Spatial coherence of DNA barcode networks

David Fernandez Bonet [1], Johanna I. Blumenthal [1], Shuai Lang [1], Simon K Dahlberg [1], and Ian T. Hoffecker [1]

[1]Science for Life Laboratory; Department of Gene Technology, KTH Royal Institute of Technology; Tomtebodavägen 23a 171 65, Solna, Sweden

Sequencing-based microscopy is an emerging family of techniques for locating the positions of molecules in a tissue or biological sample without traditional optics. Instead, spatial information is gathered through a network of DNA sequences that tag individual molecules. Such information is retrieved using high throughput sequencing technology and reconstructed computationally into an image. Multiple proposed chemistries and proof of concept experiments have established the feasibility of this approach on simple model systems. However, methods are needed to assess the validity of reconstructed images in the absence of ground truth knowledge or parallel use of optical techniques. To address this, we identified a set of ground truth-agnostic properties, spatial coherence measurements, that may be computed in simulated or experimental sequencing-based microscopy data. Spatial coherence represents a network's potential to preserve spatial relationships through its topology. The measures are based on a generalization of Euclidean geometry to spatial networks, and they indicate when such networks deviate from Euclidean laws. These deviations, e.g. due to the presence of non-spatially correlated connections, create contradictory constraints that lead to distortions in image reconstruction. We propose that spatial coherence may be employed as a generic metric of overall quality of spatial information in sequencing-based microscopy even in the absence of secondary validation, as it is based on a fundamental geometric criteria.

DNA sequencing-based microscopy | graph theory | imaging-by-sequencing | DNA microscopy | spatial biology | spatial transcriptomics | DNA computing | molecular programming

Correspondence: ***ithof@kth.se, dfb@kth.se***

Introduction

Methods for mapping spatial information in tissues are expanding our understanding of molecular physiology, development, and pathology Lewis et al. (2021); Moses and Pachter (2022). These include optical molecular imaging methods which offer finely resolved molecular images for up to tens of simultaneous targets, whereas high throughput single-cell methods (Shapiro et al., 2013; Gawad et al., 2016) provide data for a large number of genes but lack spatial information about where each cell is located in the tissue. Spatial omics techniques (Ke et al., 2013; Lee et al., 2015; Ståhl et al., 2016; Wang et al., 2018; Karaiskos et al., 2017; Satija et al., 2015; Achim et al., 2015; Halpern et al., 2017; Rodriques et al., 2019) seek to eliminate this trade-off with high multiplicity coupled with spatial resolution, but are also constrained by instrumentation-demanding bottleneck steps for generating spatially resolved reference maps (e.g. by microscopy-based *in situ* sequencing, array printing, or cyclic exchange of fluorescent markers) used to look up the positions of molecules of interest.

Sequencing-based microscopy(Zador et al., 2012; Glaser et al., 2015; Schaus et al., 2017; Boulgakov et al., 2018; Weinstein et al., 2019; Hoffecker et al., 2019; Gopalkrishnan et al., 2020; Boulgakov et al., 2020; Greenstreet et al., 2022; Karlsson et al., 2023), in contrast, is an emerging family of techniques that seeks to spatially resolve many molecular targets while also eliminating the requirement of building a spatial reference map. Instead of optics, the main idea is to rely only on DNA as a tool to capture spatial information and later sequence it. Using this "inside-out" approach instead of array-based methods offers potential advantages, including high multiplexing of molecular targets, scalability and automation, reduced instrumentation, and isotropic imaging in 3D or optically opaque samples.

Sequencing-based microscopy methods take advantage of DNA's capacity to capture spatial information. Initially, molecules of interest –such as messenger RNA or proteins from a biological sample– are tagged with unique DNA identifiers (Fig.1a). These identifiers are then clonally amplified to form polonies (Fig.1b), or local patches of uniquely identifiable DNA. Each polony interacts with neighboring polonies via covalent linkage (Fig.1c, d), effectively constructing a proximity graph Mathieson and Moscato (2019) (which, for the sake of clarity, we hereafter refer to as a spatial network). This spatial network encapsulates both individual molecular identities and their spatial information (Fig.1e). In this network, nodes represent the initial molecules of interest, while edges denote their proximity. After isolating and sequencing the DNA from the sample, this information can be extracted and used to reconstruct the original positions.

Different sequencing-based microscopy modalities have been modeled *in silico* (Hoffecker et al., 2019; Bonet and Hoffecker, 2023; Greenstreet et al., 2022; Glaser et al., 2015; Boulgakov et al., 2020) and implemented *in vitro* (Weinstein et al., 2019; Karlsson et al., 2023) and *ex vivo* (Qian and Weinstein, 2023). Although each modality comes with a model of how networks form, e.g. saturated polony forests that form tessellations or beacon-receiver communication via diffusion, these models are ultimately idealized representations of

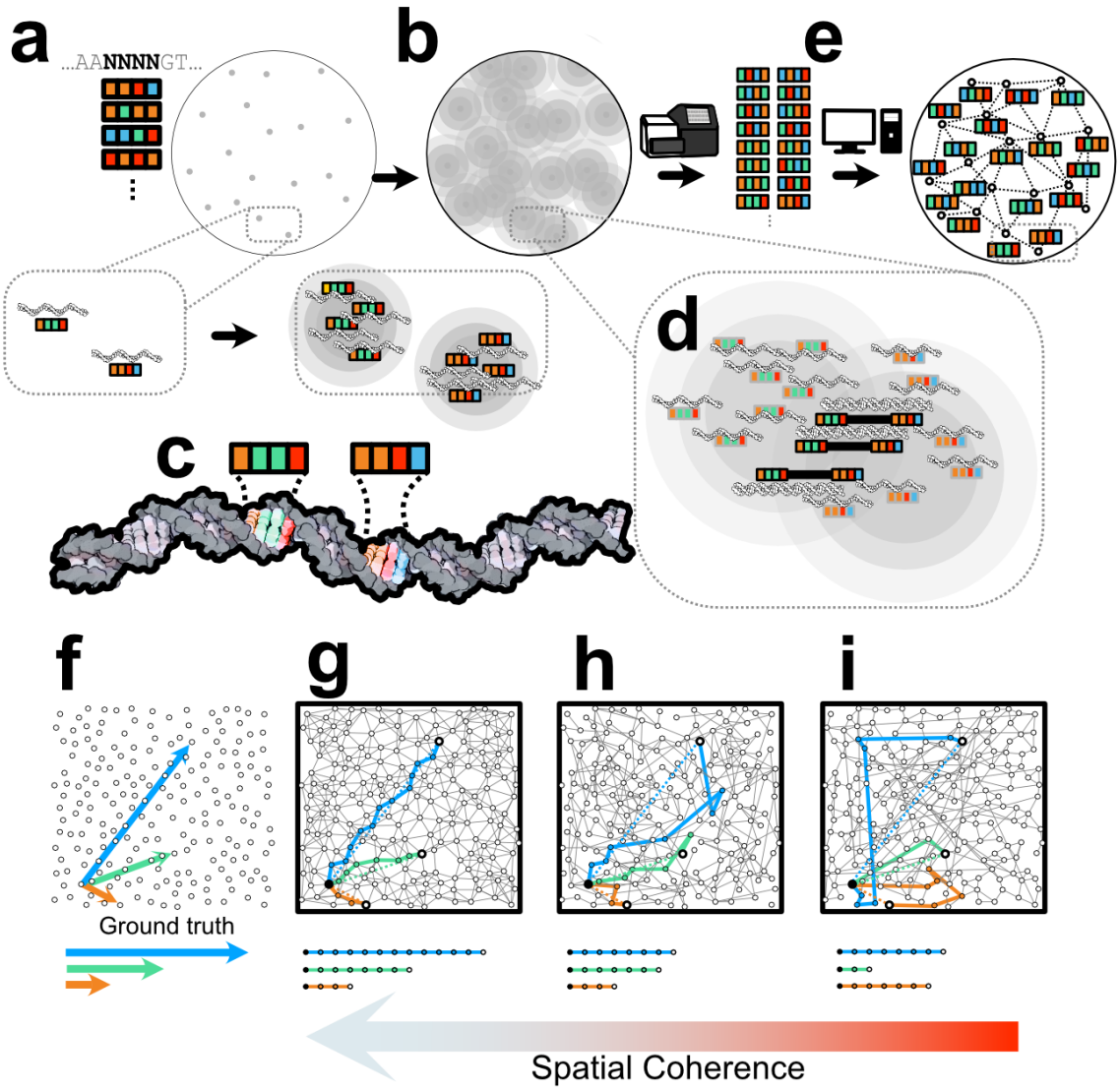

Figure 1. Sequencing-based microscopy. **a.** DNA strands are distributed across space, each with a unique DNA barcode and with a sequence design that enables interaction both with molecules of interest and with other DNA strands. **b** After seeding the space with DNA strands, such seeds are clonally amplified forming polonies. Each polony effectively represents a region covering the space. **c** Polonies later interact through covalent linkage, recording proximity in physical space. Products containing two barcodes denote polony interactions. **d** Such pairwise interactions are powered by confined spread (amplicon diffusion, or other locally confined amplification) and physical proximity and are represented by network edges that provide spatial information. **e** This information is later sequenced and filtered, providing an edge list that results in a spatial network denoting proximity associations. **f** The ground truth comprises the original disposition of DNA strands and their distances and is in general not available in experimental settings. **g** An idealized spatial network where network edges connect nodes that are originally close. Because ground truth distances and network distances are strongly correlated, the network is spatially coherent. **h** Realistic network where edges can connect distant regions. These shortcuts shorten the network distance required to reach one node from another, resulting in a discrepancy compared to the ground truth distances. As a result, the spatial coherence is reduced. **i** Pathological network with an abundance of non-spatially correlated edges.

the network formation chemistry. Real data can contain spurious interactions between spatially distant regions, potentially affecting the reconstructed image. Therefore, the quality of the reconstruction depends on the quality of the interactions defining the network, as reconstruction algorithms rely on the assumption that interactions are representative of spatial proximity.

Hence, a major challenge facing this field is addressing these network errors, and doing so without knowing the original spatial positions. These positions, or ground truth data (Fig.1f), are available in *in silico* implementations, where it is possible to perform quantitative comparisons between the original positions and the reconstructed positions. In our previous work Bonet and Hoffecker (2023), we developed quality metrics to assess both local and global reconstruction quality by direct comparison with ground truth. However, laboratory experimentation lacks ground truth access, necessitating secondary validation strategies. Examples are confirming the co-localization of molecules of interest after cell stimulation Karlsson et al. (2023), or verifying that the global morphology agrees with expected biological sample morphology Qian and Weinstein (2023); Weinstein et al. (2019). These validation strategies rely on prior knowledge of the sample, and only reveal the extent to which distortions influence the reconstruction quality so far as deviations from the ground truth can be identified. Distortions might arise from false communi-

cation between unrelated parts of the network, incomplete network data, or errors during amplification and sequencing.

Validation tools that work in the absence of ground truth or prior knowledge would facilitate the use of sequencing-based microscopy techniques that rely less on additional verification methods and aid in the development and improvement of different chemistries/modalities. We investigated the underlying mathematical relationship between network topology and spatial location. We identified measurable properties that we show are necessary features of spatially-informative networks and which are needed to reconstruct an image from them. The fulfillment of these properties, which we term "spatial coherence", is the consistency with which the network follows Euclidean geometric laws. We show that spatial coherence is a feature of well-behaved spatial networks, and that it is reduced by introducing random, non-spatially-correlated edges between nodes in the network (Fig.1g, h, i) and that it is impacted by sparse or incomplete sampling of the network. Spatial coherence is a measurable, ground-truth agnostic property that can be used to assess how well spatial information is captured in sequencing-based microscopy networks, and could aid in benchmark comparison across different techniques, or provide a metric of confidence in reconstructed images.

Results

Based on a previous observation that Euclidean and shortest path topological distances of spatial networks are correlated, Hoffecker et al. (2019); Bonet and Hoffecker (2023) we sought to extend this concept to more broadly interrogate the geometry of spatial networks, and assess whether a particular network obeys Euclidean geometric scaling laws, i.e. whether it exhibits spatial coherence. Real-world networks such as the world-wide-web often exhibit a small-world property in which the average shortest path between any two nodes is short compared to the size of the network Watts and Strogatz (1998); Barrat and Weigt (2000). However, spatial networks are fundamentally different, and exhibit a "large-world" property: distant nodes in physical space are also distant in the network Zemel and Carreira-Perpiñán (2004). For example, in the European high-speed rail network, it takes more connections to reach Rome from Stockholm than from Copenhagen.

A. Metrics for examining the geometry of spatial networks

We sought quantitative metrics for evaluating spatial coherence that would provide a way to validate the consistency with which space is represented by DNA barcode networks without relying on a ground truth reference for comparison. This unsupervised or intrinsic criterion restricts our analysis to network structure without prior biological knowledge of the sample, and should be informative on an individual network-basis without additional (e.g. calibration) experiments needed. We propose three such unsupervised measures, though note that these represent only a selection of possible Euclidean geometric relationships that could be exploited in graph space to quantify spatial coherence. We developed measures for: quantifying the intrinsic network dimension, measuring if a quantity that we refer to as spatial constant remains stable across scales, and quantifying the proportion of variance explained within the dimensions of the original space.

The first approach we explored is to quantify the intrinsic dimension of a network via an analogy to Euclidean geometric scaling laws, and to compare that dimension with that which we expect from Euclidean spatial dimensions. Spatial networks exhibit a correlation between their shortest path distance and their ground truth Euclidean distance, enabling substitution of spatial variables in familiar Euclidean geometric scaling relationships. The Euclidean geometric relationship between the number of points inside a ball (a sphere or a circle, see Fig.2a, b) and its radius r obeys the power law:

$$\frac{N}{C} = \alpha r^{\mathcal{D}} \tag{1}$$

where N is the number of points, C is the point density per unit of space, $\mathcal{D}$ is the dimension, and α is a geometric factor that depends on the dimension. In particular, this factor is $\alpha = \pi$ for a circle and $\alpha = \frac{4}{3}\pi$ for a sphere. Conversely to the Euclidean scenario, the dimension of a network can be quantified using an analogous power law Shanker (2007) (Fig.2c), that is applicable to DNA barcode networks, where N corresponds with the number of moles and C with solution concentration. The network resulting from such an experiment behaves according to Euclidean laws because network molecules connect or do not connect according to their spatial proximity to one another. The intrinsic network dimension can be found by fitting:

$$\frac{N}{\langle k \rangle} = \alpha r'^{\mathcal{D}'} \tag{2}$$

where the Euclidean distance r in Eq.1 has been substituted with the network's shortest path distance r' from a given origin node (i.e. the depth of a breadth-first-search), N is the number of nodes in the subgraph contained within r' of a given origin node, $\langle k \rangle$ is the average node degree, α is a constant of proportionality, and $\mathcal{D}'$ is intrinsic dimension of the network. (Fig.2d).

In a large spatially coherent network, the network dimension should agree with the Euclidean dimension of the space. However, boundary effects influence the measurement, making it dependent on the origin point and the depth of the scaling relationship interrogated at that point. If the measurement begins near a boundary, the perceived dimensionality will be lower compared to a central start.

Another quantity we sought is the consistency of dimensionless constants in geometric scaling laws, such as

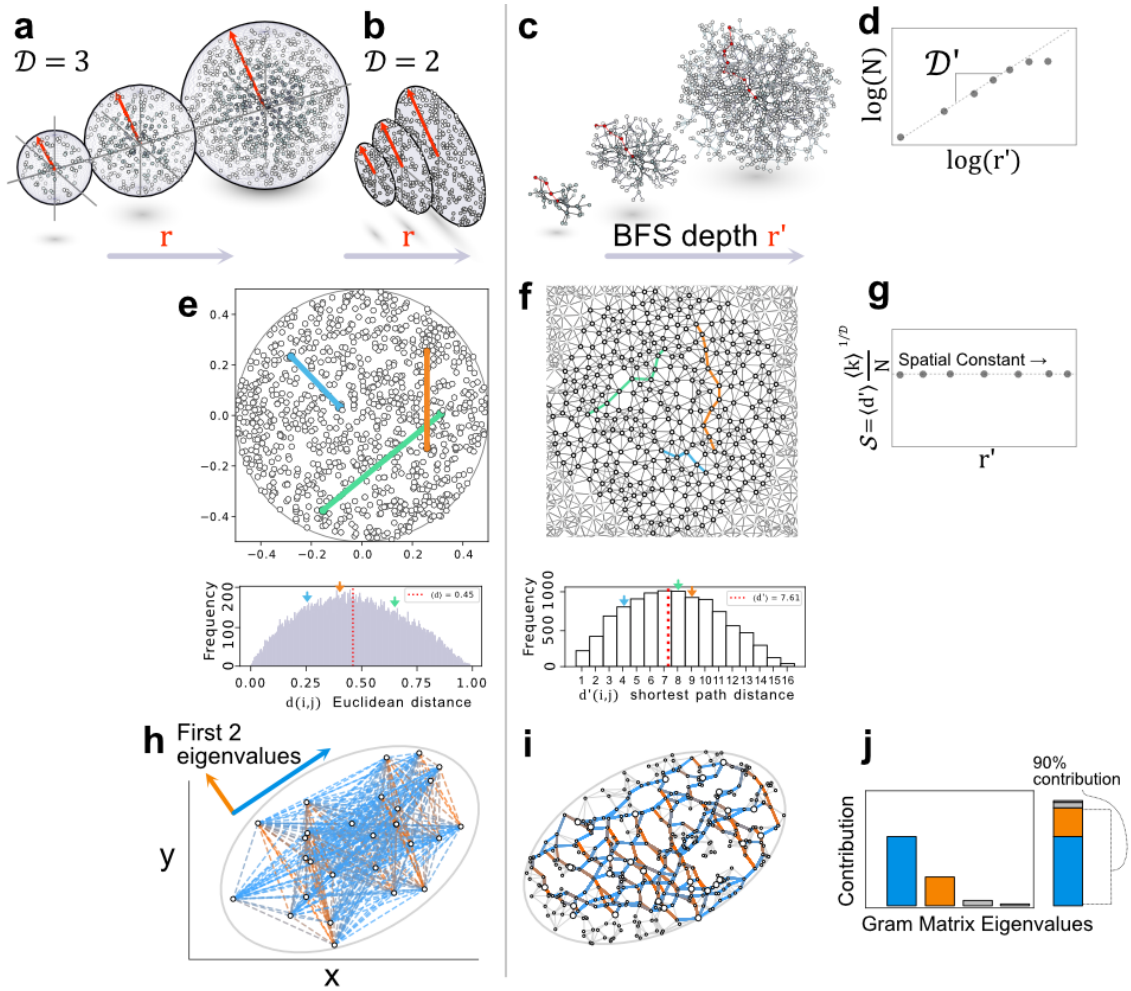

Figure 2. Measures to quantify the spatial coherence of networks. **a.** A Euclidean scaling law relates the volume of a sphere to the cube of its radius, or **b.** the square of the radius in the case of a circle. A volume with a uniform concentration of molecules or points may be represented by the number of points. **c.** A network-based analog of the Euclidean spherical volume scaling laws relates the number of nodes discovered within a certain depth r' of a breadth-first search (BFS). **d.** For spatially coherent networks, the dimension parameter of the scaling law will approximate Euclidean dimensions. **e.** The distance d between any two points within a space in Euclidean geometry follows another scaling law with an expected mean that depends on the radius or scale of the space. **f.** A spatial network exhibits an analogous scaling law for the shortest path distance d' between any two nodes within a BFS range, with a mean that scales with depth. **g.** The dimensionless constant of proportionality for the mean distance should remain constant across scales in a spatially coherent network or Euclidean space. **h.** The Gram matrix for a set of Euclidean points may be calculated from the pairwise distances of those points, the eigenvalues of which reveal the contributions to variation in the spread of points from different dimensions (2 for spatially distributed points in 2D). **i.** A Gram matrix may be computed from a set of shortest path distances substituting for Euclidean distances in the case of a spatial network. **j.** The first eigenvalues, representing the number of dimensions of the space, should account for the variation in distances in a spatially coherent network. Contributions from higher dimensions indicate variation that cannot be explained by spatial variation.

the scaling law governing the relationship between the mean distance between points (as approximated by the shortest path distance between nodes) and the size of the space. In Euclidean space (Fig.2e), the expected line segment length d for a space of dimension D is related to the size of a space by a dimensionless constant

$$\beta = \langle d \rangle \left(\frac{C}{N} \right)^{\frac{1}{\mathcal{D}}} \tag{3}$$

where $\frac{C}{N}$ is the size of the space, equivalent to volume in 3D or area in 2D. This relative quantity is constant across scales in Euclidean space for a given dimensionality and boundary geometry of the space (e.g. a cylinder vs a sphere). The network-based analog (Fig.2f) of this dimensionless spatial constant is

$$\mathcal{S} = \langle d' \rangle \left(\frac{\langle k \rangle}{N} \right)^{\frac{1}{\mathcal{D}}} \tag{4}$$

where N is the number of nodes, $\langle d \rangle$ has been substituted with mean the shortest path distance $\langle d' \rangle$, and the point density (molar concentration) has been substituted with the average node degree $\langle k \rangle$. This mean distance spatial constant may be computed either within a limited subset of a large graph, or by considering the average shortest path distance between all pairs of nodes, as an ensemble characteristic of the whole graph (Eq.4). A Euclidean spatial constant does not change across scales, and the same principle should apply to networks (Fig.2g) that have the same behavior across different network scales, but would change if topological representation of space was not consistent across scales in a network. Hence, the extent to which

this dimensionless quantity remains constant could be a good indicator of spatial coherence.

Finally, we explored a modified principal component analysis (PCA)-based approach adapted for spatial network data. Namely, we analyzed the proportion of Gram matrix eigenvalues as a tool to determine the contribution of variation in the data along each expected spatial dimension. This analysis differs from classical PCA performed on point coordinates that would require ground truth knowledge. We chose to work with the Gram matrix instead of the covariance matrix used in PCA, which is computed from distances that we substitute with shortest path network distances. The resulting eigenvalues are equivalent to principal components under the assumption that the distances are Euclidean Hastie et al. (2009). In Euclidean space (Fig.2h), the first $\mathcal{D}$-components account for all the variance, while remaining dimensions do not contribute. A network in which its shortest path distances data can be mostly or completely explained using only $\mathcal{D}$-dimensions (Fig.2i, j) would have high spatial coherence, consistent with the constraints of a Euclidean space.

B. Spatially coherent proximity graphs preserve Euclidean geometric relations

We evaluated how the Euclidean geometric relations described in the previous section behave in both Euclidean geometric spaces and simulated spatial networks. We may generate a point cloud that gives rise to either of such spaces, but while Euclidean space distances are computed with the Pythagorean theorem, a network formed from that point cloud and some proximity rule for assigning edges between points will be characterized by distances that are computed from the shortest path along network edges between two nodes. The network is generated by creating edges if a pair of points is close, e.g., according to some function of separation distance. Fig.3a shows that Euclidean and shortest path distances are correlated, confirming the intuition that a large Euclidean distance will also result in a large shortest path distance. This can be further corroborated quantitatively by computing the coefficient of determination between the two distances. While a strong correlation demonstrates that the network is spatially coherent, we relied on ground truth distances (Euclidean distances) to find that correlation.

In order to avoid using ground truth distances, which are unavailable in experimental scenarios, we assessed the networks' spatial coherence. Instead of comparing ground truth and network distances, we assessed the extent to which networks consistently follow Euclidean relationships. Fig.3b shows the spectral analysis of the Gram matrix, comparing the eigenvalues derived from both Euclidean and network distance matrices. In the Euclidean scenario, the first two eigenvalues account for all variation in the input data, while the rest of the eigenvalues have a null contribution. This reflects how the underlying space is also 2-dimensional. We observed a similar trend for the network's shortest path distances, but with a slight deviation. Here, the first two eigenvalues continued to capture the bulk of the variance, while the remaining eigenvalues exhibit a small but non-zero contribution. This difference highlights the information lost transitioning from a Euclidean space to a spatial network with discrete distances.

We measured the spatial constant (Eq. 4) of the mean segment length to volume relationship for both Euclidean and network cases across different scales (Fig.3c). This quantity remains constant in Euclidean space, as well as in network space as the spatial network is spatially coherent. This means that the growth of the average distance between points follows the same quadratic growth expected of a Euclidean point cloud (in two dimensions). Therefore, the essential spatial relationships are maintained in networks where edges represent proximity, preserving the Euclidean geometric scaling relations.

Spatial networks also exhibited close agreement in measured dimensionality compared to their Euclidean counterpart. Fig.3d illustrates the dimension prediction for a point cloud shaped as a circle along with its corresponding dimension fit, both for Euclidean and network cases originating from the same point cloud. We measured dimensionality from different origin locations in the network (shown via heat map coloring of each node) and observed depressed values close to the boundaries of the space, indicating that central points are more useful predictors of the inherent dimensionality of the space.

The synthetic spatial networks (Methods section G.1) used in Fig.3A show that these shortest path distance-based spatial coherence measures approach Euclidean results in controlled networks where points are distributed uniformly and density variations are absent. Real spatial networks, such as the US county border network, display non-uniform degree distribution, density variations, and non-isotropic boundary shape (Fig.4a). Fig.4b shows how a reconstructed approximation without crossing edges, preserving the planarity of the original image, may be produced using the STRND algorithm Bonet and Hoffecker (2023). The reconstruction is obtained without using the original point coordinates, and the absence of crossing edges suggests that the network is spatially coherent. This is corroborated by Fig.4c, which tracks the spatial constant at different scales (the size of subgraphs analyzed according to their breadth-first-search or BFS depth). The value of the spatial constant does not change significantly over different scales, suggesting spatial coherence similar to the well-behaved simulated spatial networks. However, the spatial constant declines when introducing random, non-spatially correlated "false" edges that do not reflect the true geographical proximity, connecting far-apart regions. This contrast in the characteristic behavior of the spatial constant in spatially coherent versus networks with false edges could provide an indication of distor-

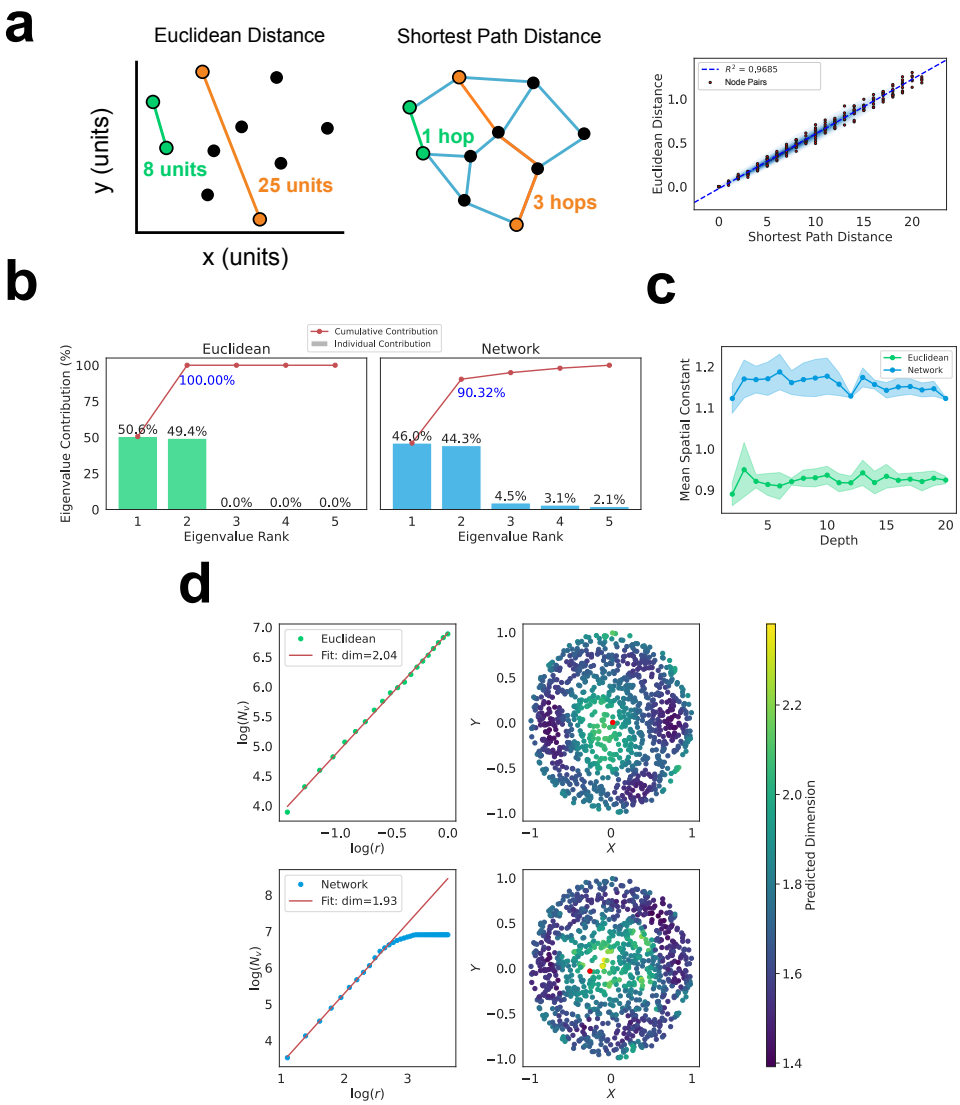

Figure 3. Spatial coherence in Euclidean and network data. **a** Euclidean data depicting ground truth positions and selected distances, along with its network counterpart. Network shortest path distances and Euclidean distances are highly correlated. **b** First Gram matrix eigenvalues and their contribution at explaining data variability. Pairwise ground truth distances are used to obtain the Gram matrix for the Euclidean case (left) shortest path distances for the network case (right). **c** Spatial constant variation with depth. For the Euclidean case, depth corresponds with the radius of included points r, while for the network case depth corresponds with the BFS depths r'. **d** Dimension prediction obtained by selecting a central origin node (in red) and computing the number of nodes at a certain depth, both for the Euclidean (top) and network (bottom) cases. The number of nodes grows as a power law, and the fit to such power law yields a dimension estimate.

tion in scenarios where the ground truth is absent.

Fig. 4d shows the spectral analysis of the Gram Matrix, where the first two eigenvalues comprise the greatest contribution. This is the expected result for a 2-dimensional spatial network. We observed that the first eigenvalue is comparably higher than the second, reflecting that more variance is captured by the network's longer x-axis dimension and the non-isotropic rectangular shape of the space.

The intrinsic dimensionality of the network was also close to two (Fig.4e), indicating that a Euclidean-like 2D scaling law could explain the relationship between the number of discovered nodes as a function of radial BFS depth. To predict the dimension, we identified 10 representative origin nodes of high network centrality (Methods Eq.10). From each origin node, we conducted a breadth-first-search and inferred the dimension by fitting the power law that relating the number of nodes discovered at depth r' (Eq. 2).

C. False edges result in spatially incoherent networks

The analysis of the US county network and simulated spatial networks shows how spatial coherence measures take on values in spatial networks approaching those of Euclidean spaces. However, these networks are well-behaved in the sense that all edges are spatially informative, i.e. their existence is correlated with spatial distance, and there are no false edges, shortcuts that connect distant regions of the space. Shortcuts could appear in DNA barcode networks, for example, via abnormal polony growth or spatially inconsistent diffusion, polymerase chain reaction (PCR) errors introducing non-spatially correlated connections, or

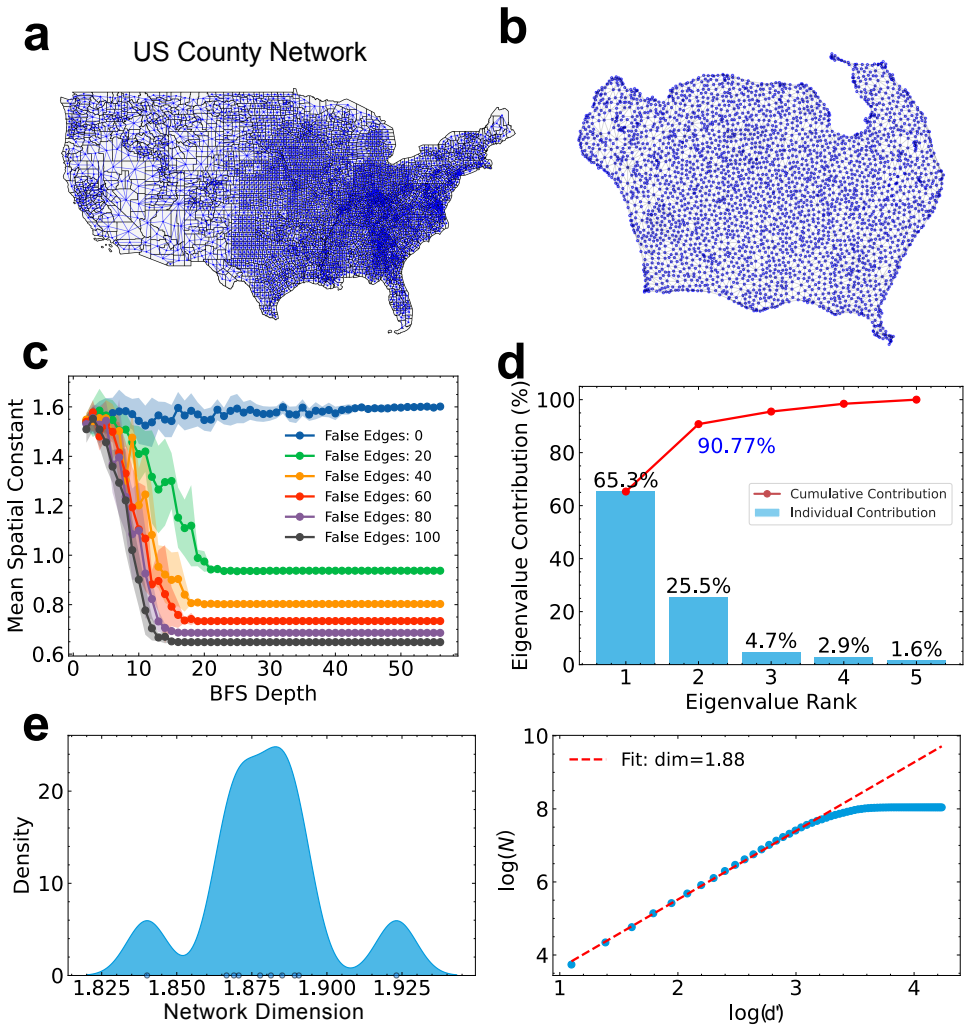

Figure 4. The US County network as a case study. a. US County network formation by county proximity, where an edge is drawn if a pair of counties is adjacent. **b.** Reconstruction of the original county positions using only network data. **c.** Spatial constant behavior against BFS depth, which can be altered by the introduction of false edges. **d.** First Gram matrix eigenvalues contribution in variance explanation. **e.** Spread of network dimension results obtained from several central origin nodes, along with an example of the fits generating such results.

barcode collisions resulting in indistinguishable identities. Such deviations from distance-dependent edges introduce contradictory constraints during reconstruction of the networks that would manifest as distortions, which we sought to detect by measuring spatial coherence. This quantitative approach could would provide an indication of both the quality of the network data and the performance of filtering correction or data preprocessing steps.

Non-spatially correlated or false edges connect distant nodes, reducing spatial coherence of the network. Fig.5a shows the shortest path distance from every node to the central node in a square lattice network. This network is spatially coherent, i.e. shortest path distances correlate with Euclidean distances. If a node is distant in Euclidean space, it is also distant in the network. In particular, distances from the center node increase radially towards the boundaries. However, introducing false edges alters this property by reducing distances. This disruption becomes more significant with each additional false edge, and there is no longer a radial progression in distance from the center. This effect is quantitatively shown in Fig.5b, where there is a strong correlation between Euclidean and network distances if false edges are absent. False edges reduces the correlation, indicating spatial incoherence (Fig.5c). Without incoherence, networks can be reconstructed through a satisfaction of spatial constraints in which there is a strong consensus among points and their relative distances to other points (Fig.5d). However, if false edges are present, reconstruction performance can be affected (Fig.5e) by the conflicting distance constraints that require higher dimensions to resolve, e.g. the folding of a sheet in 3D. While some reconstruction algorithms are more robust than others, false edges are detrimental to the reconstructed image.

An open challenge in sequencing-based microscopy is to reliably detect and quantify the presence of false edges. The task is difficult because the unavailability of ground truth prevents us from directly determining the spatial correlation of edges. The spatial coherence framework can aid in this regard, as it only requires network data. To test the framework, we simulate spatially coherent networks and incorporate false

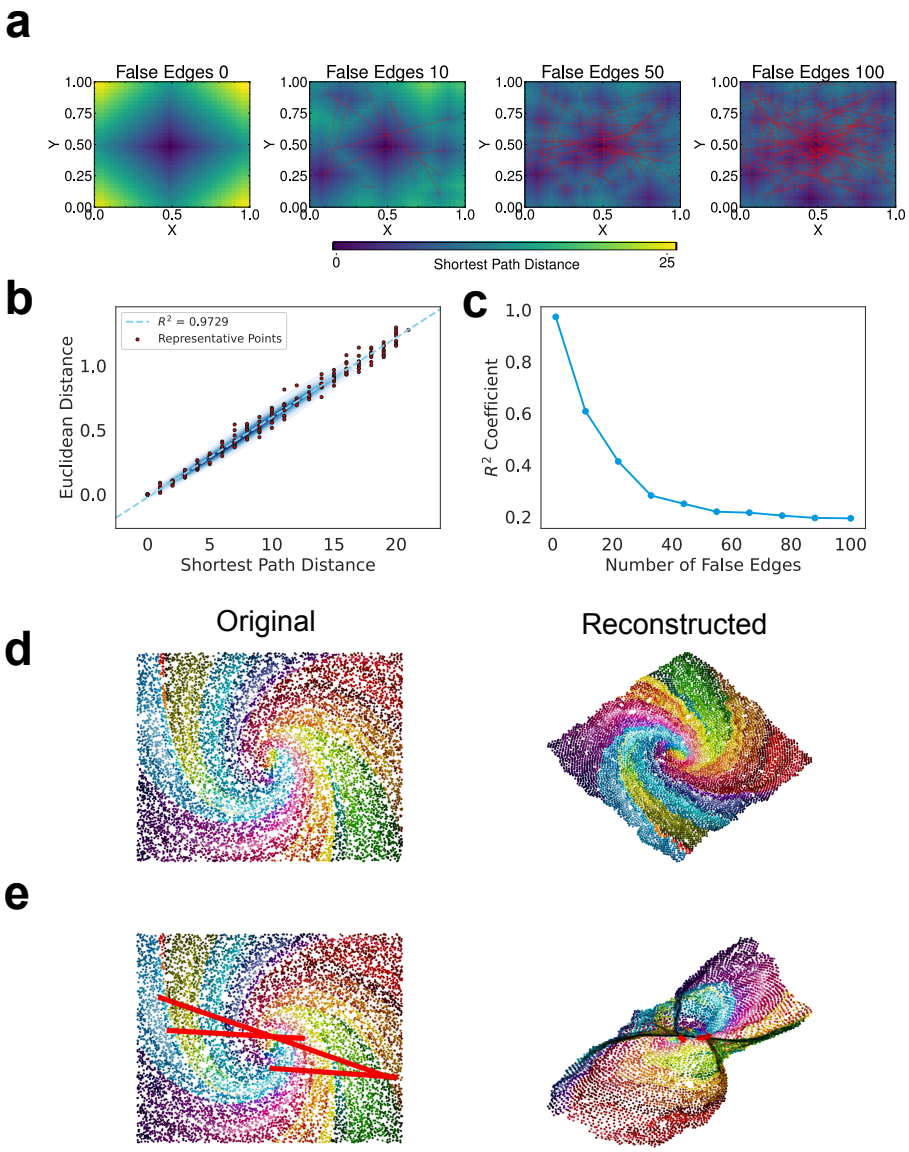

Figure 5. Impact of false edges in networks. a Network shortest path distances from the central node at varying number of false edges. **b**. Correlation between Euclidean and shortest path distance for a spatially coherent network. **c**. Behavior of the coefficient of determination R^2 against the number of false edges introduced. **d**. Ground truth original positions colored according to a spiral pattern and the reconstructed positions using network information only. **e**. Same data with the artificial introduction of false edges (highlighted in red) connecting distant regions and its corresponding reconstruction.

edges into them. Fig.6a infers the dimensionality of the network by assuming that network distances behave as Euclidean distances. When false edges are absent, the predicted dimensions are close to 2, but as false edges are added, the network dimension increases. This increase is expected, as adding false edges to a circular 1-dimensional ring makes it close to a 2-dimensional circle Shanker (2007), and adding false edges to a 2-dimensional paper sheet is reminiscent of folded 3-dimensional structures. Fig.6b shows that it is possible to detect the presence of false edges by computing the spatial constant of the network at different size scales. A drop in the spatial constant indicates spatial incoherence, with more false edges corresponding to a steeper drop. The spectral analysis of the Gram Matrix also indicates spatial coherence (Fig.6c), where the first two components (eigenvalues) account for the highest contribution in a network with no false edges. When false edges are introduced, the contribution decreases accordingly. A similar behavior can be observed for the spectral gap score, or the relative contribution between the first d components and the $d+1$ component (Methods section E.1).

D. Filtering in experimental data affects spatial coherence

After observing that false edges affect spatial coherence measures by increasing the intrinsic network dimension, redistributing Gram matrix eigenvalues, or introducing variability in spatial constant across scales, we next investigated the effect of missing edges. Whereas false edges connect distant spatial regions, missing edges represent the lack of connection between points that are close. In sequencing-based microscopy,

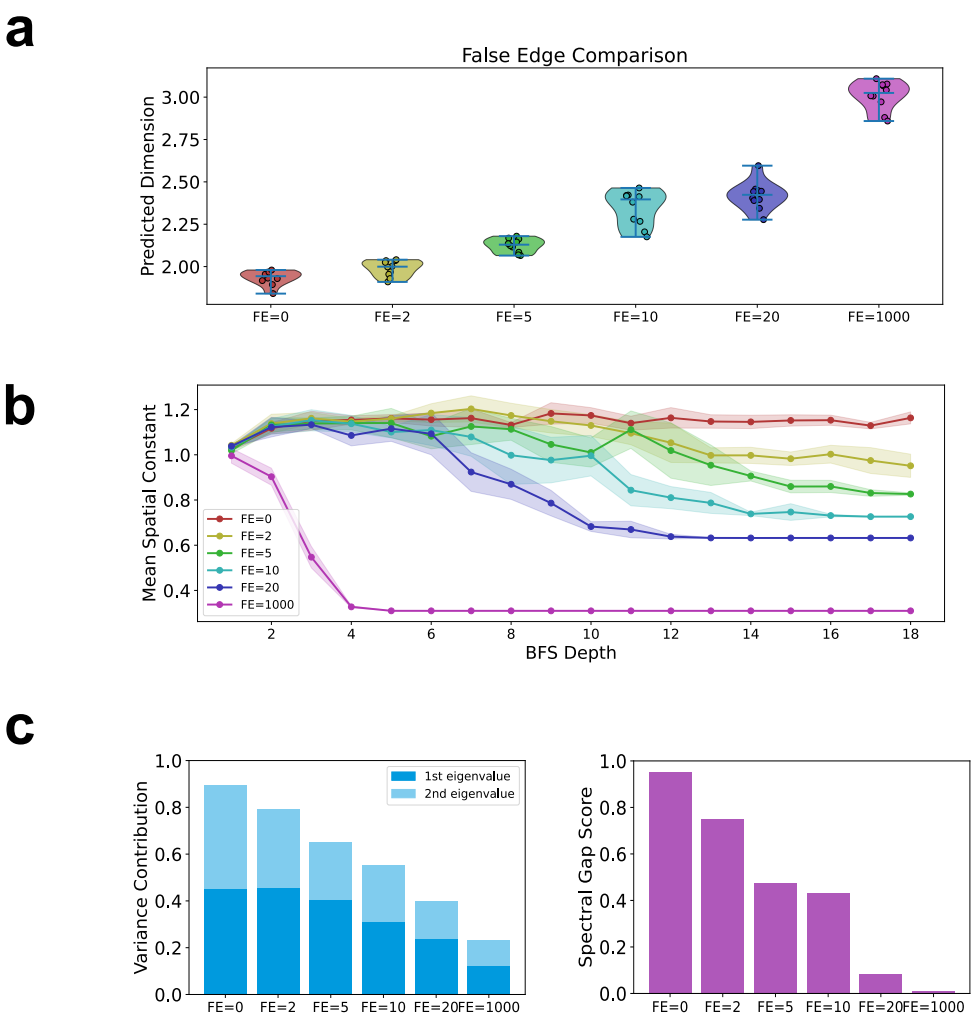

Figure 6. Spatial coherence measures in networks containing an increasing number of false edges, ranging from 0 to 1000. a Network dimension estimates originating from different central nodes, **b** spatial constant values against BFS depth, and **c** Gram matrix first two eigenvalue contributions along with its spectral gap score.

sparseness could arise when nearby polonies do not interact, or when the sequencing depth is too shallow to detect the interaction. The resulting network would therefore be missing edges.

Missing edges do not introduce erroneous data and conflicting spatial constraints like false edges. However, a lack of data introduces a lack of spatial constraints if the number of missing edges is high enough, which could be similarly problematic for reconstruction. In that case, nodes that are close in Euclidean space might lack a direct edge connection, making their shortest path distance greater than it should be and potentially affecting reconstruction quality.

False edges and missing edges can co-exist in experimental settings, so we studied their combined effect on spatial coherence measures. While Fig.7a shows examples of simulated bipartite spatial networks containing false edges and missing edges, Fig.7b shows how they affect spatial coherence measures. False edges generally increased the network dimension and decreased the Gram matrix eigenvalue contribution. In contrast, in the absence of false edges, a high number of missing edges caused a decrease in the network dimension. Specifically, for a sparse case with 60% of the edges missing, the network is closer to a measured dimensionality of 1 than 2. But when networks are both sparse and contain false edges, we notice a compensatory effect where a balance of false edges and missing edges could conceivably make the network appear to be two-dimensional. A similar effect happens for the contribution of the Gram matrix eigenvalues, where networks with false edges have higher scores if they are also sparse. Therefore, for spatial coherence measures to be used for quality assessment, knowledge as to whether the network lacks a portion of its edge content would be necessary to avoid misinterpretation by conflating quality and compensation.

Spatial coherence measures could be used to quantitatively assess the effectiveness of filtering algorithms meant to correct network distortions by removing false edges or fused nodes. The binary classification task where edges and nodes are classified as false or true involves balancing sensitivity and specificity: higher sensitivity for detecting false edges could lead to undesired removal of true edges and data loss. Identifying an optimal filtering power that maximizes false edge removal

while minimizing true edge removal, requires an evaluation metric. Spatial coherence measures could facilitate finding this optimal parameter under conditions where ground truth knowledge of spatial positions is absent, as they are a proxy to quantitatively measure false edges.
A previous study proposed an error correction algorithm to identify false edges and fused nodes Kloosterman et al. (2024), particularly in diffusing DNA amplicons Weinstein et al. (2019). We reproduced the two-dimensional simulations based on this model and applied the corresponding filtering algorithm at different levels of filtering power (sensitivity). We then measured the spatial coherence of the resulting networks (Fig.7c). The results show that increasing filtering power improves spatial coherence: the intrinsic dimension gets closer to two, and the Gram matrix eigenvalue contribution is close to one. Furthermore, the data show that higher false edge ratios require greater filtering power. The value of measuring spatial coherence is especially evident in the Gram matrix eigenvalue contribution due to its variability. For a false edge ratio of 0.02, a filtering power of 0.01 efficiently detects most false edges, while higher sensitivity would lead to excessive deletion of true edges. Conversely, a false edge ratio of 0.1 requires a high filtering power of 0.05 to achieve a spatially coherent network. This approach thus helps to select an optimal filtering power adapted to the individual dataset. This is done without requiring ground truth or extra validation steps, and it could help reduce data loss in filtering algorithms.

Discussion

This study establishes spatial coherence as a purely computational validation method for sequencing-based microscopy networks that does not require additional experiments and does not use ground truth. It provides quantitative metrics that assess if the network topology is representative of Euclidean spaces.
We identified three primary measures to evaluate spatial coherence: network dimension estimation, computing the spatial constant, and spectral analysis of the Gram matrix. Each of these measures can detect the presence of errors like false edges and fused nodes in spatial networks. We stress that they are all derived from the powerful fact that Euclidean distance is correlated with the network's shortest path distance in spatial networks. Additional measures could be developed by analogies to other relative Euclidean geometric relationships, such as the relationship between distances described by the Pythagorean theorem substituted for shortest path distances, or the relationship between the surface area of a sphere of a given radius versus its volume exchanged with the number of nodes discovered by breadth first search of a certain depth.
These methods offer a framework for detecting inconsistencies and improving data quality in sequencing-based microscopy, that could improve confidence in network reconstructions and reduce the need for extensive prior knowledge or corroborative experimental validation. However, the current metrics, while effective in identifying false edges, could be refined to address their sensitivity to non-uniform sparse network topologies and the high computational complexity requirements for large networks. While this can be partially addressed by sampling the network, future work could focus on computational complexity performance.
Next steps could also involve exploring improved filtering algorithms to not only detect the presence of false edges but also correct them while simultaneously reducing the deletion of true edges, strengthening data reliability even more. All in all, this framework enables the possibility to assess different sequencing-based microscopy techniques under a common umbrella, potentially facilitating benchmarking both for the techniques and for their error correction algorithms. Moreover, such algorithms are not incompatible with spatial coherence measures: future work could incorporate spatial coherence to improve filtering algorithms. We believe that the insights derived from this study could improve the reproducibility and quality of network reconstruction in the new spatial omics techniques that are based on DNA barcode networks.

Methods

E. Algorithm description

E.1. Spectral Analysis of the Gram Matrix. The rank of a Gram matrix derived from the pairwise distances of a set of points X is indicative of the dimension $\mathcal{D}$ of X. Therefore, the first-$\mathcal{D}$ eigenvalues of the Gram matrix account for all the contribution while the rest of the spectra have null values. In particular, a Gram matrix is defined by the dot product of its set of points X as $G = XX^T$. However, the point coordinates are not necessary to compute the Gram matrix as it is also related to the set's pairwise distances D (Eq.5), known as the Euclidean Distance Matrix (EDM) Dokmanic et al. (2015):

$$D^2 = \text{diag}(XX^T)\mathbf{1}^T - 2XX^T + \mathbf{1}\text{diag}(XX^T)^T \quad \textbf{(5)}$$

Here, $\mathbf{1}$ represents a column vector of all ones, and diag$X$ denotes the column vector of the diagonal entries of the point coordinates matrix X.
This framework is generalized to networks, using the Shortest Path Distance Matrix D' (SPDM) as opposed to the EDM D. To obtain the network's Gram matrix $G = XX^T$ Eq.5 is simplified by using the geometric centering matrix J

$$J = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \quad \textbf{(6)}$$

where I_n is the $n \times n$ identity matrix and $\mathbf{1}_n$ is an n-dimensional column vector of ones. Applying J results in positioning the centroid of X at the origin, effectively removing the terms diag$(XX^T)\mathbf{1}^T$ and $\mathbf{1}$diag$(XX^T)^T$

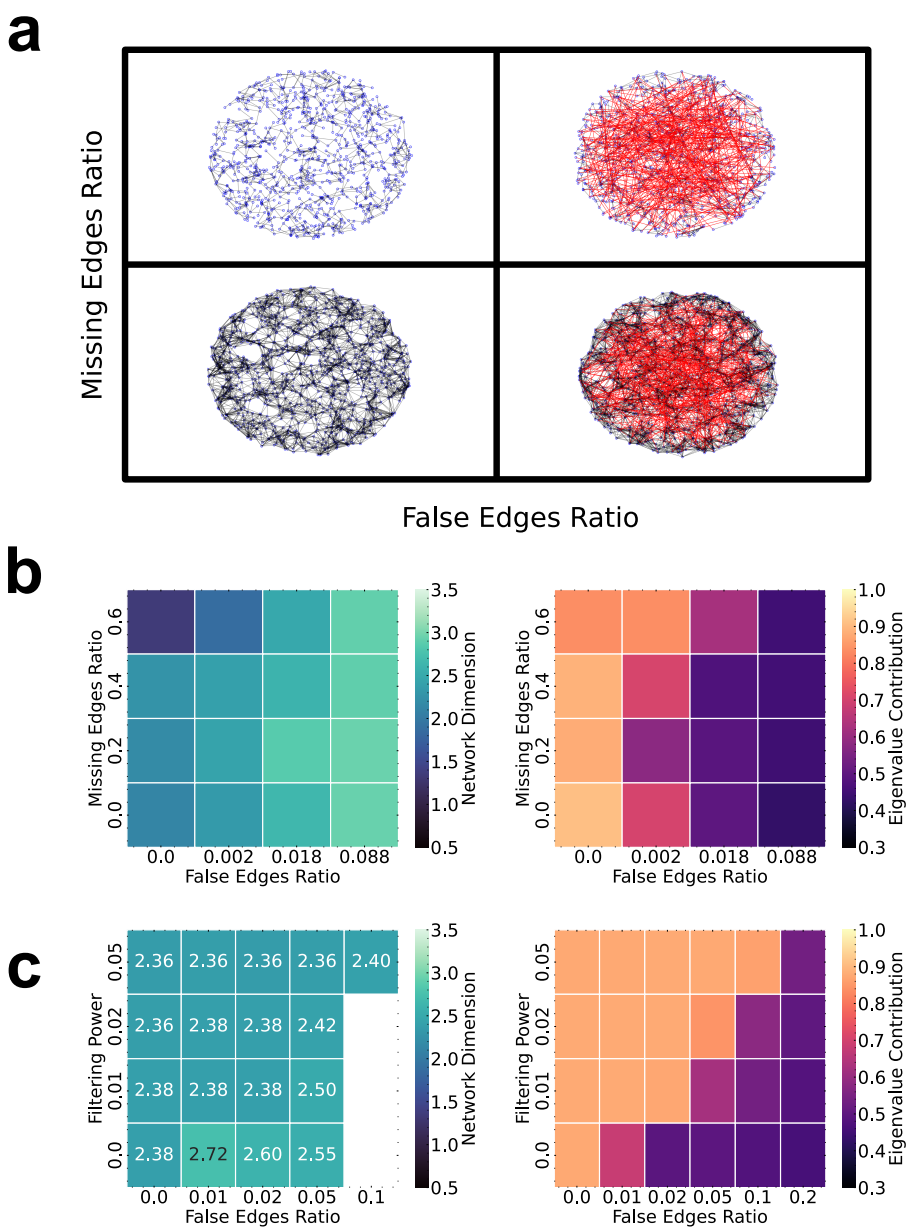

Figure 7. Effect of missing edges and error correction filtering on spatial coherence measures. a. Visual representation of networks in four quadrants defined by varying ratios of false edges and missing edges, illustrating several scenarios. **b.** Heatmaps showing spatial coherence across two axes: false edges ratio and missing edges ratio; the color intensity represents the magnitude of the network dimension (left), and the Gram matrix eigenvalue contribution (right). **c.** Similar heatmaps as in **b**, but displaying the interaction between false edge ratio and filtering power.

in Eq.5. This is known as double centering D'^2. Therefore, the Gram Matrix can be obtained from D'^2 by applying Eq.7:

$$G = -\frac{1}{2} J D'^2 J \tag{7}$$

The eigenvalues of G coincide with the principal components obtained via Principal Component Analysis, up to a scaling factor Hastie et al. (2009). Therefore, the magnitude of the eigenvalues is linked with the variance explained by each eigenvector. The normalized contribution of the first $\mathcal{D}$ eigenvalues $(\lambda_1, ..., \lambda_{\mathcal{D}})$ is expressed as:

$$C_{\mathcal{D}} = \frac{\sum_{i=1}^{\mathcal{D}} \lambda_i}{\sum_{j=\mathcal{D}+1}^{N} \lambda_j} \tag{8}$$

This measure is bounded between 0 and 1, and it reflects the extent to which the first $\mathcal{D}$ dimensions capture the variance of the data.

An additional measure is that of the spectral gap, defined as the normalized difference between the last $\mathcal{D}$ and the $\mathcal{D}+1$ eigenvalue:

$$\Delta\lambda = 1 - \frac{\lambda_{\mathcal{D}+1}}{\lambda_{\mathcal{D}}} \tag{9}$$

where $\Delta\lambda$ denotes the spectral gap. A larger gap indicates a greater distinction between the relevant dimensions and the residual dimensions, with the largest gap occurring in a case where the residual dimensions take null values.

E.2. Network Dimension. The network's dimension is computed based on the Euclidean principle that the number of points (nodes) within a certain distance r increases according to a power-law, where the exponent corresponds to the dimension of the space. To estimate the network's dimension, the shortest path distances d' act as proxies for Euclidean distances d.
First, several central network nodes are identified in order to minimize finite-size effects. Centrality in our network analysis is defined by the closeness centrality measure, which identifies nodes that are in proximity to all other nodes in the network, thereby minimizing finite-size effects. Closeness centrality for a given node is typically calculated as the reciprocal of the sum of the shortest path distances from that node to all other nodes in the network. Mathematically, for a node i, the closeness centrality $C(i)$ is defined as:

$$C(i) = \frac{1}{\sum_{j=1}^{N} d'(i,j)} \tag{10}$$

where $d'(i,j)$ is the shortest path distance between node i and node j, and N is the total number of nodes in the network. Nodes with higher closeness centrality are used as starting points for a BFS of depth r', as they are statistically more likely to represent the central structure of the network and reduce boundary biases in network traversal.
After central nodes are identified, the number of nodes N within a depth r' is computed for all possible values of r', starting at $r' = 1$, where N represents the number of immediate neighbors one hop away from the origin. The gathered data is used to compute the fit $\log(N)$ against $\log(r')$ that enables finding the power defining the power-law behavior, or the network's dimension. To ensure linearity, a sliding window of points is iteratively selected, and the segment providing the highest coefficient of determination is chosen. The slope of this linear fit is used to estimate the network's dimension, which in the case of a proximity network closely aligns with the dimensionality of the underlying Euclidean space.

E.3. Spatial Constant. Analogous to the mean line segment in Euclidean space, the spatial constant $\mathcal{S}$ measurement is concerned with the mean distance between all pairs of nodes.
In Euclidean space, the mean line segment of a circle is Solomon (1978):

$$\langle d \rangle \approx 0.9r \tag{11}$$

where r is the radius of the circle. But $N = \rho V = \rho \pi r^2$, and Eq.11 becomes

$$\langle d \rangle = 0.9 \left(\frac{N}{\pi \rho} \right)^{\frac{1}{2}} \tag{12}$$

The spatial constant is chosen by definition as the dimensionless parameter $\mathcal{S} = \frac{0.9}{\sqrt{\pi}}$ that does not depend on density nor the number of points. By rearranging the terms, it can be generalized to the network's case:

$$\mathcal{S} = \langle d' \rangle \left(\frac{\langle k \rangle}{N} \right)^{\frac{1}{\mathcal{D}}} \tag{13}$$

where $\langle d' \rangle$ is the mean shortest path distance between nodes, N is the number of nodes, $\mathcal{D}$ is the network's original dimension, and $\langle k \rangle$ is the average degree of the network. The reason for using $\langle k \rangle$ as an analog of the Euclidean density ρ is because the average degree accounts for how many nodes there are on average at one space unit.
The assessment of the spatial constant's behavior is done by sampling the network at different size scales. First, origin nodes are selected at random and a BFS traversal of fixed depth is conducted. This depth is increased iteratively, allowing for the discovery of more nodes and effectively increasing the size N. Different BFS depths represent different network sizes. Therefore, this approach measures how the spatial constant changes with network growth. A consistent spatial constant that mirrors the behavior in Euclidean spaces would suggest the absence of false edges in the network. Conversely, a varying spatial constant indicates that the network's average shortest path length varies across different scales, pointing out that the behavior is not consistent and could potentially have false edges.

F. Algorithm implementation

All algorithms are implemented in Python 3.10 and the code will be publically available both as a GitHub repository and as a Python package. The implementation relies at its core on efficient network representations, shortest path distance computations, and BFS traversals.

F.1. Network representation. This study leverages sparse matrix representations for network data, specifically using the Compressed Sparse Row (CSR) format. The choice is motivated by the efficiency in storing and manipulating sparse graphs, which have only a fraction of the potential $N(N-1)/2$ edges. Spatial networks are sparse in that sense because the number of edges is usually bounded by spatial constraints and scales proportionally to the number of nodes $N \propto E$. Using a CSR format can accelerate procedures to compute network properties, such as degree distribution and shortest path distance matrices. However, for the latter, they are memory intensive to store as the space requirements scale quadratically with the number of nodes.

F.2. Shortest Path Distance Matrix. The SPDM D' denotes the shortest path distance D'_{ij} between nodes i and j. To compute the SPDM from the CSR representation of a network, we employ Dijkstra's algorithm, optimized for sparse matrices. In particular for this work, the shortest path distance matrix of a network is used to

obtain the eigenvalues of its Gram matrix and to compute the network's dimension.

F.3. Breadth-First Search. Breadth-First Search (BFS) is implemented by selecting an origin node and using a queue to keep track of the next nodes to visit. The depth of the traversal can be chosen to study the network at different scales, where higher depths result in bigger scales. Exploring different scales is used for the spatial constant computation, where several origin nodes are selected and BFS traverals with varying depths are run from them. This results in multiple subgraphs representing different network regions, each with its corresponding spatial constant.

F.4. Large networks. In scenarios where the network's scale makes computations such as the shortest path matrix or eigenvalue decompositions prohibitively expensive, especially in terms of space complexity, the network can be sampled. This is done by selecting a starting node and running a BFS traversal until a certain number of nodes is reached. This threshold is user-customizable and can be adapted depending on hardware.

G. Simulated data processing

G.1. Synthetic data set. Networks are simulated by generating uniformly distributed points confined in a region, by default a circle in 2D and a sphere in 3D. We remark that changing the shape is not impactful on the bulk properties of the network, as it only affects the boundaries. After point generation, the network is created following the general criteria that points that are close in space interact and form an edge. If not stated otherwise, the K-Nearest Neighbor proximity is used with $K = 6$ in 2 dimensions and $K = 15$ in 3 dimensions, a choice motivated by the average degree in Voronoi tesselations. The generated networks are bipartite, meaning that the nodes are divided into two distinct sets such that no two nodes within the same set share an edge. This mimics popular experimental settings in sequencing-based microscopy.
For the specific experimental setting of DNA microscopy on diffusing amplicons Weinstein et al. (2019), the simulated data set is obtained by running the pipeline at https://doi.org/10.5281/zenodo.10256692 Kloosterman et al. (2024) with a default amplitude and spread of 10. The filtering algorithm from the same pipeline is used to study the spatial coherence before and after filtering.

G.2. False edges and missing edges. Modifications in the original network's topology are introduced in the form of false edges and missing edges. False edges are created by selecting pairs of nodes that are not originally connected and linking them. This is done until the desired false edges number or ratio is satisfied. Conversely, missing edges are created by removing existing edges. If the deletion disconnects the network, another pair of nodes is chosen. As with false edges, the process is iterated until the desired sparsity of ratio is satisfied.

G.3. US County dataset. Data was acquired from https://census.gov/cgi-bin/geo/shapefiles/index.php?year=2023&layergroup=Counties+%28and+equivalent%29. The resulting shapefile was processed using Python's package geopandas. The inferred network was created by connecting mainland counties based on adjacency.

Author Contributions

DFB implemented the algorithms, characterization, and computational exploration in the study. JIB and SL retrieved and analyzed data from the literature and run its algorithms. All authors contributed conceptual insights. DFB, ITH and SKD wrote the manuscript. DFB and ITH conceived the study.

Acknowledgements

We acknowledge support from the Swedish Research Council (no. 2020-05368 to ITH), and the European Research Council ERC (no. 949624 to ITH). The authors wish to thank Erik Benson (Karolinska Institutet), Antti Elonen and Pekka Orponen (Aalto University), and Ragnar Thobaben (KTH Royal Institute of Technology) for helpful discussions and insights.

Bibliography

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J. C. High-throughput spatial mapping of single-cell rna-seq data to tissue of origin. *Nature biotechnology*, 33(5):503–509, 2015.

Barrat, A. and Weigt, M. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13:547–560, 2000.

Bonet, D. F. and Hoffecker, I. T. Image recovery from unknown network mechanisms for dna sequencing-based microscopy. *Nanoscale*, 15(18):8153–8157, 2023.

Boulgakov, A. A., Xiong, E., Bhadra, S., Ellington, A. D., and Marcotte, E. M. From space to sequence and back again: Iterative dna proximity ligation and its applications to dna-based imaging. *bioRxiv*, page 470211, 2018.

Boulgakov, A. A., Ellington, A. D., and Marcotte, E. M. Bringing microscopy-by-sequencing into view. *Trends in biotechnology*, 38(2):154–162, 2020.

Dokmanic, I., Parhizkar, R., Ranieri, J., and Vetterli, M. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

Gawad, C., Koh, W., and Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 2016.

Glaser, J. I., Zamft, B. M., Church, G. M., and Kording, K. P. Puzzle imaging: Using large-scale dimensionality reduction algorithms for localization. *PloS one*, 10(7):e0131593, 2015.

Gopalkrishnan, N., Punthambaker, S., Schaus, T. E., Church, G. M., and Yin, P. A dna nanoscope that identifies and precisely localizes over a hundred unique molecular features with nanometer accuracy. *bioRxiv*, 2020.

Greenstreet, L., Afanassiev, A., Kijima, Y., Heitz, M., Ishiguro, S., King, S., Yachie, N., and Schiebinger, G. The dna-based global positioning system—a theoretical framework for large-scale spatial genomics. *bioRxiv*, 2022.

Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., Massasa, E. E., Baydatch, S., Landen, S., Moor, A. E., et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542(7641):352–356, 2017.

Hastie, T., Tibshirani, R., and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Hoffecker, I. T., Yang, Y., Bernardinelli, G., Orponen, P., and Högberg, B. A computational framework for dna sequencing microscopy. *Proceedings of the National Academy of Sciences*, 116(39):19282–19287, 2019.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R. P. The drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, 2017.

Karlsson, F., Kallas, T., Thiagarajan, D., Karlsson, M., Schweitzer, M., Navarro, J. F., Leijonancker, L., Geny, S., Pettersson, E., Rhomberg-Kauert, J., et al. Molecular pixelation: Single cell spatial proteomics by sequencing. *bioRxiv*, pages 2023–06, 2023.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. In situ sequencing for rna analysis in preserved tissue and cells. *Nature methods*, 10(9): 857–860, 2013.

Kloosterman, A., Baars, I., and Högberg, B. An error correction strategy for image reconstruction by dna sequencing microscopy. *Nature Computational Science*, pages 1–9, 2024.

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., Turczyk, B. M., Yang, J. L., Lee, H. S., Aach, J., et al. Fluorescent in situ sequencing (fisseq) of rna for gene expression profiling in intact cells and tissues. *Nature protocols*, 10(3):442–458, 2015.

Lewis, S. M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., and Naik, S. H. Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9):997–1012, 2021.

Mathieson, L. and Moscato, P. An introduction to proximity graphs. *Business and Consumer Analytics: New Ideas*, pages 213–233, 2019.

Moses, L. and Pachter, L. Museum of spatial transcriptomics. *Nature methods*, 19(5):534–546, 2022.

Qian, N. and Weinstein, J. A. Volumetric imaging of an intact organism by a distributed molecular network. *bioRxiv*, 2023.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434): 1463–1467, 2019.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

Schaus, T. E., Woo, S., Xuan, F., Chen, X., and Yin, P. A dna nanoscope via auto-cycling proximity recording. *Nature communications*, 8(1):1–9, 2017.

Shanker, O. Defining dimension of a complex network. *Modern Physics Letters B*, 21(06): 321–326, 2007.

Shapiro, E., Biezuner, T., and Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

Solomon, H. *Geometric probability*. SIAM, 1978.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

Wang, G., Moffitt, J. R., and Zhuang, X. Multiplexed imaging of high-density libraries of rnas with merfish and expansion microscopy. *Scientific reports*, 8(1):1–13, 2018.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature*, 393 (6684):440–442, 1998.

Weinstein, J. A., Regev, A., and Zhang, F. Dna microscopy: optics-free spatio-genetic imaging by a stand-alone chemical reaction. *Cell*, 178(1):229–241, 2019.

Zador, A. M., Dubnau, J., Oyibo, H. K., Zhan, H., Cao, G., and Peikon, I. D. Sequencing the connectome. 2012.

Zemel, R. and Carreira-Perpiñán, M. Proximity graphs for clustering and manifold learning. *Advances in neural information processing systems*, 17, 2004.