bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Wilson Statistics: Derivation, Generalization, and Applications to Electron Cryomicroscopy

Amit Singer*

May 14, 2021

Abstract

The power spectrum of proteins at high frequencies is remarkably well described by the flat Wilson statistics. Wilson statistics therefore plays a significant role in X-ray crystallography and more recently in electron cryomicroscopy (cryo-EM). Specifically, modern computational methods for three-dimensional map sharpening and atomic modelling of macromolecules by single particle cryo-EM are based on Wilson statistics. Here we provide the first rigorous mathematical derivation of Wilson statistics. The derivation pinpoints the regime of validity of Wilson statistics in terms of the size of the macromolecule. Moreover, the analysis naturally leads to generalizations of the statistics to covariance and higher order spectra. These in turn provide theoretical foundation for assumptions underlying the widespread Bayesian inference framework for three-dimensional refinement and for explaining the limitations of autocorrelation based methods in cryo-EM.

1 Introduction

The power spectrum of proteins is often modelled by the Guinier law at low frequencies and the Wilson statistics at high frequencies. At low frequencies, there is a quadratic decay of the power spectrum characterized by the moment of inertia of the molecule (e.g., its radius of gyration). At high frequencies, the power spectrum is approximately flat. In structural biology, it is customary to plot the logarithm of the spherically-averaged power spectrum of a three-dimensional structure as a function of the squared spatial frequency. This Guinier plot typically depicts the two different frequency

*Department of Mathematics and PACM, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, amits@math.princeton.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

regimes. It is not surprising that these laws are of critical importance in structural biology, with applications in X-ray crystallography [1] and cryo-EM [2]. However, while Guinier law has a very simple mathematical derivation based on a Taylor expansion, in the literature we could only find heuristic arguments in support of Wilson statistics, such as the original argument provided by Wilson in his seminal 1-page Nature paper [3]. Here we provide a rigorous mathematical derivation of Wilson statistics in the form of Theorem 3 and derive other forms of statistics with potential application to cryo-EM. The main ingredients to our analysis are a scaling argument, basic probability theory, and modern results in Fourier analysis that have found various applications within mathematics (such as the distribution of lattice points in domains), but their application to structural biology appears to be new.

1.1 Random bag of atoms

The model underlying Wilson statistics is a random "bag of atoms", where the random "protein" consists of N atoms whose locations $X_1, X_2, \ldots, X_N$ are independent and identically distributed (i.i.d). For example, each X_i could be uniformly distributed inside a container $\Omega \subset \mathbb{R}^3$ such as a cube or a ball, though other shapes and non-uniform distributions are also possible. The electron scattering potential $\phi : \mathbb{R}^3 \to \mathbb{R}$ of the protein is modelled as

$$\phi(x) = \sum_{i=1}^{N} f(x - X_i) \tag{1}$$

where f is a bump function such as a Gaussian, or a delta function in the limit of an ideal point mass. For simplicity of exposition, we assume that the atoms are identical. Otherwise, one can use different f's to describe the scattering from each atom type. The Fourier transform of (1) is given by

$$\hat{\phi}(\xi) = \sum_{i=1}^{N} \hat{f}(\xi) e^{-2\pi\imath\langle \xi, X_i\rangle} = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, X_i\rangle}. \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1.2 Wilson statistics

Wilson's original argument [3] uses (2) to evaluate the power spectrum as follows

$$
\begin{aligned}
|\hat{\phi}(\xi)|^2 &= |\hat{f}(\xi)|^2 \left| \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, X_i\rangle} \right|^2 \\
&= |\hat{f}(\xi)|^2 \left(\sum_{i,j=1}^{N} e^{-2\pi\imath\langle \xi, (X_i - X_j)\rangle} \right) \\
&= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} e^{-2\pi\imath\langle \xi, (X_i - X_j)\rangle} \right) \qquad (3) \\
&\approx N|\hat{f}(\xi)|^2. \qquad (4)
\end{aligned}
$$

Wilson argued that the sum of the complex exponentials in (3) is negligible compared to N, as those terms wildly oscillate and cancel each other, especially for high frequency ξ. We shall make this hand wavy argument more rigorous and the term "high frequency" mathematically precise. Note that for an ideal point mass $\hat{f}(\xi) = 1$, and (4) implies that the power spectrum is flat, i.e., $|\hat{\phi}(\xi)|^2 \approx N$.

The challenge is to show that there is so much cancellation that adding $O(N^2)$ oscillating terms of size $O(1)$ in (3) is negligible compared to N. For a random walk, the sum of $O(N^2)$ i.i.d zero-mean random variables of variance $O(1)$ is $O(N)$ (the square-root of the number of terms). In order to show that the sum is negligible compared to N, additional cancellation must be happening. The role that ξ plays also needs to be carefully analyzed, as for $\xi = 0$, clearly $|\hat{\phi}(0)|^2 = N^2$. What is the mechanism by which $|\hat{\phi}(\xi)|^2$ decays from N^2 to N as ξ increases?

2 Derivation of Wilson Statistics

2.1 $N^{1/3}$ scaling

Since $X_1, \ldots, X_N$ are i.i.d, one might be tempted to apply the Central Limit Theorem (CLT) to (2) and conclude that $\hat{\phi}(\xi)$ is approximately a Gaussian, for which the mean and variance can be readily calculated as done in [4]. However, one should proceed with caution, because if the container Ω is fixed, then in the limit $N \to \infty$, the density of the atoms also grows indefinitely, whereas the density of atoms in a protein is clearly bounded. If the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

density of the atoms is to be kept fixed, the container Ω has to grow with N. To make this dependency explicit, we denote the container by Ω_N. The volume of the container Ω_N must be proportional to N. The length-scale is therefore proportional to $N^{1/3}$, that is, $\Omega_N = N^{1/3}\Omega_1$, or $X_i = N^{1/3}Y_i$ with $Y_i \sim U(\Omega_1)$ in the uniform case, and i.i.d in general. The Fourier transform (2) is rewritten as

$$\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, N^{1/3}Y_i\rangle}, \tag{5}$$

but now the CLT can no longer be applied in a straightforward manner, because the summands in (5) are random variables that depend on N.

2.2 Shape of container and decay rate of the Fourier transform

The representation (5) facilitates the calculation of any moment of $\hat{\phi}(\xi)$. The expectation (first moment) of $\hat{\phi}(\xi)$ is given by

$$\begin{aligned} \mathbb{E}[\hat{\phi}(\xi)] &= N\hat{f}(\xi)\mathbb{E}[e^{-2\pi\imath\langle \xi, N^{1/3}Y_1\rangle}] \\ &= N\hat{f}(\xi)\int_{\mathbb{R}^3} e^{-2\pi\imath\langle \xi, N^{1/3}y\rangle} g(y)\, dy \\ &= N\hat{f}(\xi)\hat{g}(N^{1/3}\xi), \end{aligned} \tag{6}$$

where $g(y)$ is the probability density function of Y_1, and $\hat{g}$ is its Fourier transform. The dependency on $\hat{g}(N^{1/3}\xi)$ and N being a large parameter together suggest that the decay rate of $\hat{g}$ at high frequencies is critical for analyzing Wilson statistics.

Different container shapes and choices of g can lead to different behavior of its Fourier transform $\hat{g}$. Before stating known theoretical results, it is instructive to consider a couple of examples.

- A uniform distribution in a ball. Here Ω_1 is a ball of radius 1, denoted B, and the uniform density is $g_B(x) = \frac{1}{4\pi/3}\chi_B(x)$, where χ_B is the characteristic function of the ball. It is a radial function, a property that can readily be used to calculate its Fourier transform as

$$\hat{g}_B(\xi) = -\frac{3\cos(2\pi|\xi|)}{4\pi^2|\xi|^2} + \frac{3\sin(2\pi|\xi|)}{8\pi^3|\xi|^3}. \tag{7}$$

  In particular, (7) implies that $|\hat{g}_B(\xi)| \leq \frac{C}{|\xi|^2}$ for some constant C.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

- A uniform distribution in a cube. Here $\Omega_1 = [-\frac{1}{2}, \frac{1}{2}]^3$ is the unit cube, and g_C is a product of three rectangular window functions whose Fourier transform is the sinc function. As a result,

$$\hat{g}_C(\xi) = \prod_{i=1}^{3} \mathrm{sinc}(\xi_i) = \prod_{i=1}^{3} \frac{\sin(\pi\xi_i)}{\pi\xi_i}. \tag{8}$$

  Taking ξ along one of the axes, e.g., $\xi = (|\xi|, 0, 0)$ gives $\hat{g}_C(|\xi|, 0, 0) = \frac{\sin(\pi|\xi|)}{\pi|\xi|}$. In this case, $|\hat{g}_C(\xi)| \leq \frac{C}{|\xi|}$ for some $C > 0$. Notice that the decay of $\hat{g}_C$ in directions not normal to its faces is faster. For example, for $\xi = \frac{1}{\sqrt{3}}(|\xi|, |\xi|, |\xi|)$ we have $|\hat{g}_C(\frac{1}{\sqrt{3}}(|\xi|, |\xi|, |\xi|))| = \frac{|\sin^3(\frac{1}{\sqrt{3}}\pi|\xi|)|}{(\frac{1}{\sqrt{3}}\pi|\xi|)^3} \leq \frac{C}{|\xi|^3}$.

We are now ready to state existing theoretical results about the decay rate of the Fourier transform for containers of general shape.

Theorem 1. *(see [5, p. 336])*

1. *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded region whose boundary $M = \partial\Omega$ has non-vanishing Gauss curvature at each point, then*

$$|\hat{\chi}_\Omega(\xi)| = O(|\xi|^{-\frac{d+1}{2}}), \quad \text{as } |\xi| \to \infty. \tag{9}$$

2. *If M has m non-vanishing principal curvatures at each point, then*

$$|\hat{\chi}_\Omega(\xi)| = O(|\xi|^{-(m+2)/2}), \quad \text{as } |\xi| \to \infty. \tag{10}$$

The decay rates previously observed for the three-dimensional ball ($d = 3$ or $m = 2$) and the cube ($m = 0$) are particular cases of Theorem 1.

Although the decay rate in different directions could be different (as the example of the cube illustrates), for a large family of containers (convex sets and open sets with sufficiently smooth boundary surface), the following Theorem asserts that the spherical average of the power spectrum has the same decay rate as that of the ball.

Theorem 2. *(see [6]) Suppose $\Omega \subset \mathbb{R}^d$ is a convex body or an open bounded set whose boundary $\partial\Omega$ is $C^{3/2}$. Then,*

$$\int_{S^{d-1}} |\hat{\chi}_\Omega(k\omega)|^2 \, d\omega = O(k^{-(d+1)}), \quad \text{as } k \to \infty. \tag{11}$$

Here $k = |\xi|$ is the radial frequency and S^{d-1} is the unit sphere in $\mathbb{R}^d$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.3 Validity regime of Wilson statistics

We are now in position to state and prove our main result that fully characterizes the regime of validity of Wilson statistics.

Theorem 3. *1. For the random bag of atoms model, the expected power spectrum is given by*

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right). \tag{12}$$

2. If the container is a convex body or an open set with a $C^{3/2}$ boundary surface, and the atom locations are uniformly distributed in the container, then the expected spherically-averaged power spectrum satisfies

$$\mathbb{E}\left[\frac{1}{4\pi}\int_{S^2} |\hat{\phi}(k\omega)|^2\, d\omega\right] = |\hat{f}(k)|^2 \left(N + o(N)\right), \tag{13}$$

for $k \gg N^{-1/12}$.

3. If the Fourier transform of the density g satisfies $|\hat{g}(\xi)| \leq C|\xi|^{-\alpha}$, then

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + o(N)\right), \quad \text{for } |\xi| \gg N^{-\beta}, \tag{14}$$

where $\beta = \dfrac{2\alpha - 3}{6\alpha}$.

Proof. Starting with Wilson's original approach, from (5) it follows that the power spectrum of ϕ is given by

$$\begin{aligned} |\hat{\phi}(\xi)|^2 &= |\hat{f}(\xi)|^2 \sum_{i,j=1}^{N} e^{-2\pi\imath\langle \xi, N^{1/3}(Y_i - Y_j)\rangle} \\ &= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} e^{-2\pi\imath\langle \xi, N^{1/3}(Y_i - Y_j)\rangle}\right). \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Since the Y_i's are i.i.d, the expected power spectrum satisfies

$$\begin{aligned}
\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] &= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} \mathbb{E}\left[e^{-2\pi\imath\langle \xi, N^{1/3}(Y_i - Y_j)\rangle}\right]\right) \\
&= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} \mathbb{E}\left[e^{-2\pi\imath\langle \xi, N^{1/3}Y_i\rangle}\right] \overline{\mathbb{E}\left[e^{2\pi\imath\langle \xi, N^{1/3}Y_j\rangle}\right]}\right) \\
&= |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\mathbb{E}\left[e^{-2\pi\imath\langle \xi, N^{1/3}Y\rangle}\right]\right|^2\right) \\
&= |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right), \qquad (15)
\end{aligned}$$

establishing (12). Assuming f (hence also $\hat{f}$) are radial functions, the expectation of the spherically-averaged power spectrum satisfies

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{4\pi}\int_{S^2} |\hat{\phi}(k\omega)|^2\, d\omega\right] \qquad (16) \\
&= |\hat{f}(k)|^2 \left(N + N(N-1)\frac{1}{4\pi}\int_{S^2} \left|\hat{g}(N^{1/3}k\omega)\right|^2\, d\omega\right).
\end{aligned}$$

Theorem 2 with $d = 3$ implies

$$N(N-1)\frac{1}{4\pi}\int_{S^2} \left|\hat{g}(N^{1/3}k\omega)\right|^2\, d\omega = O(N^{2/3}k^{-4}). \qquad (17)$$

This term is negligible compared to N in (16) for $k \gg N^{-1/12}$, proving (13). Finally, if $|\hat{g}(\xi)| \leq C|\xi|^{-\alpha}$, then $N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2 = O(N^{2-\frac{2\alpha}{3}}|\xi|^{-2\alpha})$, which is $o(N)$ for $|\xi| \gg N^{-\frac{\frac{2\alpha}{3}-1}{2\alpha}}$. □

Theorem 3 (specifically, (13)) implies that the transition in Guinier plot occurs around $k_0 = N^{-1/12}$, and for higher radial frequencies the spherically-averaged power spectrum is approximately flat. Suppose a protein of radius 100Å with $N = 10^4$ atoms. In dimensionless units, where each unit cell occupies a single atom on average, Wilson statistics holds for frequencies above $k_0 = (10^4)^{-1/12} = 10^{-1/3}$. The radius R_0 of the unit cell satisfies $10^4 R_0^3 = 100^3 \text{Å}^3$. Therefore, $R_0 = 10^{-4/3}100\text{Å}$. The cutoff frequency is therefore $k_0/R_0 = 10^{-1/3}10^{4/3}0.01\text{Å}^{-1} = 0.1\text{Å}^{-1}$, corresponding to resolution of 10Å. For a protein of the same radius but with 10 times more atoms

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$N = 10^5$, the cutoff frequency is 0.178Å^{-1}, corresponding to 5.6Å resolution. These resolutions are in agreement with empirical evidence about the validity regime of Wilson statistics [2].

Moreover, (16) and (17) suggest that the spherically-averaged power spectrum decays to its high frequency limit as k^{-4}. This decay rate at high frequencies is reminiscent of Porod's law in SAXS [7, 8]. At first, the 1/12 exponent of the cutoff frequency $k_0 = N^{-1/12}$ might seem mysterious. In hindsight, it is simply the product of the dimension $d = 3$ that resulted in the scaling of $N^{1/3}$ and the decay rate exponent of k^{-4}.

Note that in our derivation of Wilson statistics, we first took expectation with respect to the atom positions followed by spherically-averaging the power spectrum. On the other hand, spherically-averaging (3) first gives

$$\frac{1}{4\pi}\int_{S^2} |\hat{\phi}(k\omega)|^2 \, d\omega = |\hat{f}(k)|^2 \left(N + \sum_{i \neq j} \frac{\sin(2\pi k|X_i - X_j|)}{2\pi k|X_i - X_j|} \right) \tag{18}$$

as in Debye's scattering equation [9], due to the identity

$$\frac{1}{4\pi}\int_{S^2} e^{-2\pi\imath\langle k\omega, x\rangle} \, d\omega = \frac{\sin(2\pi k|x|)}{2\pi k|x|}. \tag{19}$$

Although the $1/k$ decay of the sinc function in (18) sheds some light on the mechanism by which the sum over atom pairs decreases with k, it does not seem to provide a good starting point for a rigorous derivation of Wilson statistics, nor does it provide a clear path for the generalizations considered later in this paper.

While Theorem 3 characterizes the expected power spectrum, one may wonder whether the statistical fluctuations of the power spectrum could overwhelm its mean. This turns out not to be the case. Similar to the derivation of Wilson statistics, one can show that if $|\hat{g}(\xi)| \leq C|\xi|^{-2}$ then

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^4\right] = N^2|\hat{f}(\xi)|^4 + o(N^2), \quad \text{for } |\xi| \gg N^{-1/12}. \tag{20}$$

Since $\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = N|\hat{f}(\xi)|^2 + o(N)$ for $|\xi| \gg N^{-1/12}$, it follows that for $|\xi| \gg N^{-1/12}$

$$\mathrm{Var}(|\hat{\phi}(\xi)|^2) = \mathbb{E}\left[|\hat{\phi}(\xi)|^4\right] - \mathbb{E}\left[|\hat{\phi}(\xi)|^2\right]^2 = o(N^2). \tag{21}$$

In other words, the standard deviation of the power spectrum is $o(N)$, so the fluctuation is smaller than the mean value.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3 Generalizations and applications to cryo-EM

3.1 Existing applications to cryo-EM

A common practice in single particle cryo-EM is to apply a filter to the reconstructed map. The filter boosts medium and high frequencies such that the power spectrum of the sharpened map is approximately flat and consistent with Wilson statistics [2, 10]. The filter is an exponentially growing filter whose parameter is estimated using the Guinier plot. The boost of medium and high frequency components increases the contrast of many structural features of the map and helps to model the atomic structure. This is the so-called B-factor correction, B-factor flattening, or B-factor sharpening. It is a tremendously effective method to increase the interpretability of the reconstructed map. In fact, most map depositions in the Electron Microscopy Data Base (EMDB) only contain sharpened maps [11]. Map sharpening is still an active area of research and method development, see, e.g. [12, 13] and references therein. Wilson statistics is also used to reason about and extrapolate the number of particles required to high resolution [2].

3.2 Generalization of Wilson statistics to covariance with application to 3-D iterative refinement

We now highlight a certain generalization of Wilson statistics with potential application to 3-D iterative refinement, arguably the main component of the computational pipeline for single particle analysis [14]. Specifically, the Bayesian inference framework underlying the popular software toolbox RELION [15] requires the covariance matrix of $\hat{\phi}$ and approximates it with a diagonal matrix [16]. For tractable computation, the variance (the diagonal of the covariance matrix) is further assumed to be a radial function.

The random bag of atoms model underlying Wilson statistics provides the covariance matrix

$$\mathrm{Cov}[\hat{\phi}](\xi_1, \xi_2) = \mathbb{E}[\hat{\phi}(\xi_1)\overline{\hat{\phi}(\xi_2)}] - \mathbb{E}[\hat{\phi}(\xi_1)]\mathbb{E}[\overline{\hat{\phi}(\xi_2)}] \tag{22}$$

in closed form as

$$\begin{aligned} &\mathrm{Cov}[\hat{\phi}](\xi_1, \xi_2) \\ &= N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[\hat{g}(N^{1/3}(\xi_1 - \xi_2)) - \hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right]. \end{aligned} \tag{23}$$

Before proving this result, note that it implies a vast reduction in the number of parameters needed to describe the covariance matrix. In general,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for a 3-D map represented as an array of L^3 voxels, the covariance matrix is of size $L^3 \times L^3$ which requires $O(L^6)$ entries, which is prohibitively large. However, (23) suggests that the covariance depends on only $O(L^3)$ parameters. Furthermore, approximating $\hat{g}(\xi)$ by a radial function implies that the covariance depends on just $O(L)$ parameters, the same number of parameters in the existing Bayesian inference method for 3-D iterative refinement. Moreover, comparing the two terms in (23), the decay of $\hat{g}$ implies that $|\hat{g}(N^{1/3}(\xi_1 - \xi_2))| \gg |\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}|$ whenever $|\xi_1|, |\xi_2| \gg N^{-1/3}$. Therefore, for $|\xi_1|, |\xi_2| \gg N^{-1/3}$

$$\mathrm{Cov}[\hat{\phi}](\xi_1, \xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\hat{g}(N^{1/3}(\xi_1 - \xi_2))(1 + o(1)). \tag{24}$$

Since $\hat{g}(N^{1/3}(\xi_1 - \xi_2))$ is largest for $\xi_1 = \xi_2$ and decays with increasing distance $|\xi_1 - \xi_2|$, it follows from (24) that the covariance matrix restricted to frequencies above $N^{-1/3}$ is approximately a band matrix with bandwidth $O(N^{-1/3})$, such that the diagonal is dominant and matrix entries decay when moving away from the diagonal. Note that $N^{-1/3}$ is a very low frequency corresponding to resolution of the size of the protein (as implied by the $N^{1/3}$ scaling). Therefore, the covariance is well approximated by a band matrix with a very small number of diagonals. This serves as a theoretical justification for the diagonal approximation in the Bayesian inference framework [16], as correlations of Fourier coefficients with $|\xi_1 - \xi_2| \gg N^{-1/3}$ are negligible. On the flip side, correlations for which $|\xi_1 - \xi_2| \ll N^{-1/3}$ should not be ignored and correctly accounting for them could potentially lead to further improvement of [16].

To prove (23), we evaluate the two terms in the right hand side of (22) separately. The second term is directly obtained from (6) as

$$\mathbb{E}[\hat{\phi}(\xi_1)]\mathbb{E}[\overline{\hat{\phi}(\xi_2)}] = N^2\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}. \tag{25}$$

To evaluate the first term, we substitute $\hat{\phi}(\xi_1)$ and $\hat{\phi}(\xi_2)$ by (5), separate the summation into diagonal terms ($i = j$) and off-diagonal terms ($i \neq j$)

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

as in Wilson's original argument, and use that Y_i's are i.i.d, resulting

$$\begin{aligned}
&\mathbb{E}[\hat{\phi}(\xi_1)\overline{\hat{\phi}(\xi_2)}] \\
&= \hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\mathbb{E}\left[\sum_{i=1}^{N} e^{-2\pi\imath\langle N^{1/3}\xi_1, Y_i\rangle} \sum_{j=1}^{N} e^{2\pi\imath\langle N^{1/3}\xi_2, Y_j\rangle}\right] \\
&= \hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\mathbb{E}\left[\sum_{i=1}^{N} e^{-2\pi\imath\langle N^{1/3}(\xi_1-\xi_2), Y_i\rangle}\right. \\
&\qquad\qquad \left. + \sum_{i\neq j} e^{-2\pi\imath\langle N^{1/3}\xi_1, Y_i\rangle} e^{2\pi\imath\langle N^{1/3}\xi_2, Y_j\rangle}\right] \\
&= \hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[N\hat{g}(N^{1/3}(\xi_1-\xi_2))\right. \\
&\qquad \left. + N(N-1)\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right]. \qquad (26)
\end{aligned}$$

Subtracting (25) from (26) proves (23). This is a generalization of Wilson statistics, as setting $\xi_1 = \xi_2$ reduces (26) to (12).

Note that the diagonal of the covariance matrix satisfies

$$\mathrm{Var}(\hat{\phi}(\xi)) = \mathrm{Cov}[\hat{\phi}](\xi,\xi) = N|\hat{f}(\xi)|^2\left[1 - |\hat{g}(N^{1/3}\xi)|^2\right]. \qquad (27)$$

The variance vanishes for $\xi = 0$ because $\hat{\phi}(0) = N$ regardless of the atoms positions. The small variance at very low frequencies shares the same origins of Guinier law.

In existing Bayesian inference approaches [16], the mean of each frequency voxel is assumed to be zero. However, comparing (6) and (27) for the mean $\mathbb{E}[\hat{\phi}(\xi)]$ and the variance $\mathrm{Var}(\hat{\phi}(\xi))$, we see that the variance dominates the squared mean only for $|\xi| \gg N^{-1/12}$, which is the validity regime of Wilson statistics. It follows that it is justified to assume a zero-mean signal only for high frequencies, but not at low frequencies. Including an explicit (approximately radial) non-zero mean in the Bayesian inference framework may therefore bring further improvement.

3.3 Generalization of Wilson statistics to higher order spectra with application to autocorrelation analysis

Autocorrelation analysis, originally proposed by Kam [17, 18], has recently found revived interest in X-ray free electron laser (XFEL) [19, 20, 21] and cryo-EM [22, 23, 24]. In autocorrelation analysis, the three-dimensional

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

molecular structure is determined from the correlation statistics of the noisy images. Typically, the second or third order correlation functions are sufficient in principle to uniquely determine the structure [25, 22]. It is therefore of interest to derive a third order statistics analogue of (12). Specifically, $\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right]$ is given by

$$\begin{aligned}
&\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right] \\
&= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3) \\
&\quad \mathbb{E}\left[\sum_{i,j,k=1}^{N} e^{-2\pi\imath\langle\xi_1, N^{1/3}Y_i\rangle} e^{-2\pi\imath\langle\xi_2, N^{1/3}Y_j\rangle} e^{-2\pi\imath\langle\xi_3, N^{1/3}Y_k\rangle}\right] \\
&= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3)\left[N\hat{g}\left(N^{1/3}(\xi_1+\xi_2+\xi_3)\right)\right. \\
&\quad +N(N-1)\hat{g}\left(N^{1/3}(\xi_1+\xi_2)\right)\hat{g}\left(N^{1/3}\xi_3\right) \\
&\quad +N(N-1)\hat{g}\left(N^{1/3}(\xi_1+\xi_3)\right)\hat{g}\left(N^{1/3}\xi_2\right) \\
&\quad +N(N-1)\hat{g}\left(N^{1/3}(\xi_2+\xi_3)\right)\hat{g}\left(N^{1/3}\xi_1\right) \\
&\quad \left.+\, N(N-1)(N-2)\hat{g}\left(N^{1/3}\xi_1\right)\hat{g}\left(N^{1/3}\xi_2\right)\hat{g}\left(N^{1/3}\xi_3\right)\right].
\end{aligned} \tag{28}$$

This result is obtained by separating the sum over all triplets i, j, k into five groups: $i = j = k$, $i = j \neq k$, $i = k \neq j$, $j = k \neq i$, and $i \neq j \neq k \neq i$.

Similar to the power spectrum $|\hat{\phi}(\xi)|^2$ which is the Fourier transform of the autocorrelation function, the bispectrum $\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(-(\xi_1+\xi_2))$ is the Fourier transform of the triple-correlation function. The bispectrum, like the power spectrum, is also shift-invariant. As such, it plays an important role in various autocorrelation analysis techniques. The expected bispectrum under the random bag of atoms model is obtained by setting $\xi_1+\xi_2+\xi_3 = 0$ in (29)

$$\begin{aligned}
\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right] &= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3) \\
&\left[N + N(N-1)\left|\hat{g}\left(N^{1/3}\xi_1\right)\right|^2\right. \\
&+N(N-1)\left|\hat{g}\left(N^{1/3}\xi_2\right)\right|^2 + N(N-1)\left|\hat{g}\left(N^{1/3}\xi_3\right)\right|^2 \\
&\left.+\, N(N-1)(N-2)\hat{g}\left(N^{1/3}\xi_1\right)\hat{g}\left(N^{1/3}\xi_2\right)\hat{g}\left(N^{1/3}\xi_3\right)\right],
\end{aligned} \tag{29}$$

for $\xi_1 + \xi_2 + \xi_3 = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The bispectrum drops from N^3 for $\xi_1 = \xi_2 = \xi_3 = 0$ to N at high frequencies. This drop is even more pronounced than that of the power spectrum that decreases from N^2 to N. This may lead to numerical difficulties in inverting the bispectrum as it has a large dynamic range, e.g., it spans eight orders of magnitude for $N = 10^4$.

The terms in the first two lines of (29) have similar behavior to the power spectrum (12). The last term depends on the decay rate of $\hat{g}$. If $|\hat{g}(\xi)| \leq C|\xi|^{-2}$ as for the ball, then

$$\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right] = N\hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3) + o(N), \tag{30}$$

for $|\xi_1|, |\xi_2|, |\xi_3| \gg 1$, which can be regarded as a generalization of Wilson statistics (e.g., (13)) to higher order spectra. However, for higher order spectra such as the bispectrum the behavior at high frequencies is more involved. For example, taking ξ_1 and ξ_2 to be high frequencies does not imply $\xi_3 = -(\xi_1 + \xi_2)$ is necessarily a high frequency, as can be readily seen by taking $\xi_2 = -\xi_1$ for which $\xi_3 = 0$. For this particular choice of $\xi_2 = -\xi_1$ the expected bispectrum is always greater than N^2.

4 Discussion

This paper provided the first formal mathematical derivation of Wilson statistics, offered generalizations to other statistics, and highlighted potential applications in structural biology.

The assumption underlying Wilson statistics of independent atom locations is too simplistic as it ignores correlations between atom positions in the protein. It is well known that the power spectrum deviates from Wilson statistics at frequencies that correspond to interatomic distances associated with secondary structure such as α-helices. A more refined model that includes such correlations is beyond the scope of this paper.

From the computational perspective, we note that numerical evaluation of Fourier transforms and power spectra associated with Wilson statistics involves computing sums of complex exponentials of the form (1). These can be efficiently computed as a Type-1 3-D non-uniform fast Fourier transform (NUFFT) [26]. The computational complexity of a naïve procedure is $O(NM)$, where M is the number of target frequencies, whereas the asymptotic complexity of NUFFT is $O(N + M)$ (up to logarithmic factors).

Wilson statistics is an instance of a universality phenomenon: all proteins regardless of their shape and specific atomic positions exhibit a similar power spectrum at high frequencies. From the computational standpoint in

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

cryo-EM, this universality is a blessing and a curse at the same time. On the one hand, it enables to correct the magnitudes of the Fourier coefficients of the reconstructed map so they agree with the theoretical prediction. On the other hand, it implies that the high frequency part of the power spectrum is not particularly useful for structure determination, as it does not discriminate between molecules. The generalization of Wilson statistics to the higher order spectra shows that the bispectrum also becomes flat at high frequencies. These observations may help explain difficulties of the autocorrelation approach as a high resolution reconstruction method [23].

Acknowledgements

The author is indebt to Nicholas Marshall, Fred Sigworth, Ti-Yen Lan, Tamir Bendory, and Joe Kileel for valuable discussions and comments. This work was supported in part by AFOSR Awards FA9550-17-1-0291 and FA9550-20-1-0266, the Simons Foundation Math+X Investigator Award, the Moore Foundation Data-Driven Discovery Investigator Award, NSF BIGDATA Award IIS1837992, NSF Award DMS-2009753, and NIH/NIGMS Award R01GM136780-01.

References

[1] J. Drenth, *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

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

[3] A. Wilson, "Determination of absolute from relative X-ray intensity data," *Nature*, vol. 150, no. 3796, p. 152, 1942.

[4] A. Wilson, "The probability distribution of X-ray intensities," *Acta Crystallographica*, vol. 2, no. 5, pp. 318–321, 1949.

[5] E. M. Stein and R. Shakarchi, *Functional analysis: introduction to further topics in analysis*, vol. 4. Princeton University Press, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[6] L. Brandolini, S. Hofmann, and A. Iosevich, "Sharp rate of average decay of the Fourier transform of a bounded set," *Geometric & Functional Analysis GAFA*, vol. 13, no. 4, pp. 671–680, 2003.

[7] G. Porod, "The X-ray small-angle scattering of close-packed colloidal systems," *Kolloid Zeitschrift*, vol. 124, pp. 83–114, 1951.

[8] G. Porod, "General theory," in *Small Angle X-ray Scattering*, pp. 17–51, Academic Press, 1982.

[9] P. Debye, "Zerstreuung von Röntgenstrahlen," *Annalen der Physik*, vol. 351, no. 6, pp. 809–823, 1915.

[10] J. Fernandez, D. Luque, J. Caston, and J. Carrascosa, "Sharpening high resolution information in single particle electron cryomicroscopy," *Journal of structural biology*, vol. 164, no. 1, pp. 170–175, 2008.

[11] J. Vilas, J. Vargas, M. Martínez, E. Ramírez-Aportela, R. Melero, A. Jimenez-Moreno, E. Garduño, P. Conesa, R. Marabini, D. Maluenda, *et al.*, "Re-examining the spectra of macromolecules. current practice of spectral quasi B-factor flattening," *Journal of structural biology*, vol. 209, no. 3, p. 107447, 2020.

[12] A. J. Jakobi, M. Wilmanns, and C. Sachse, "Model-based local density sharpening of cryo-EM maps," *Elife*, vol. 6, p. e27131, 2017.

[13] S. Kaur, J. Gomez-Blanco, A. A. Khalifa, S. Adinarayanan, R. Sanchez-Garcia, D. Wrapp, J. S. McLellan, K. H. Bui, and J. Vargas, "Local computational methods to improve the interpretability and analysis of cryo-EM maps," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.

[14] A. Singer and F. J. Sigworth, "Computational methods for single-particle electron cryomicroscopy," *Annual Review of Biomedical Data Science*, vol. 3, pp. 163–190, 2020.

[15] S. H. Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," *Journal of structural biology*, vol. 180, no. 3, pp. 519–530, 2012.

[16] S. H. Scheres, "A Bayesian view on cryo-EM structure determination," *Journal of molecular biology*, vol. 415, no. 2, pp. 406–418, 2012.

[17] Z. Kam, "Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations," *Macromolecules*, vol. 10, no. 5, pp. 927–934, 1977.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.14.444177; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[18] Z. Kam, "The reconstruction of structure from electron micrographs of randomly oriented particles," *J. Theor. Biol.*, vol. 82, no. 1, pp. 15–39, 1980.

[19] B. von Ardenne, M. Mechelke, and H. Grubmüller, "Structure determination from single molecule x-ray scattering with three photons per image," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[20] R. P. Kurta, J. J. Donatelli, C. H. Yoon, P. Berntsen, J. Bielecki, B. J. Daurer, H. DeMirci, P. Fromme, M. F. Hantke, F. R. Maia, *et al.*, "Correlations in scattered X-ray laser pulses reveal nanoscale structural features of viruses," *Physical review letters*, vol. 119, no. 15, p. 158102, 2017.

[21] H. Liu, B. K. Poon, D. K. Saldin, J. C. Spence, and P. H. Zwart, "Three-dimensional single-particle imaging using angular correlations from X-ray laser data," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 69, no. 4, pp. 365–373, 2013.

[22] N. Sharon, J. Kileel, Y. Khoo, B. Landa, and A. Singer, "Method of moments for 3D single particle ab initio modeling with non-uniform distribution of viewing angles," *Inverse Problems*, vol. 36, no. 4, p. 044003, 2020.

[23] T. Bendory, N. Boumal, W. Leeb, E. Levin, and A. Singer, "Toward single particle reconstruction without particle picking: Breaking the detection limit," *arXiv preprint arXiv:1810.00226*, 2018.

[24] T. Bendory, N. Boumal, W. Leeb, E. Levin, and A. Singer, "Multi-target detection with application to cryo-electron microscopy," *Inverse Problems*, vol. 35, no. 10, p. 104003, 2019.

[25] A. S. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. S. Wein, "Estimation under group actions: recovering orbits from invariants," *arXiv preprint arXiv:1712.10163*, 2017.

[26] A. Dutt and V. Rokhlin, "Fast Fourier transforms for nonequispaced data," *SIAM Journal on Scientific computing*, vol. 14, no. 6, pp. 1368–1393, 1993.