Neighbour-Counting Statistic for Multitype 3 Dimensional Protein Point Patterns

Chase Osborne

Abstract

We present a statistically principled and computationally efficient method for quantifying local clusters of *target* proteins around *reference* proteins in three–dimensional cryo-EM data. Our framework (i) handles an arbitrary number of protein types, (ii) provides analytic expectations under a complete spatial randomness (CSR) null model or any Monte-Carlo surrogate, (iii) outputs z-scores, confidence intervals, and p-values, and (iv) scales to millions of points via spatial indexing.

1 Introduction

Recent single-particle cryo-EM experiments generate cloud-like sets of 10^4–10^6 protein coordinates per sample. Detecting protein–protein interactions from such data reduces to testing whether specific *target* types occur unusually often in the neighbourhood of *reference* types. We formalise this as a multitype marked point-process problem and build a hypothesis-testing pipeline around efficient neighbour counts.

2 Data model and notation

First working with N types of proteins, which we assume to be point particles whose points are located at the protein's centroid. We will denote the protein type by τ_k where k indexes the protein type. We will also define the 3-Dimensional set of coordinates for a given protein to be $\mathbf{x}$. Given this, the set of all protein coordinates along with their indexed type is:

$$\mathcal{P} = \big\{(\mathbf{x}, \tau_k) \in \mathbb{R}^3 \times \{1, \dots, N\}\big\}$$

From here, our procedure for determining clustering between two arbitrary protein types $i, j \in \{1, ..., N\}$ is as follows:
Take i and j from the set of all possible protein types and define a reference set

$$R_i = \{\mathbf{x} \mid \tau_k = i\}$$

and a target set

$$T_j = \{\mathbf{x} \mid \tau_k = j\}$$

along with a set of radii of interest for binning

$$\mathcal{R} = \{r_1, \dots, r_L\}$$

We can define our neighbor-counting function. Characterizing how many reference proteins fall within a radius of r of a given target protein $\mathbf{y} \in T_j$:

$$C_{\mathbf{y}}^{(i,j)}(r) = \sum_{\mathbf{x} \in R_i} \mathbf{1}\big(\|\mathbf{x} - \mathbf{y}\| \leq r\big). \tag{1}$$

Since we are mainly focused on local clustering between one or two particles, we define a function that counts the target proteins that have exactly m particles within a given radius r

$$H_m^{(i,j)}(r) = \sum_{\mathbf{y}\in T_j} \mathbf{1}\big(C_{\mathbf{y}}^{(i,j)}(r) = m\big) \tag{2}$$

This is the function that is used to determine the experimental data from null models. Under Complete Spatial Randomness with reference particle density ρ_i. The neighbor-count for any given volume $V(r)$ will follow a poisson distribution with parameter λ as:

$$\lambda_{CSR}^{(i,j)}(r) := \lambda_{ij}(r) = \rho_j\, V_3(r), \quad \text{given } V_3(r) = \frac{4\pi}{3} r^3.$$

We can then define (2) in terms of the poisson PDF

$$\mathbb{E}[H_m^{(i,j)}(r)] = |T_j| \frac{(\lambda_{ij}(r))^m e^{-\lambda_{ij}(r)}}{m!}$$

When analytic assumptions fail (aperiodic boundry conditions) we approximate $\mathbb{E}[H_m^{(i,j)}(r)]$ and its variance from the expectation through Monte-Carlo methods. In our model, this numerical approximation is not analytic. And proves to be computationally expensive. Optimization of this problem will be covered in later sections.

Dimensionality Note. An important methodological consideration is the dimensionality of the spatial analysis. Although experimental CryoEM data is often visualized as two-dimensional projections, the underlying protein distributions are inherently three-dimensional. Early analyses using 2D projections of particle positions showed consistent inflation in clustering statistics—particularly in Ripley's K-function—due to artificial compression of distances in the projection plane. These artifacts manifest as uniformly elevated neighbor counts and can lead to false rejection of the Complete Spatial Randomness (CSR) hypothesis. To avoid this geometric bias, all analysis presented in this work is conducted directly in $\mathbb{R}^3$ using the full 3D centroid coordinates of the protein particles.

3 Statistical Method

For each pair (i,j) and radius r we compute the z–score

$$z_{ij}(r;m) = \frac{\widehat{\mu}_{ij}(r) - \mu_{ij}^{\text{null}}(r)}{\sigma_{ij}^{\text{null}}(r)}, \tag{3}$$

Using the formula(s):

$$\mu_{ij}(r) = \mathbb{E}[H_m^{(i,j)}(r)], \quad \sigma_{ij}^{\text{null}}(r) = \sqrt{\frac{\sum_{k=1}^{S} (H_m^{(i,j)}(r) - \mu_{ij}^{null}(r))^2}{S-1}}$$

Here *null* stands for the Monte-Carlo surrogate, and the carat denotes the value derived from experimental data. Assuming approximate normality (justified by the Lyapunov CLT because counts are sums of weakly dependent Bernoulli's in dilute samples) the two-sided p-value is $p = 2\Phi\big(-|z|\big)$. A $100(1-\alpha)\%$ confidence interval for the null mean is

$$\big[\mu^{\text{null}} - z_{\alpha/2}\,\sigma^{\text{null}},\ \mu^{\text{null}} + z_{\alpha/2}\,\sigma^{\text{null}}\big].$$

This allows for us to create a confidence envelope for any m and r of our choosing. These statistics laid out in (3) are directly based on using Monte-Carlo to sample from the underlying distribution. Hence the sample variance being used. In the next section we will cover how to do this sampling.

4 Neighbour-count algorithm

```
Input: reference set $R_i$, target set $T_j$, radii $\mathcal{R}$
Output: $\{\widehat{\mu}_{ij}(r)\}_{r\in\mathcal{R}}$
Build a KD-tree $K$ on $T_j$;
Initialise $C[r] \leftarrow 0$ $(\forall r \in \mathcal{R})$;
foreach $x \in R_i$ do
    foreach $r \in \mathcal{R}$ do
        $C[r]$ += $|\,\textsc{QueryBall}(K, x, r)|$;
    end
end
foreach $r \in \mathcal{R}$ do
    $\widehat{\mu}_{ij}(r) \leftarrow C[r]/|R_i|$;
end
```

Algorithm 1: Mean neighbour count for one (i, j) pair

Complexity. Let $n_i = |R_i|$, $m_j = |T_j|$, and $\overline{k}$ be the average number of neighbours returned.

- KD-tree construction: $\mathcal{O}(m_j \log m_j)$.
- One radius query: $\mathcal{O}(\log m_j + \overline{k})$.
- Total for all radii: $\mathcal{O}\big(n_i L(\log m_j + \overline{k})\big)$.

By reusing the same tree for every reference type, the full multitype run costs

$$\mathcal{O}\Big(\sum_j m_j \log m_j + L \log\Big(\sum_j m_j\Big) \sum_i n_i + L \sum_{i,j} n_i\, \overline{k}_{ij}\Big),$$

which is near-linear in input size when $\overline{k}_{ij}$ stays bounded.

The naïve double loop visible in the first prototype (`for ref in ['C3', 'Cx', 'CP']: ...`) is $\mathcal{O}(n_i m_j L)$ and becomes prohibitive for $N \gtrsim 10^5$.

5 Results

Using the neighbour-counting statistic, we compared experimental protein distributions to a Monte Carlo null model of complete spatial randomness (CSR) for all protein–protein type pairs and a range of search radii. Across all tested combinations, the observed counts fell within the 95% confidence envelope of the null distribution, indicating no statistically significant deviations from CSR at the measured length scales. This result suggests that, within the resolution limits of the data and the sensitivity of our current statistical framework, there is no evidence of strong protein–protein interactions.

The two panels in Figures 1 and 2 summarise the main findings. Figure 1 presents the fully processed results, including mask-based spatial restrictions, isotropic edge correction, and multiple-testing control. All observed neighbour counts remain close to the null expectation, with fluctuations consistent with Monte Carlo sampling noise. Figure 2 shows the same analysis before spatial filtering and statistical corrections, where apparent deviations from CSR can be seen; these disappear after applying the full set of methodological controls, underscoring their importance in avoiding false positives.

While this study demonstrates the capabilties of the current CSR-based pipeline for detecting strong clustering between protein types, several variations of the method presented remain for future work. These include the adoption of log-spaced radius grids to better balance small and large scale sensitivity, additional replicate stability tests to quantify Monte Carlo variance across parameter regimes, and more refined multiple-testing correction strategies tailored to spatial point processes.

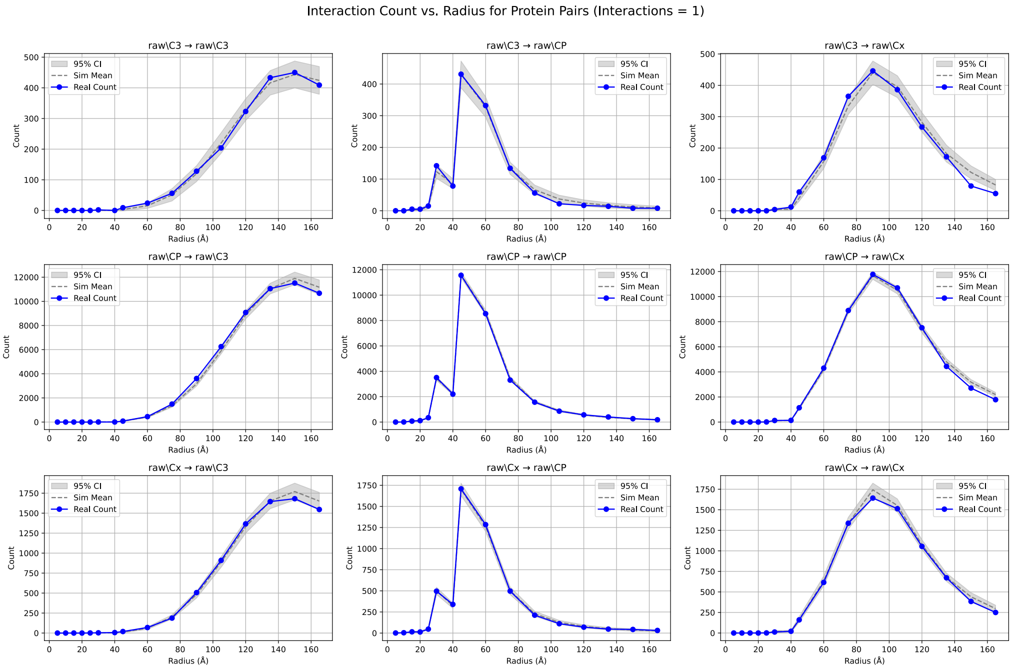

Figure 1: Filtered interaction counts compared to null expectations. Shaded regions represent the 95% confidence envelope from Monte Carlo simulations under CSR. All measured counts remain within the envelope, consistent with random spatial arrangement.

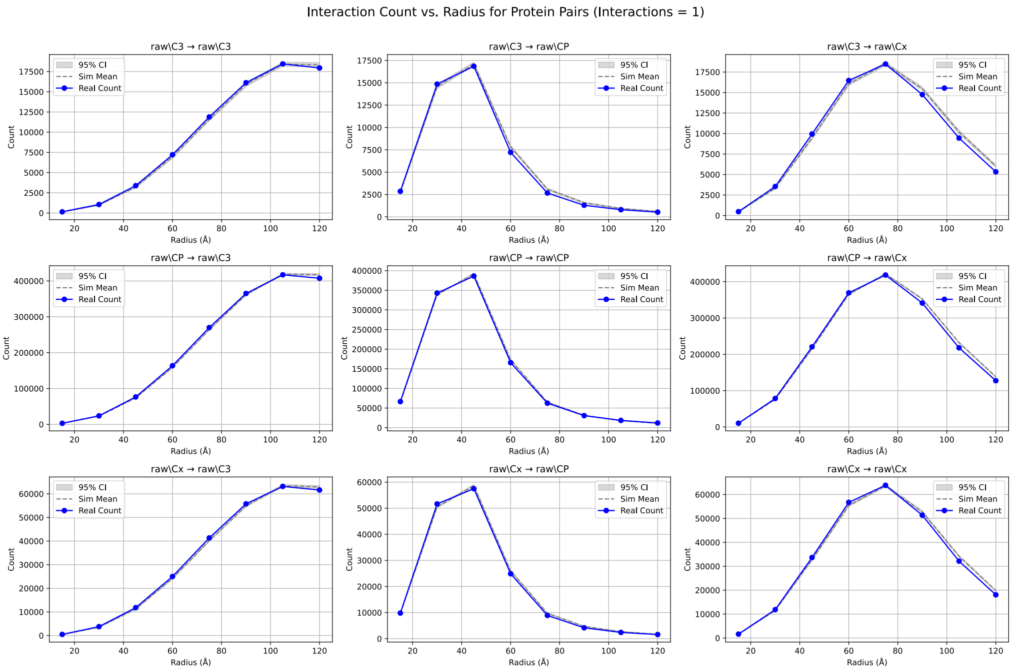

Figure 2: Unfiltered interaction counts prior to edge correction and multiple-testing control. Deviations outside the 95% confidence envelope are visible here but are removed after applying the full correction pipeline.

6 Conclusion

We have developed and applied a statistically principled neighbour-counting framework to three-dimensional CryoEM point-pattern data, enabling direct comparison of experimental protein distributions to CSR-based null models. Within the scope of our current dataset and analysis pipeline, no significant protein–protein clustering was detected, implying that any interactions at this spatial scale are either absent or below the detectable threshold of our method.

These results provide a rigorous statistical baseline for future interaction studies. The modular structure of the framework readily accommodates methodological refinements—such as log-spaced radius sampling, higher-order interaction tests, and alternative null models—that may enhance sensitivity to subtler effects. As a result, the approach presented here not only establishes current limits on detectable clustering but also sets the stage for improved spatial statistics in high-resolution structural biology.