}, \beta_1, \ldots, \beta_{K_d}, \mu_{\alpha d}, \mu_{\beta d}, \tau_{\alpha d}, \tau_{\beta d}\}$, we see that J has a block form corresponding to

$$J = \begin{bmatrix} a & b & e \\ c & d & f \\ g & h & i \end{bmatrix},$$

where the blocks a, b, c, and d are only composed of second derivatives of the form $\frac{\partial^2 \ln \mathcal{J}}{\partial \alpha_{k'} \partial \alpha_k}$, $\frac{\partial^2 \ln \mathcal{J}}{\partial \beta_{k'} \partial \alpha_k}$, $\frac{\partial^2 \ln \mathcal{J}}{\partial \alpha_{k'} \partial \beta_k}$, $\frac{\partial^2 \ln \mathcal{J}}{\partial \beta_{k'} \partial \beta_k}$, and are therefore diagonal because these second derivatives are only non-zero when $k' = k$. Taking

advantage of this fact, we can simplify the calculation of the inverse of the Jacobian, J^{-1}. Re-organizing J into a blocked-block form

$$J = \begin{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} & \begin{bmatrix} e \\ f \end{bmatrix} \\ \begin{bmatrix} g & h \end{bmatrix} & i \end{bmatrix} \equiv \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

we can greatly simplify the calculation J^{-1} using the identity

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + (A^{-1}B)(D - CA^{-1}B)^{-1}(CA^{-1}) & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}(CA^{-1}) & (D - CA^{-1}B)^{-1} \end{bmatrix}.$$

In this situation, because

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and a, b, c, and d are all diagonal and therefore easily invertible, A^{-1} may be easily calculated through a first application of the above identity, which consequently simplifies the second application that comprises the full calculation of J^{-1}. In addition to its use in the update equation for the Newton-Raphson optimization process, J^{-1} is the inverse of the Hessian of $\ln \mathcal{J}$, which, when calculated at the MAP point, is the covariance matrix for the Laplace approximation to the posterior $P(\theta|D_d)$ (see above).

4.5 Newton-Raphson Maximization Protocol

Practically, successful use of the Newton-Raphson method to find the MAP point requires a good initial choice of θ_0. We initialize the α_ks and β_ks by taking a log-uniform random initialization of the precision $s_k = \alpha_k + \beta_k$ between 10^{-2} to 10^2 followed by moment matching $\frac{1}{J_k}\sum_{j=1}^{J_k} \ln P_{jk} = \mathbb{E}\left[\ln P_{jk}\right] = \psi(\alpha_k) - \psi(s_k)$. Solving for α_k gives a good preliminary initialization for a subsequent maximum likelihood estimation of α_k using a few iterations of the Newton-Raphson method to obtain the value used in θ_0 [11]. A similar procedure is performed for β_k using instead a moment-matching procedure for $\mathbb{E}\left[\ln(1 - P_{jk})\right]$. The initializations for μ_α and τ_α are then determined by moment-matching the mean and precision of the $\{\ln \alpha_k\}$, and an equivalent procedure is performed for μ_β and τ_β. From this initialization, successive iterations of the Newton-Raphson method described above are performed until the value of $\ln \mathcal{J}$ converges to a relative change of 10^{-10}, and at least five restarts were performed for each D_d.

References

(1) Rullgård, H.; Öfverstedt, L.-G.; Masich, S.; Daneholt, B.; Öktem, O. *Journal of Microscopy* **2011**, *243*, 234–256.

(2) Vulović, M.; Ravelli, R. B.; van Vliet, L. J.; Koster, A. J.; Lazić, I.; Lücken, U.; Rullgård, H.; Öktem, O.; Rieger, B. *Journal of Structural Biology* **2013**, *183*, 19–32.

(3) Peng, L.-M. *Micron* **1999**, *30*, 625–648.

(4) Peng, L.-M. *Acta Crystallographica Section A Foundations of Crystallography* **1998**, *54*, 481–485.

(5) Sorzano, C.; de la Fraga, L.; Clackdoyle, R.; Carazo, J. *Ultramicroscopy* **2004**, *101*, 129–138.

(6) Scheres, S. H. *Journal of Molecular Biology* **2012**, *415*, 406–418.

(7) Crowther, R. A.; Amos, L. A.; Finch, J. T.; De Rosier, D. J.; Klug, A. *Nature* **1970**, *226*, 421–425.

(8) Ray, K. K.; Verma, A. R.; Gonzalez, R. L.; Kinz-Thompson, C. D. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2022**, *478*, 20220177.

(9) Hartmann, H.; Parak, F.; Steigemann, W.; Petsko, G. A.; Ponzi, D. R.; Frauenfelder, H. *Proceedings of the National Academy of Sciences* **1982**, *79*, 4967–4971.

(10) Bishop, C. M., *Pattern recognition and machine learning*, New York, 2006.

(11) Minka, T. Estimating a Dirichlet Distribution `https://tminka.github.io/papers/dirichlet/` (accessed 10/11/2023).

(12) Hattne, J.; Shi, D.; Glynn, C.; Zee, C.-T.; Gallagher-Jones, M.; Martynowycz, M. W.; Rodriguez, J. A.; Gonen, T. *Structure* **2018**, *26*, 759–766.e4.

Supplemental Figures

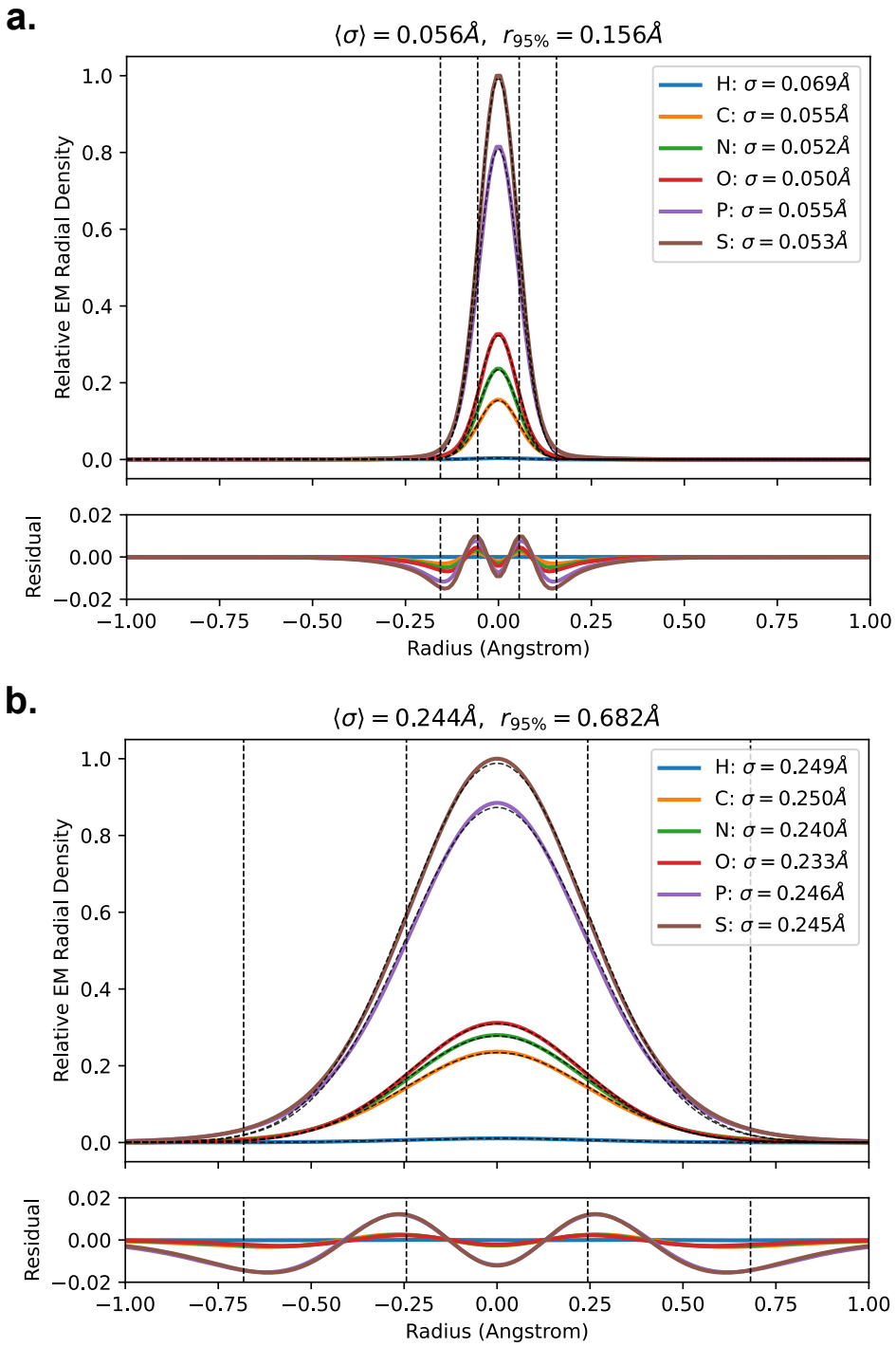

Figure S1: Element-specific Intensity Profiles. Plots of intensity profiles in a transmission electron microscopy image for different elements (color lines) along with Gaussian fits to intensity profiles (dashed lines). **(a)** Profiles calculated using no distortion factor yield an element-averaged fitted width of $\sigma = 0.056$ Å. **(b)** Profiles calculated with a Debye-Waller factor of 5.0 Å yield an element-averaged fitted width of $\sigma = 0.244$ Å.

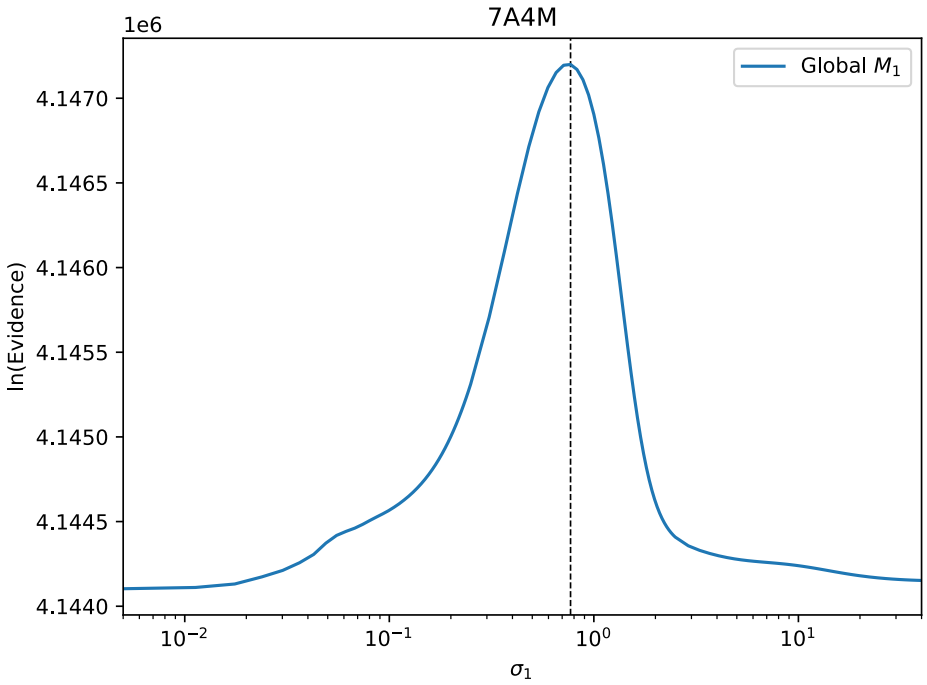

Figure S2: Residue Width probability for Global Residue-level Model. A global M_1 level model of all of the residues in a structure of apoferritin (PDB ID 7A4M) was created, and the probability $P(Y \mid \sigma, X, M_1)$ was calculated as a function of σ varying from 0.05 Å to 40 Å. The maximum is locate at $\sigma = 0.769$ Å.

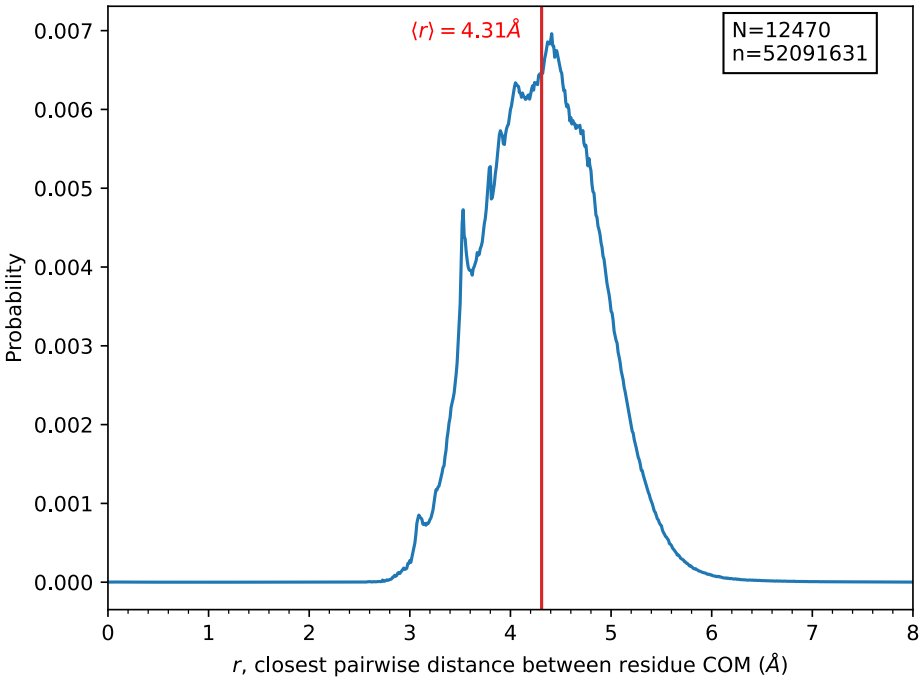

Figure S3: Histogram of distances between nearest residues Each of the cryoEM-derived PDB structures (selection criterion detailed in the main text) was taken, the center-of-mass (COM) locations for each residue in those structures were calculated, the distance between each residue COM in the structure was calculated, and then for each residue the distance to the closest residue was determined. For each structure, those closest distances were histogrammed between $r = 0.0$ Å to $r = 20.0$ Å with 0.1 Å bin widths. All of those histograms were taken, each was normalized, and then the average normalized counts in each bin was plotted here. The expectation value of the closest distance based on this histogram is marked by a vertical line at $\langle r \rangle = 4.31$ Å.

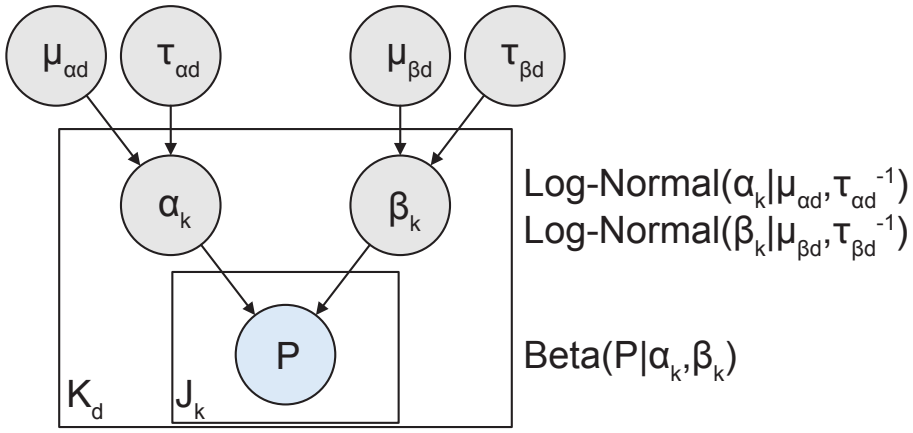

Figure S4: Schematic Diagram of the Statistical Model. Plate diagram of the directed acyclic graph (DAG) of the statistical model used for capturing the dispersion in P_{atomic} within a set of structures.

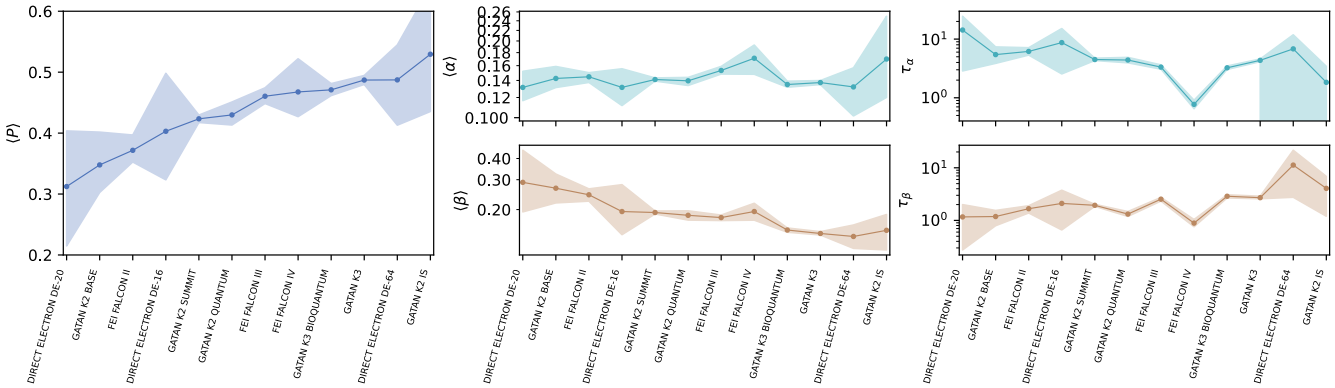

Figure S5: Statistical Model of Camera-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the microscope detector. Structures were grouped by their PDBx/mmCIF metadata entry *_em_image_recording.film_or_detector_model*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 $\mathring{\mathrm{A}}$, and that were deposited into the PDB in 2018 or later were used. Groups with less than five structures were not analyzed. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

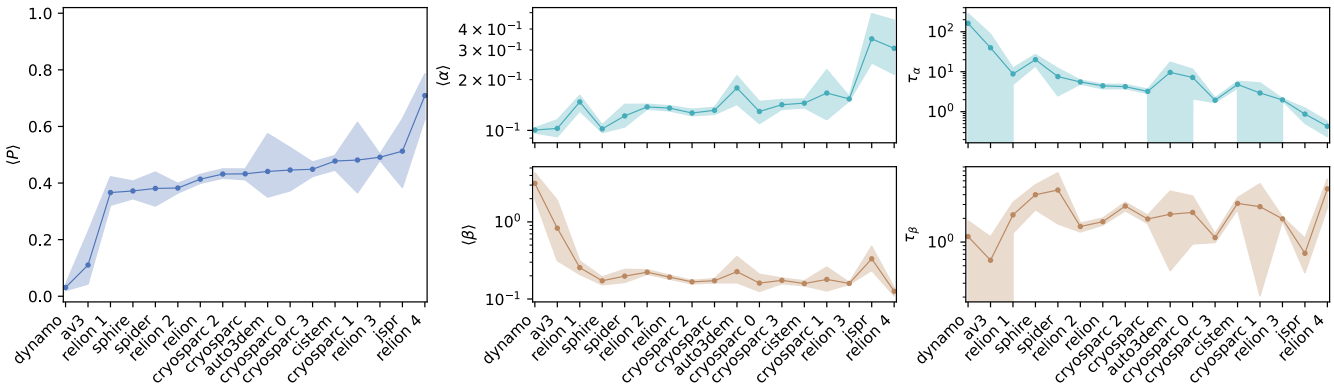

Figure S6: Statistical Model of Reconstruction Software-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the reconstruction software. Structures were grouped by their PDBx/mmCIF metadata entry *software name - reconstruction* in the category *_em_software.category*. Software was grouped by major version (*e.g.*, Relion 3.1 $\rightarrow$ Relion 3). Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 $\mathring{\text{A}}$, and that were deposited into the PDB in 2018 or later were used. Groups with less than five structures were not analyzed. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

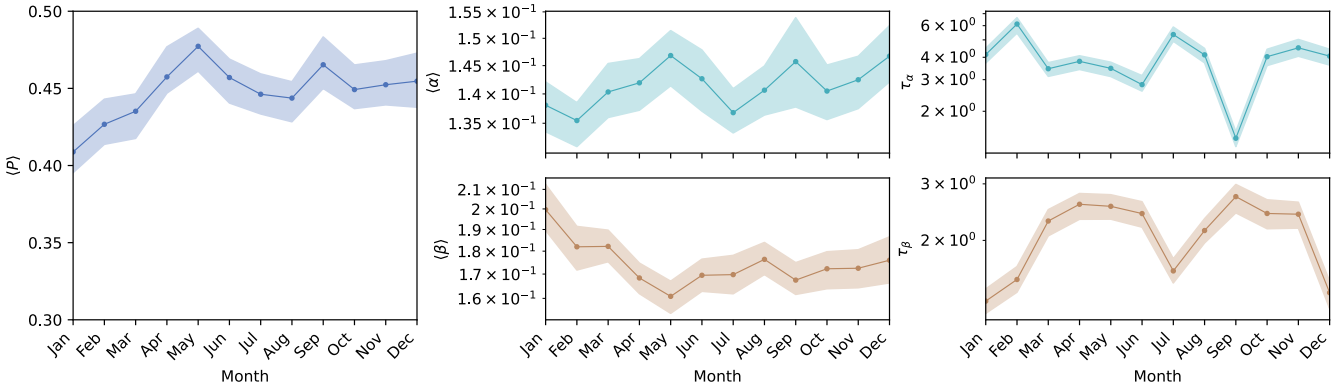

Figure S7: Statistical Model of Deposit Month-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the month of the year that a structure was deposited into the PDB. Structures were grouped by the month of their PDBx/mmCIF metadata entry *_pdbx_database_status.recvd_initial_deposition_date*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 Å, and that were deposited into the PDB in 2018 or later were used. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

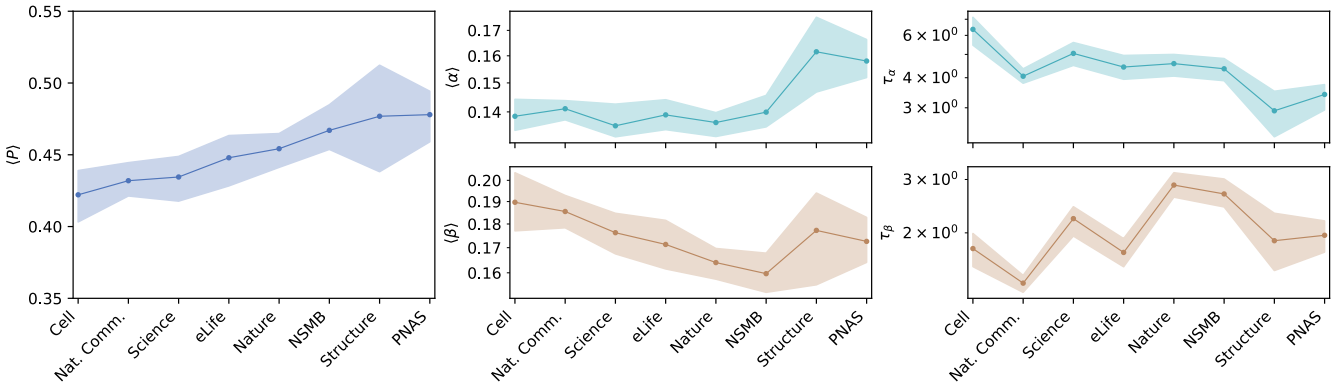

Figure S8: Statistical Model of Publication Journal-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the journal in which the associated paper was published. Structures were grouped by their PDBx/mmCIF metadata entry *_citation.journal_abbrev* and only those shown were analyzed. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 Å, and that were deposited into the PDB in 2018 or later were used. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

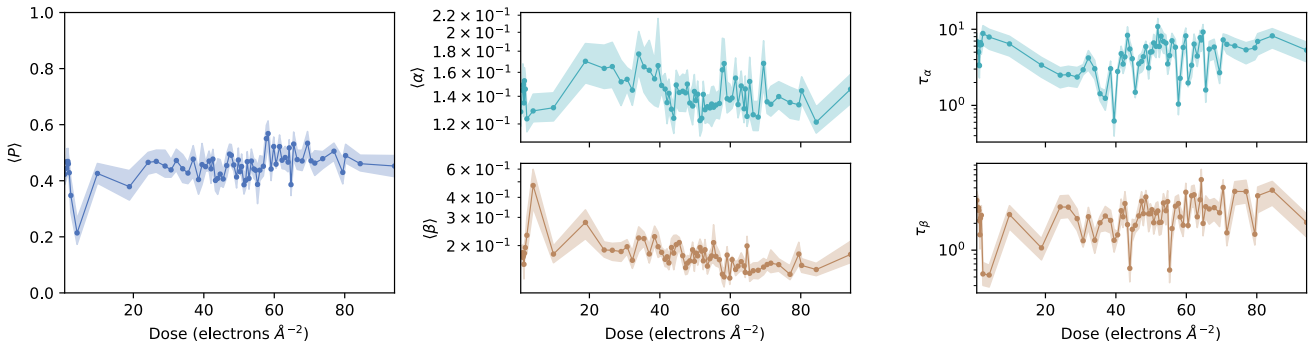

Figure S9: Statistical Model of Electron dose-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the electron dose (electrons per square Angstrom) used during imaging. Structures were grouped by their PDBx/mmCIF metadata entry *_em_image_recording.avg_electron_dose_per_image*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 Å, and that were deposited into the PDB in 2018 or later were used. Groups with less than five structures were not analyzed. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

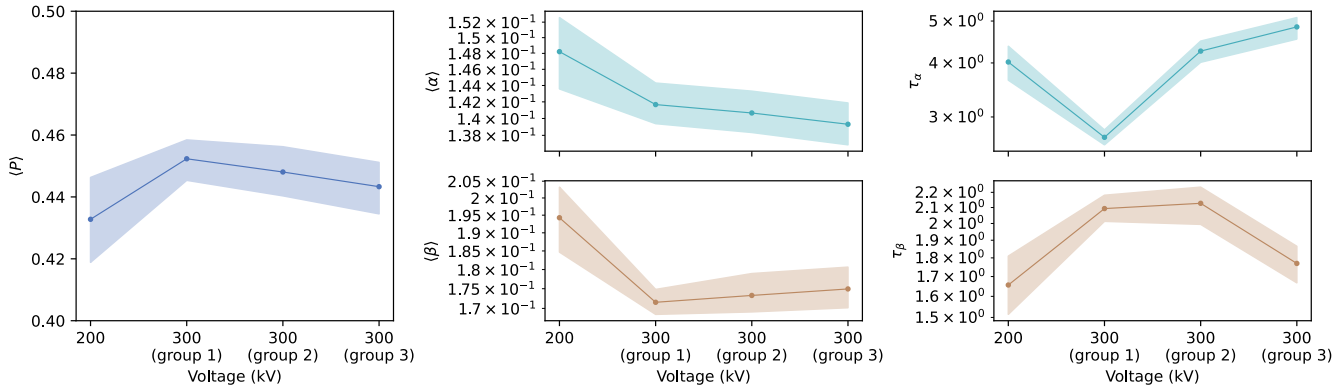

Figure S10: Statistical Model of Accelerating Voltage-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the electron accelerating dose (kV) used during imaging. Structures were grouped by their PDBx/mmCIF metadata entry *_em_imaging.accelerating_voltage*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 Å, and that were deposited into the PDB in 2018 or later were used. Groups with less than five structures were not analyzed; group sizes were capped at 3000 structures and split when more were present. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

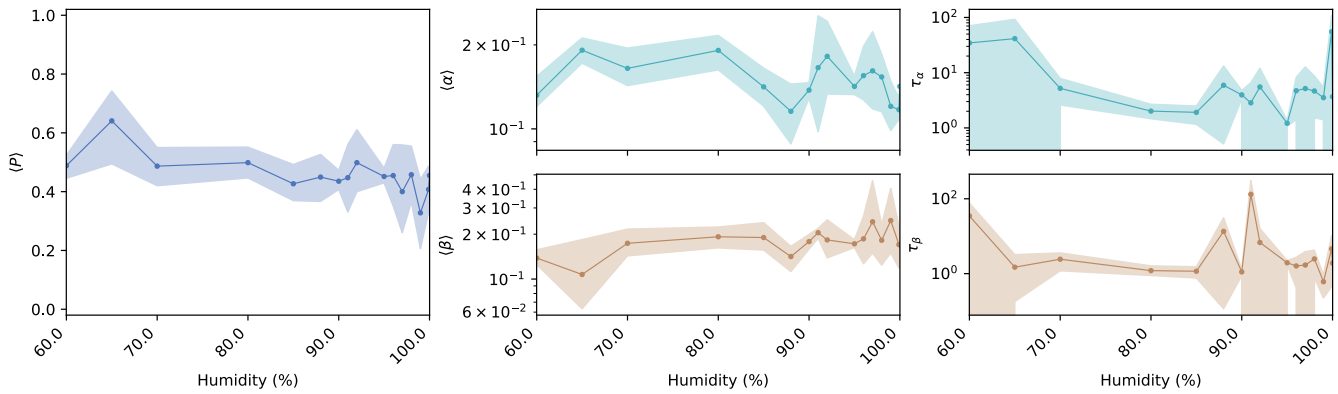

Figure S11: Statistical Model of Humidity-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the humidity during vitrification. Structures were grouped by their PDBx/mmCIF metadata entry *_em_vitrification.humidity*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 8.0 Å, and that were deposited into the PDB in 2018 or later were used. Groups with less than five structures were not analyzed. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

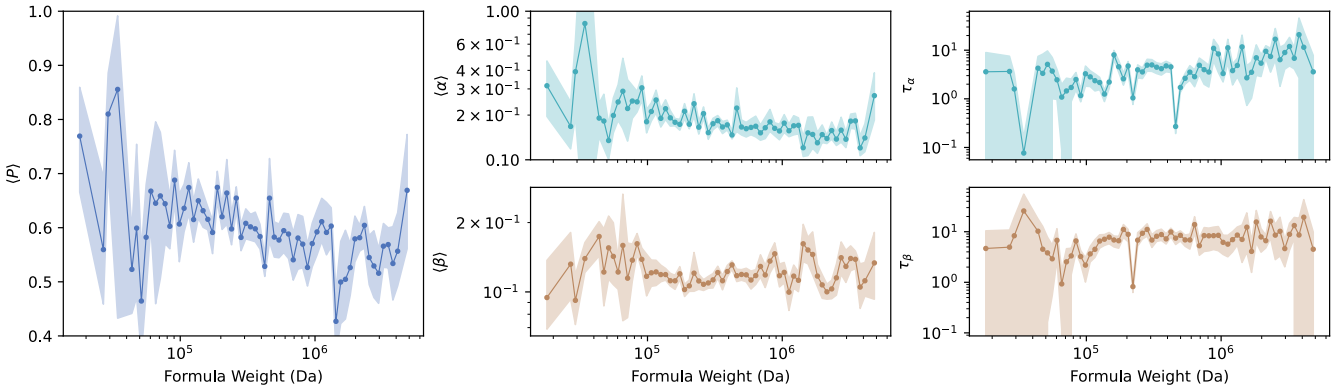

Figure S12: Statistical Model of Formula Weight-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the formula weight of the molecule and/or molecular complex. Structures were grouped by their PDBx/mmCIF metadata entry *_entity.formula_weight*. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 3.2 Å were used. Groups with less than five structures were not analyzed. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

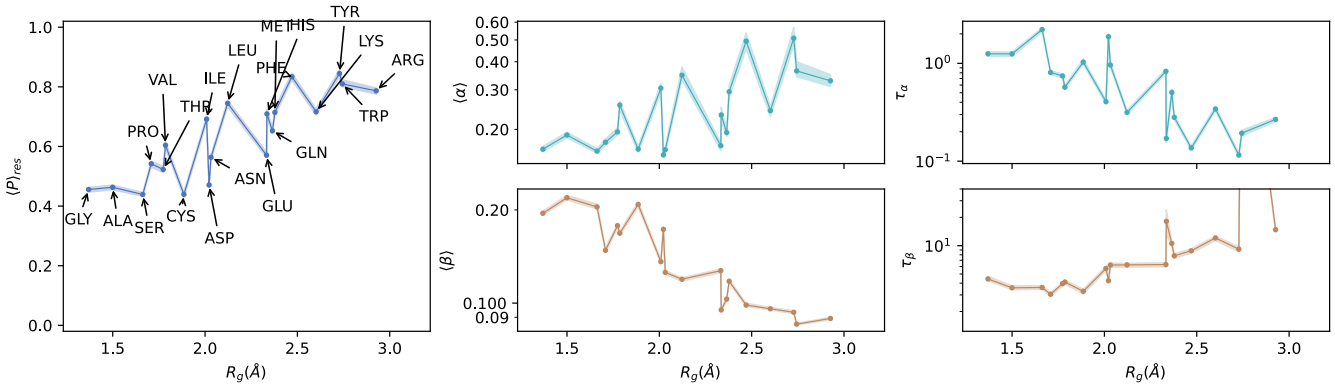

Figure S13: Statistical Model of Amino Acid Size-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the amino acid residue size, which was quantified by its radius of gyration, R_g. For each residue, R_g was calculated as the mean R_g of that residue in a high-resolution, bacterial ribosome structure (PDB ID: 8B0X, 1.55 Å) as $R_g = \sqrt{\sum_i (m_i r_i^2)/(\sum_i m_i)}$, where m_i is the mass of the i^{th} atom and r_i^2 is the square of the distance from the center-of-mass of the residue of the i^{th} atom. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 3.2 Å were used, and groups were formed by choosing all of each type of that residue within each structure. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv_{res} \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

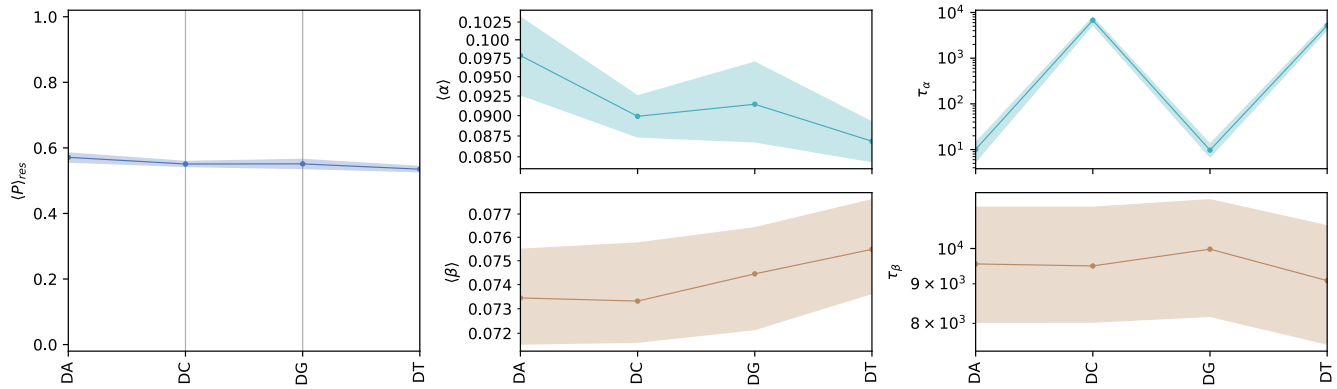

Figure S14: Statistical Model of DNA Residue identity-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the DNA residue identity. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 3.2 Å were used, and groups were formed by choosing all of each type of that residue within each structure. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv_{res} \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

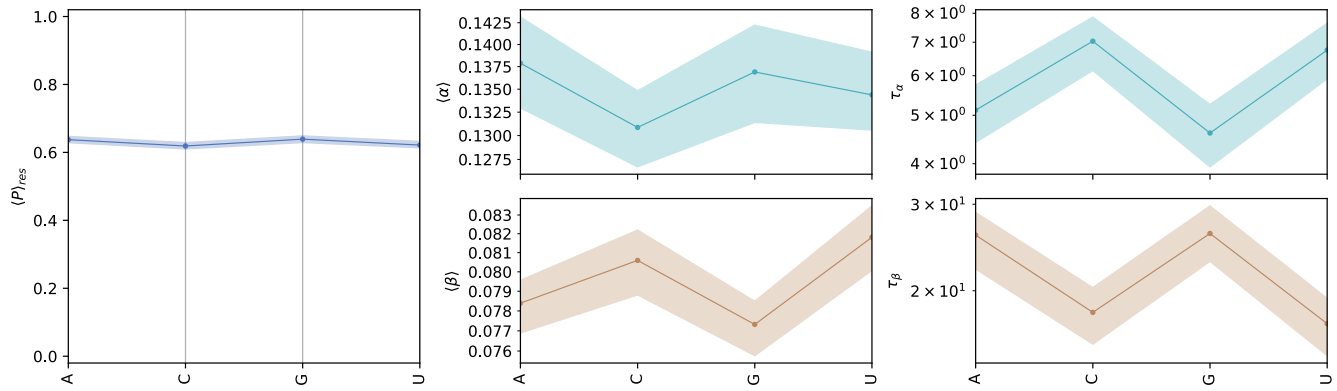

Figure S15: Statistical Model of RNA Residue identity-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of the RNA residue identity. Only structures released before Jan. 1, 2023 with reported FSC resolution less than 3.2 Å were used, and groups were formed by choosing all of each type of that residue within each structure. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv_{res} \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

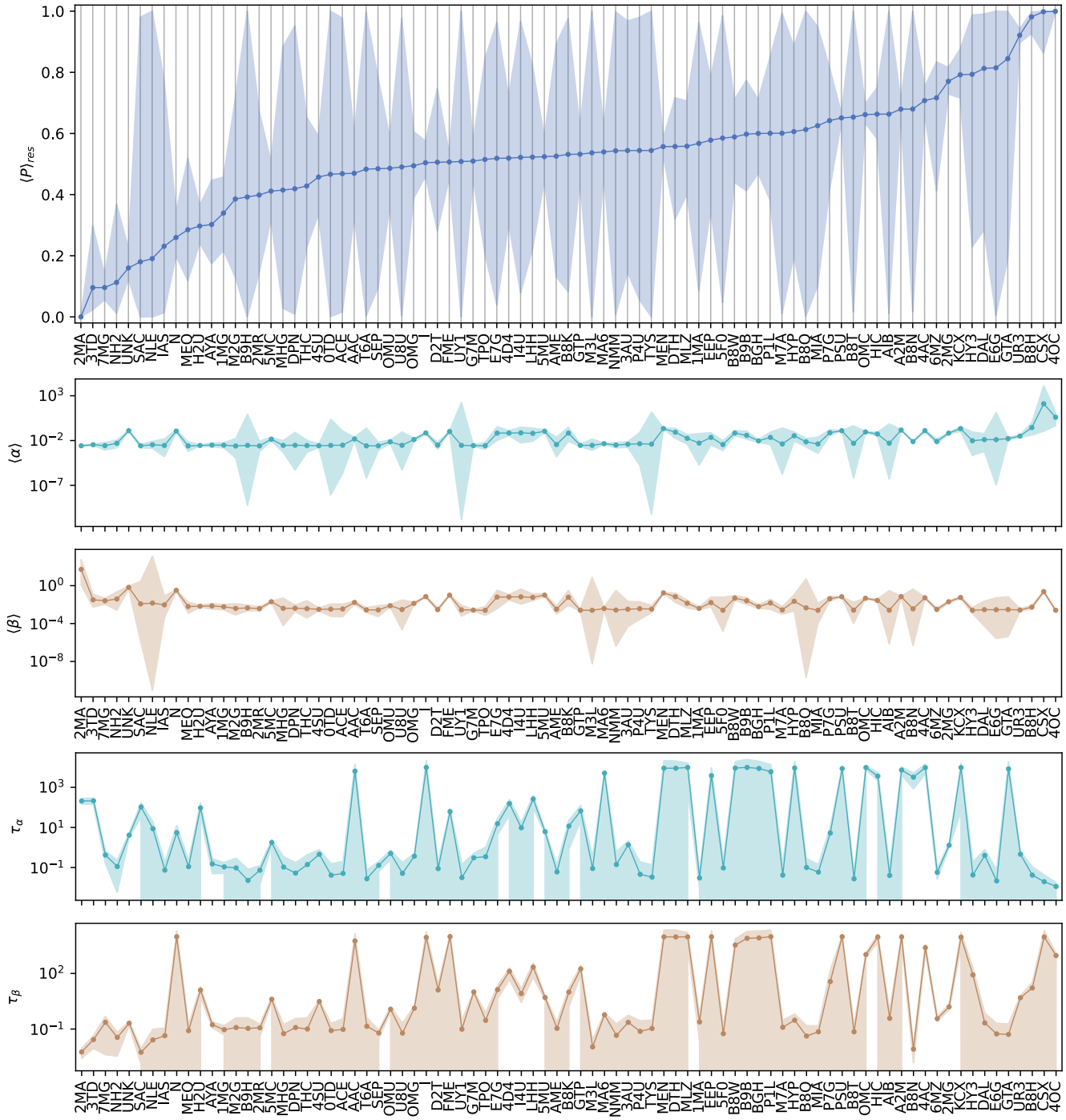

Figure S16: Statistical Model of Modified Residue Identity-dependence of Atomic Resolution. The statistical model for HARP results was used to analyze the effect of modified residues (*e.g.*, post-translational modifications). Only structures released before Jan. 1, 2023 with reported FSC resolution less than 3.2 Å were used, and groups were formed by choosing all of each type of that residue within each structure. Plots are of MAP parameters (dots) and 95% HPDI (shaded regions), while lines are provided to guide the eye. $\langle P \rangle \equiv_{res} \langle \alpha \rangle / (\langle \alpha \rangle + \langle \beta \rangle)$, $\langle \alpha \rangle \equiv \exp[\mu_\alpha]$, and $\langle \beta \rangle \equiv \exp[\mu_\beta]$.

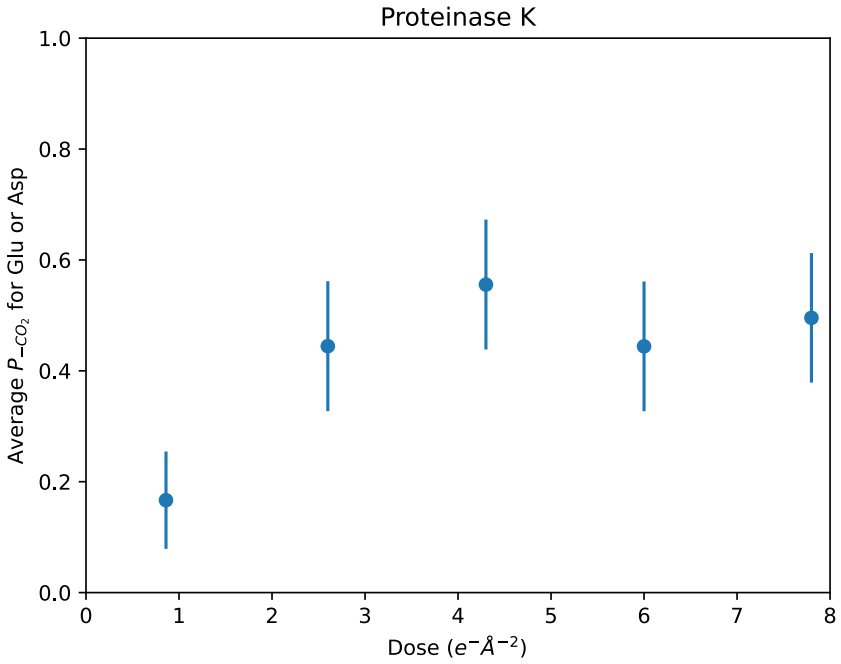

Figure S17: Electron Dose Dependence of Asp and Glu Decarboxylation in Proteinase K. Plot of average P_{-CO2} for all glutamic acid or aspartic acid residues *vs.* electron dose for Proteinase K from the micro-crystal electron diffraction study of Gonen and coworkers (PDB IDs: 6CL7, 6CL8, 6CL9, 6CLA, and 6CLB) [12]. P_{-CO2} is calculated by model selection using equal *a prior* model priors for an M_0 model of the residue with the side-chain carboxylate group present (*i.e.*, $M_{0,+CO2}$) and again with the side-chain carboxylate group removed and the 'CD' carbon (Glu) or the 'CG' carbon (Asp) replaced with a hydrogen atom (*i.e.*, $M_{0,-CO2}$). A P_{atomic}-like calculation is performed for each residue, including marginalizing σ out, as $P_{-CO2} = 1./\left(1 + \int(P(Y \mid X_{+CO2}))/\int(P(Y \mid X_{-CO2}))\right)$. Data points are the mean $1/N \sum_{\hat{T}(R_j)\in\{Glu,Asp\}} P_{i,-CO2}$, and error bars are ± 1 standard error of the mean (SEM).