PNAS
www.pnas.org

Supplementary Information for

A Computational Framework for DNA Sequencing-Based Microscopy

Ian T. Hoffecker, Yunshi Yang,Giulio Bernardinelli,Pekka Orponen, and Björn Högberg

Corresponding author Björn Högberg.
E-mail: bjorn.hogberg@ki.se

This PDF file includes:

Supplementary text
Figs. S1 to S15
References for SI reference citations

Supporting Information Text

A. Polony Seeding and Voronoi Tessellations. Our basic mathematical model comprises a family of polonies generated by a finite set of N seed points $P = \{p_1, \ldots, p_N\}$ placed on a 2D circular disk $C \subseteq \mathbb{R}^2$ according to a Poisson point process of intensity λ. (Hence in particular the expected number of seed points and respectively polonies is $E[N] = \lambda A$, where $A = A(C)$ is the area of C.) The polonies expand at a uniform rate from their seed points until their boundaries meet, so that at saturation they form a (Poisson-)Voronoi tessellation T of the disk C.

Denoting the Euclidean distance metric on $\mathbb{R}^2$ by $d(x, y)$, the polony generated by seed point p_k, $k = 1, \ldots, N$, is thus represented by the Voronoi cell $T_k \in T$ containing the set of points in C relatively closest to p_k:

$$T_k = \{x \in C \mid d(x, p_k) < d(x, p_j) \text{ for all } j \neq k\}$$

B. Chance of Barcode Degeneracy. An individual polony is identified by its barcode, a sequence over the alphabet $\Sigma = \{A, T, G, C\}$. The set of all possible barcode sequences w of length b is denoted Σ^b and has cardinality $B = 4^b$.

The probability of all the $N \sim \lambda A$ polonies on C having unique barcodes for a sequence of length b can be represented in terms of binomial coefficients by the expression

$$\Pr(nondegenerate) = \frac{\binom{B}{N}}{\binom{B+N-1}{N}}$$

which is approximated by the series

$$\Pr(nondegenerate) \approx 1 - \frac{(N-1)N}{B} + \frac{(N-1)^2N^2}{2B^2} - \frac{(N-1)^2N^2(N^2-N+1)}{B^3} \cdots = 1 - O(N^2/B)$$

From the expression one can see that for $B \gg N^2$, i.e. $b \gg \log_2 N$, the probability of nonunique barcodes goes rapidly to zero. For instance, for an ensemble of $N = 30,000$ polonies as in the example in the main paper and a $b = 20$ nucleotide barcode, and assuming a naive, uniformly distributed, base composition, the probability of all the codes being unique is $p \approx 0.9992$. We thus consider the ideal scenario where $b \gg \log_2 N$ and each polony T_k is uniquely distinguishable by its barcode sequence w_k. Then the existence of a barcode pair $\{w_i, w_j\}$ implies physical adjacency of two polonies T_i and T_j, between which crosslinks have associated their unique identifiers.

C. Heuristic Rationale for the Tutte Reconstruction Method. Any given point set P in general position on the disk C has a unique Delaunay triangulation $D = \langle P, L \rangle$, determined by the important *empty circle property*: no circumcircle of any triangle constituted from the points in P and their connecting line segments L contains a point from P in its interior (1, Section 9).

Now, for our point set reconstruction problem, we only know the Delaunay triangulation D of P as an abstract graph $G = (V, E)$. Since G is triangulated, it has a topologically unique plane embedding on the disk C, given a counterclockwise (say) listing of the vertices on the outer boundary. However, information about the vertex neighbourhoods in $D = \langle P, L \rangle$ as given by G does not yet fully determine the geometric locations of the points in P.

Statistically speaking, the problem concerns determining the conditional distribution $\Pr(P \mid G)$ of the locations of the Poisson-generated point set P, given that the graph of its Delaunay triangulation is G. (Or maybe computing a maximum-probability arrangement for P under this distribution.) However this distribution is not easy to characterize explicitly. It could be sampled by a Markov Chain Monte Carlo technique based on the observation that the local distribution of a single point $p \in P$ conditioned on the locations of the other points, $\Pr(p \mid P \setminus \{p\}, G)$, is uniform within the *feasible region* $F_p(P \setminus \{p\}, G) \subseteq C$ available for p in the given arrangement of the points in $P \setminus \{p\}$.

The feasible region $F_p = F_p(P \setminus \{p\}, G)$ for p is obviously contained within the polygon spanned by the present locations of its immediate Delaunay-graph neighbors $N_1(p) = \{p_1, \ldots, p_k\} \subseteq P \setminus \{p\}$. Let us call this the *Delaunay polygon* $V_p = V(p_1, \ldots, p_k)$. Because of the characteristics of Delaunay triangulations, the feasible region F_p does not cover all of V_p, and its shape actually depends on the present arrangement of the points in $N_1(p)$ and *their* Delaunay-graph neighbors, i.e. the set $N_2(p)$ of points at graph distance 2 from p – but no further. One could thus simplify the prospective MCMC sampling of the distribution $\Pr(P \mid G)$ by the observation that for every $p \in P$, $\Pr(p \mid P \setminus \{p\}, G) = \Pr(p \mid N(p), G)$, where $N(p) = N_1(p) \cup N_2(p)$; but it would still be a complex and time-consuming process.

Now the Tutte reconstruction method can heuristically be seen as a fast deterministic approximation to sampling the distribution $\Pr(P \mid G)$ by simply picking for each point p a location likely to be in F_p, namely the barycenter of the polygon V_p. Or to be more precise, the reconstruction algorithm explicitly arranges the points in P in such a way that every point p is located at the barycenter of the Delaunay polygon V_p spanned by its neighbors $N_1(p)$, thereby aiming to provide a feasible sample (albeit only one) from the distribution $\Pr(P \mid G)$.

D. A Topological Hierarchy in Poisson-distributed Delaunay Triangulations. Here we aim to establish the connection between the Euclidean metrics of the Delaunay triangulation, whose major spatial constraints directly result in the topological relationships that form, and the purely topological metrics of the untethered graph. While the main text contains a proof of concept essentially demonstrating this fact, underlying this phenomenon are basic principles of stochastic geometry.

The relationship between Euclidean distance and topological distance was investigated empirically by generating Delaunay triangulations and measuring both Euclidean and topological distances between points. In order to avoid bias due to boundary effects, vertices were not chosen randomly from Delaunay triangulations. This could result in points near the boundary to be selected, and topological distances may then be skewed. Instead, after generating a Delaunay triangulation for a given random point set bounded by a circle, we used betweenness centrality to select one of the central vertices as an origin point, and topological distances (and their corresponding Euclidean distances) were sampled by proceeding one edge at a time away from the origin.

Given a known boundary geometry, e.g. a circle with known radius a, and knowledge that polonies are Poisson-distributed, we conjecture that for any two vertices in the non-embedded graph G with a N length topological shortest path distance, there exists a vertex along that shortest path with $N-1$ topological distance that is also closer to the origin in *Euclidean* distance.

Conjecture 1 Let $\{k_0, k_N, k_{N-1}\} \subseteq K$ be any subset of three vertices in K that satisfies the property that k_N has at least one topological shortest path (i.e. hop count or number of edges) leading to a so-called origin k_0 equal to N steps, and let k_{N-1} be a vertex located on a topological shortest path to k_N with its own topological shortest path to k_0 equal to $N-1$ steps. Let the set $\{p_{k_0}, p_{k_N}, p_{k_{N-1}}\} \subseteq P$ be the corresponding set of real coordinate points in D in $\mathbb{R}^2$. The points are Poisson-distributed.

Then it is conjectured that

$$\langle d(p_{k_0}, p_{k_N})\rangle > \langle d(p_{k_0}, p_{k_{N-1}})\rangle \forall k_0, k_N \in K.$$

In other words, the average Euclidean distance between an origin point and a destination point is greater than that of the average distance from the origin to a point 1 topological step closer to the origin than the destination.

E. Tutte Embedding Respects the Topological Hierarchy of the Delaunay Triangulation. We make a conjecture analogous to Conjecture 1 about the coupling of Euclidean and topological metrics in the Tutte embedding, a suggestion that a Tutte embedding satisfies one of the major properties of the *a priori* distribution.

Conjecture 2 Let $\{k_0, k_N, k_{N-1}\} \subseteq K$ be any subset of three vertices in K that satisfies the property that k_N has at least one topological shortest path (i.e. hop count or number of edges) leading to a so-called origin k_0 equal to N steps, and let k_{N-1} be a vertex located on a topological shortest path to k_N with its own topological shortest path to k_0 equal to $N-1$ steps. Let the set $\{p'_{k_0}, p'_{k_N}, p'_{k_{N-1}}\} \subseteq P$ be a corresponding set of real coordinate points in the embedded graph D' in $\mathbb{R}^2$. Then it is conjectured that

$$\exists\, k_{N-1} | d(p'_{k_0}, p'_{k_N}) > d(p'_{k_0}, p'_{k_{N-1}}) \forall k_0, k_N \in K.$$

We make this conjecture based on the observation that attempts to violate this rule by artificially placing a distance N vertex closer in Euclidean distance to some origin point than a distance $N-1$ vertex invariably result in the formation concave structures that are forbidden by the laws of the Tutte embedding. Another related observation is that the topological contour lines surrounding a vertex may never cross each other in Euclidean space. A topological contour n may not be closer at all points than some points of topological contour $n-1$, however there will always be some point on contour $n-1$ that is closer to the origin in Euclidean distance.

F. Description of the topological distance matrix reconstruction. A topological path $path(i,j)$ in a graph is defined as a fully connected ordered set of edges $e_1, e_2, ...e_n$ connecting two vertices i and j in which all vertices visited are unique. The path length $\mid paths(i,j) \mid$ is then the number of edges comprising a path. We define a topological distance function as the length of the path with the fewest edges connecting two points or $t(i,j) = min(\mid path(i,j) \mid \forall path(i,j))$.

Let us compile a distance matrix $\boldsymbol{t}$ using $t(i,j)$ for the untethered graph G composed of N columns and N rows where N is the number of vertices in G, and each element $\boldsymbol{t}_{i,j}$ is a topological distance $t(i,j)$ between vertices corresponding to the ith row and jth column, taking the form:

$$\boldsymbol{t} = \begin{pmatrix} t_{0,0} & \dots & t_{N,0} \\ \vdots & \ddots & \\ t_{0,N} & & t_{N,N} \end{pmatrix}$$

F.1. total topological distance matrix reconstruction. Where each topological distance element is obtained using the Floyd-Warshall algorithm (2). We then perform a linear transform to a new coordinate system by determining the eigenvectors $v_1 ... v_N$ and their corresponding eigenvalues $\lambda_1 ... \lambda_N$ of the covariance matrix $\boldsymbol{K_{tt}}$ compiled by computing the covariances of $\boldsymbol{t}$. The 2 vectors with the highest-valued eigenvalues are used to approximate x and y coordinate vectors.

F.2. peripheral topological distance matrix reconstruction. Alternatively we utilize only the topological distances between all vertices and each of the peripheral vertices. Peripheral vertices can be found using face enumeration (above) to determine the face with the most edges.

G. Sequencing benchmark estimates for experimental implementation. Bridge PCR methods can achieve densities of one million $polonies/mm^2$ for an approximate 1 μm^2 average polony size (3). For 1000 copies per polony, a 1 mm^2 tissue slice would generate approximately 1,000,000,000 strands with an expected value of 6,000,000 unique edges (number of polonies times the expected number of neighbors of a typical 2D Poisson Voronoi cell (4)), and about 32,000,000 total cross-pairing events assuming the theoretical limit of linking yield based on our scaling measurements from Figure S3. Assuming no changes in these proportions during subsequent amplification and processing steps, an Illumina NextSeq which can deliver up to 400 million reads, would be expected to capture about 12,800,000 crosslinking reads, the rest being self-pairing reads. Let us estimate coverage with a simplified formula, sampling with replacement from a pool of unique items:

$$E(k, N) = N\left(1 - \left(\frac{N-1}{N}\right)^k\right)$$

Sampling from a pool of $N = 6,000,000$ unique edges a total of $k = 12,800,000$ times, we get an expected network coverage of 88%. We emphasize here that this is largely speculative, and meant to provide a rough expectation of order of magnitude.

H. Resolution benchmark estimates for experimental implementation. The maximum theoretical resolution depends on polony size (Figure 5, main text), and thus depends primarily on the polony generation approach used. With bridge PCR, a typical polony size could be on the order of 1 μm^2 with 1,000,000 $polonies/mm^2$, and thus an estimated FWHM (Figure 5i) point spread function of $\approx 0.9\,\mu m$. However denser polony generation techniques such as template walking amplification (5) rely on a much more locally confined mechanism of amplification requiring denser primer surface. Polony composition for these techniques is likely to be close to the immobilization density of primers, and thus should be smaller for a given ρ. Furthermore, a lower ρ value such as 125 sites per polony within an acceptable range for reconstruction could be used to push polony size to its smallest usable size.An estimate for high density primer immobilization is $1.1 \times 10^{13} molecules/cm^2$ (6), and assuming a maximal packing of sites per polony and $\rho = 125$ this would give an average polony size of approximately 0.001 μm^2 or roughly 1000 nm^2 and corresponding estimated FWHM of approximately 45 nm. Again, it should be emphasized here that this is an informal estimate designed mainly to provide an educated guess as to the order of magnitude for a best-achievable resolution.

I. Towards an optimal graph approximation algorithm. We desire an approximation approach that maximally utilizes the available information while obeying a maximum entropy or minimally presumptive criterion. That is, given an untethered graph and the knowledge that it is derived from a Poisson-Delaunay generating process, what is the best possible positioning of vertices? Such a model should be derived from properties of the originating Delaunay triangulation.

A property that intuitively should be satisfied is what we may call the same-Delaunay-topology criterion, in other words a new Delaunay diagram D' generated from the *a posteriori* point positions P' should have the same topological structure as D, i.e. $G' = G$. In this case P' shares in common with P all of the spatial constraints expressed in D. In fact our Levenshtein distance metric provides this, offering a route to possibly iteratively reconstructing a best-possible reconstruction, i.e. one which when a Delaunay triangulation is computed from the new positions, produces the exact same set of edges as those used to begin the reconstruction in the first place.

For small graphs (on the order of 10's of points) we could perform iterative adjustment to the positions, initialized with the Tutte embedding, until they form a Delaunay diagram with the same topology as the original one. This was accomplished with a simulated annealing computation where with each iteration, the Delaunay diagram generated by the current set of Euclidean coordinates (Figure S13b orange lines) is compared with the untethered graph (Figure S13b red lines), its topology derived from the Delaunay diagram of the *a priori* seed distribution. After adjustment, a final graph satisfies the property that its Delaunay diagram has the same topology as the untethered graph (Figure S13c). The *a posteriori* positions that satisfy this same-Delaunay topology criterion thus satisfy all of the constraints of the original Delaunay diagram.

Due to the computational inefficiency of the simulated annealing approach however, a much more desirable direction for future work would be the development of a Tutte-like embedding which, deterministically and in linear time, computes a non-degenerate distribution that satisfies the major properties of the Delaunay diagram (for example that vertices should be placed at the centers of mass of their topological neighbors), and ideally computes in one shot a graph satisfying the same-Delaunay criterion with $lev_{G,G'} = 0$.

References

1. de Berg M, van Krefeld M, Overmars M, Cheong O (2008) *Computational Geometry: Algorithms and Applications, 3rd Rev. Ed.* (Springer-Verlag).
2. Floyd RW (1962) Algorithm 97: shortest path. *Communications of the ACM* 5(6):345.
3. Adessi C, et al. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* 28(20):e87–e87.
4. Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) *Stochastic geometry and its applications.* (John Wiley & Sons).
5. Ma Z, et al. (2013) Isothermal amplification method for next-generation sequencing. *Proceedings of the National Academy of Sciences* 110(35):14320–14323.
6. Peterson AW, Heaton RJ, Georgiadis RM (2001) The effect of surface probe density on dna hybridization. *Nucleic acids research* 29(24):5163–5168.

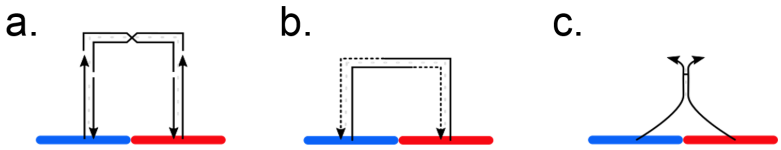

Fig. S1. Schematic representation of strategies for the formation of network connections. We identified both enzymatic (a-b) and non-enzymatic (c) strategies for the connection of the information that reside in the locally amplified DNA strands. (a) The first proposed strategy, takes advantage ofa bridging double stranded DNA to be added at the end of the polony formation reaction together with a ligase enzyme. The ligated molecules could then be released from the solid support by stripping or hydrolysis of the anchor chemistry and sequenced. (b) A second enzymatic strategy is characterized by a partial annealing of neighboring DNA strands and polymerase-mediated elongation according to a similar principal as that of bridge-PCR (3). (c) Alternatively, incorporating an unnatural nucleotide (7) in polony strands would enable the possibility to induce - in this case with the help of a 365 nm light irradiation - a covalent connection with an adjacent strand. A sequencing-ready library could be generated from this material surface isolation in an emulsion-PCR like setup (8).

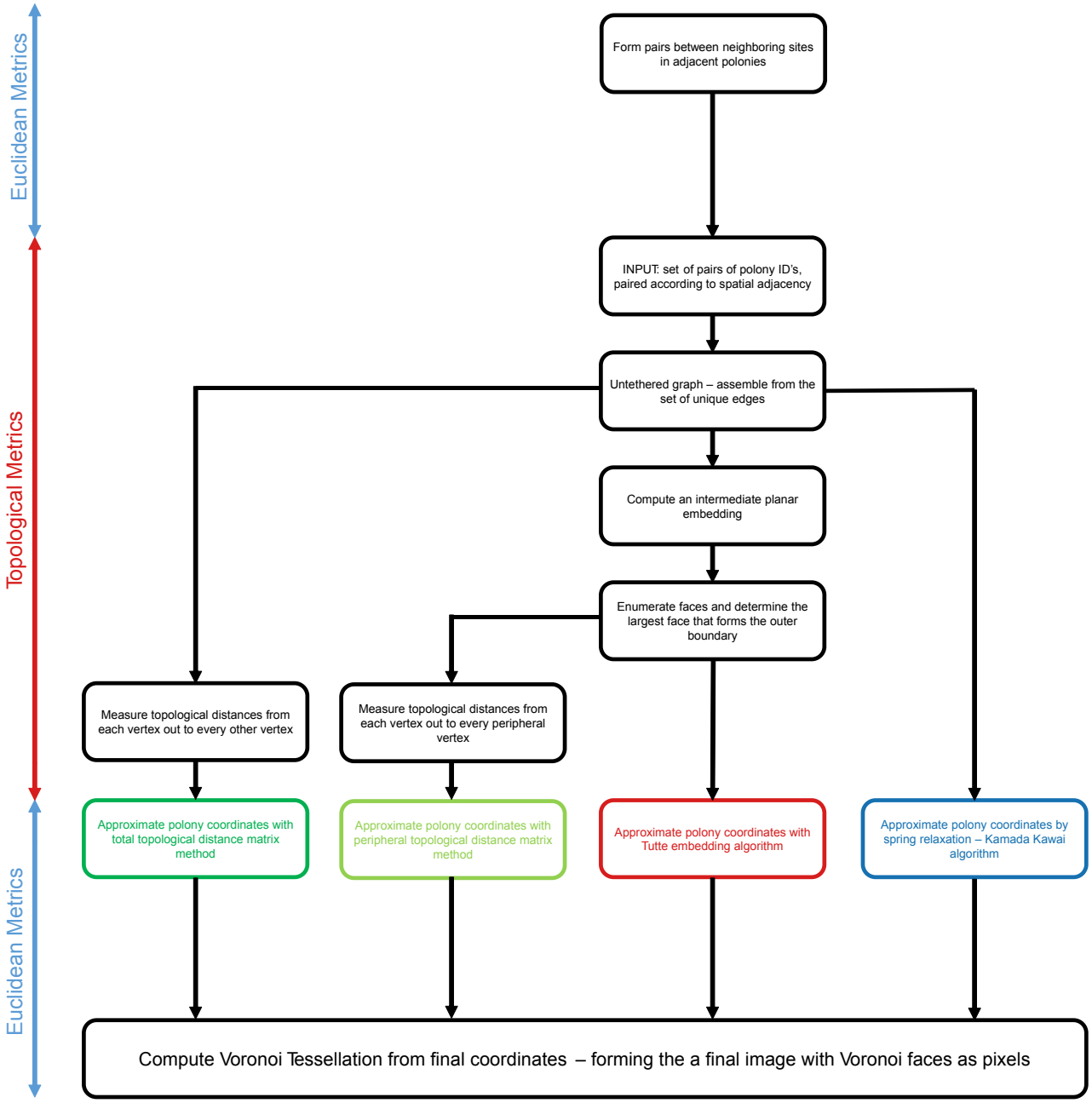

Fig. S2. Flow chart of the spatial adjacency mapping algorithm. Initial input data comes from a Euclidean metric space but is converted into a topological metric space. From topological metrics, we assemble a planar embedding, identify a largest face, and arrange the largest face on an outer boundary to then compute the final positions of interior points with the Tutte embedding algorithm. Alternatively, the spring relaxation approach of Kamada and Kawai can be used to compute final positions. And in the final two approaches, a pseudo-distance matrix is compiled from topological distances (shortest paths between vertices). In the case of total distance matrix reconstruction, all topological distances in the network are included in an NxN matrix. In the alternative variant, only peripheral vertices are used as reference points, and a matrix of all the topological distances of internal points out to these peripheral points is compiled. In both cases, the 2 principal component vectors of these matrices are used for the final coordinates. All four embedding approaches result in a return to the Euclidean metric space and planar representation of the graph.

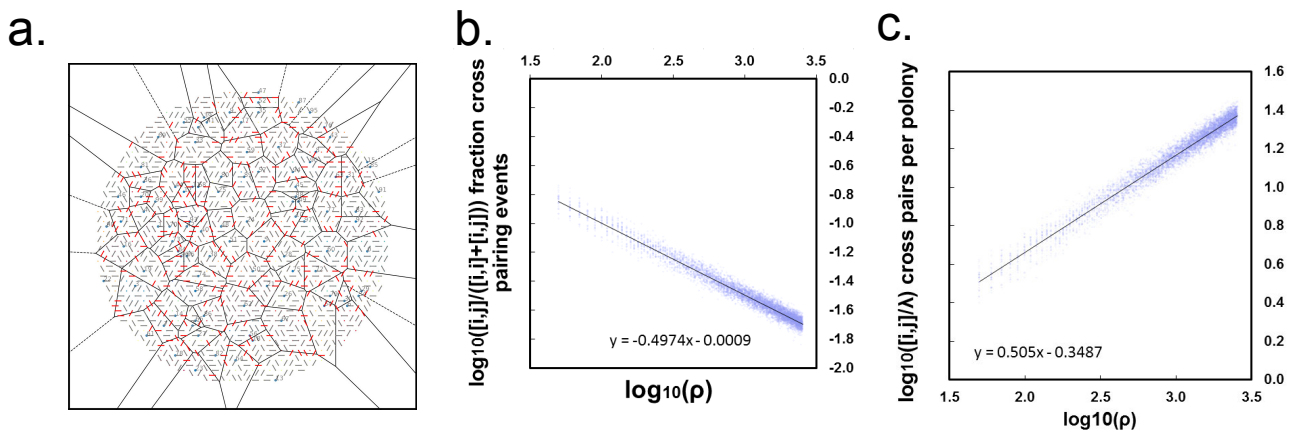

Fig. S3. Analysis of crosslinking efficiency. As the average amplicon copy number, i.e. the number of oligonucleotide sites comprising a polony ρ is increased, there is an increasing chance that a crosslinking event between two different, adjacent polonies occurs. As maximization of this characteristic is desirable, it would seem reasonable to use a high ρ. However this has the unwanted side effect of diluting the usable information content on a per read basis due to the non-proportional scaling of crosslinking events with self-pairing events as ρ is increased. (a) Example of a simulated polony surface with crosslinking events highlighted in red and self-pairing events shown in gray. We can appreciate here the probabilistic nature of edge formation which, above a certain average number of sites per polony, becomes highly probable, but at critically low values may result in missed edges. (b) Log-log plot of the fraction of crosslinking events out of total pairing events versus sites per polony ρ showing that the fraction of crosslinking events scales according to $\propto 1/\sqrt{\rho}$. Moreover, the number of crosslinking events per polony yields diminishing returns with increasing ρ: (c) Log-log plot of the number of crosslinking events per polony versus ρ showing $\propto\sqrt{\rho}$ scaling, indicating that to double the number of crosslinking events, one should square the number of additional average sites per polony.

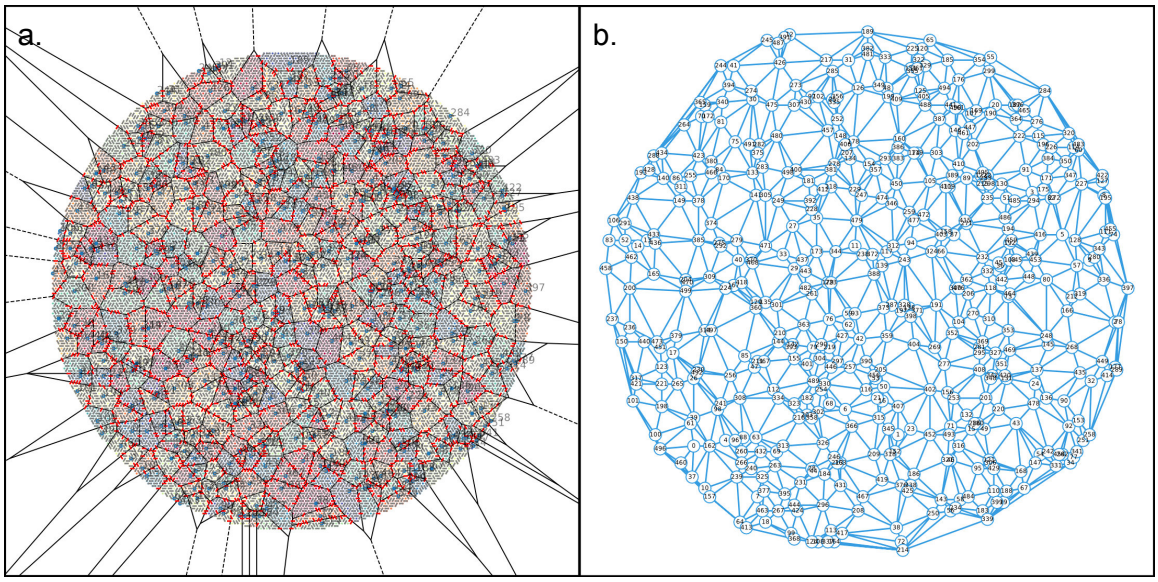

Fig. S4. (a) Example of a simulated saturated polony surface with λ = 500 polonies/ROI and relative site density of ρ = 100 sites per polony after undergoing random neighbor crosslinking. Self-pairing events are colored in gray and crosslinking events are colored in red. (b) The corresponding Delaunay triangulation computed from the positions of polony seed molecules.

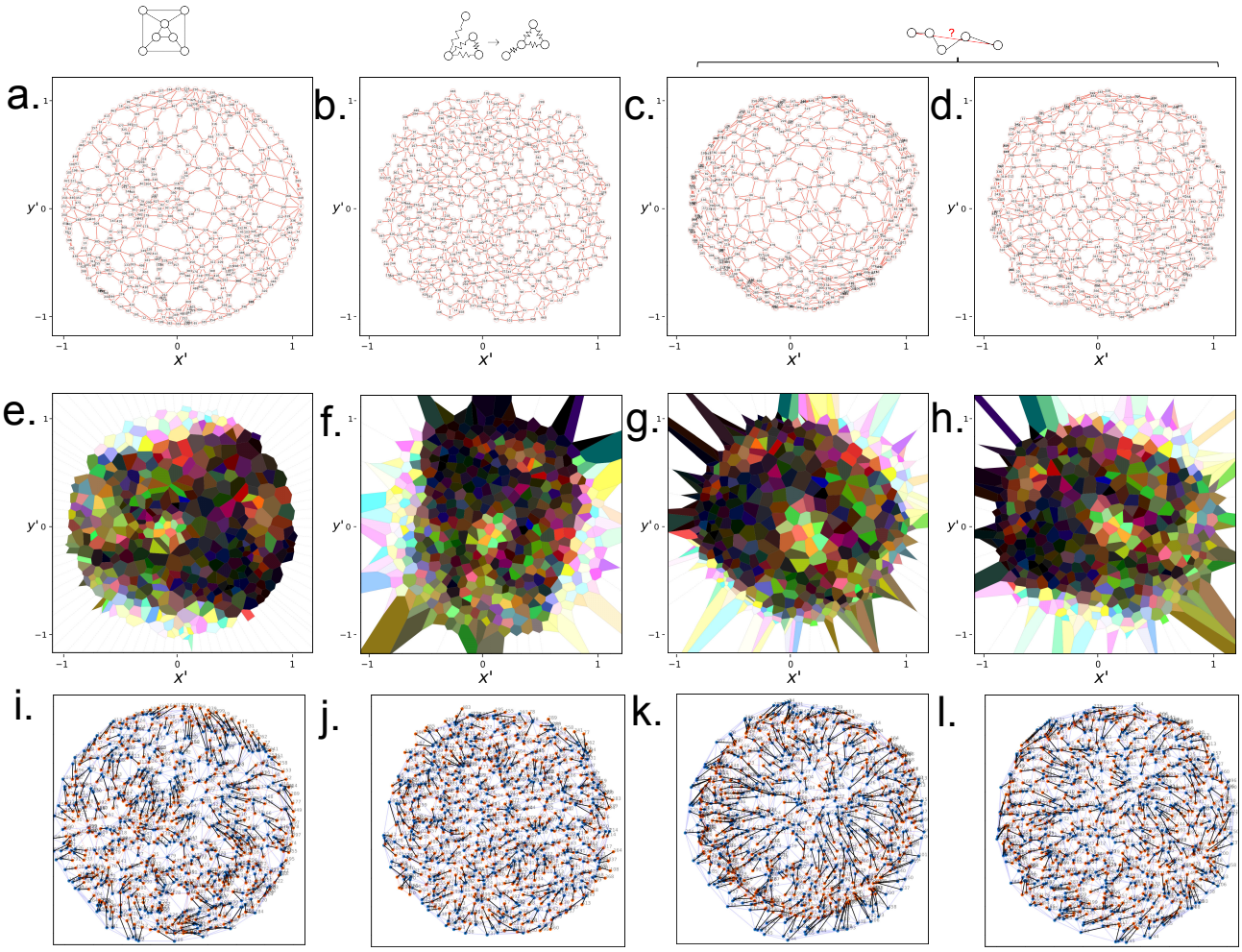

Fig. S5. Visual comparison of topological reconstruction approaches for λ = 500 polonies/unit area and ρ = 100 average sites per polony. (a-d) Final reconstructed graph embeddings produced, respectively, by Tutte embedding, spring relaxation, total distance matrix, and peripheral distance matrix approaches (e-h) Reconstructed Voronoi images corresponding to the reconstructed graphs from a-d generated using the 3-target color sampling procedure described in the main text for the four reconstruction approaches. (i-l) Alignment of *a priori* seed positions with the final *a posteriori* reconstructed positions for the four reconstruction techniques.

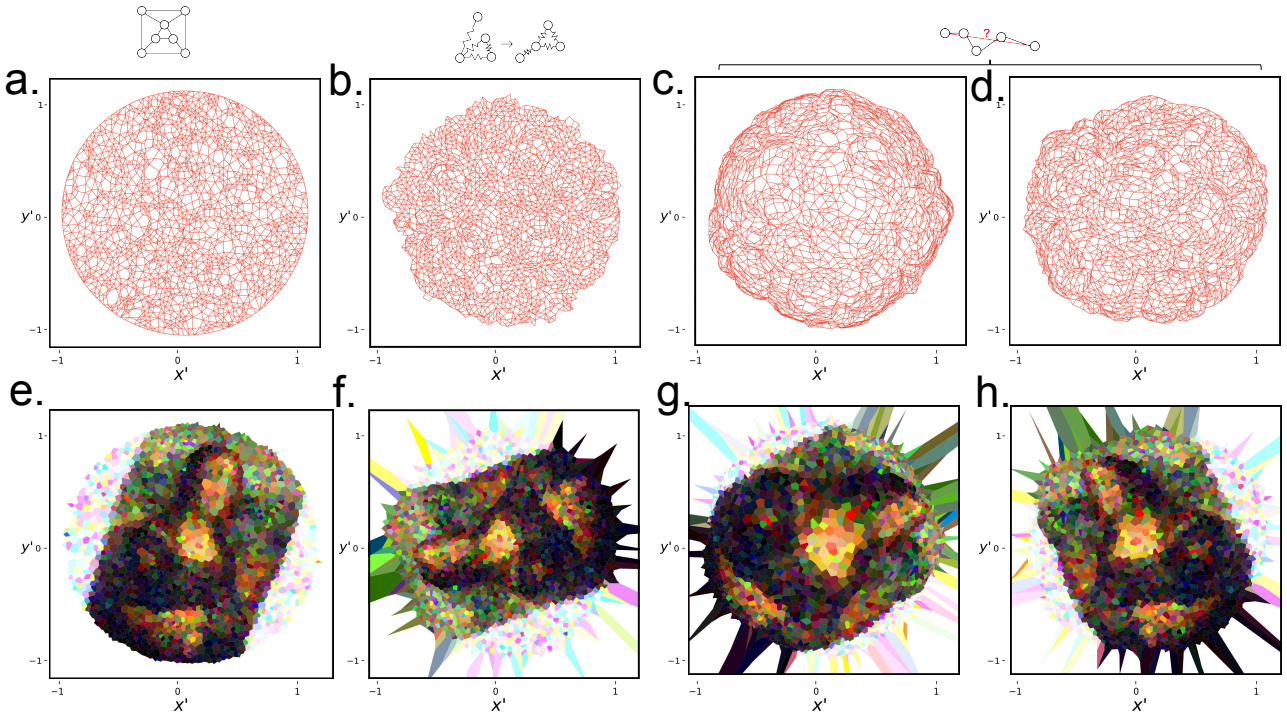

Fig. S6. Visual comparison of the four reconstruction approaches for a λ = 2000 polonies/unit area scenario and ρ = 500 average sites per polony.(a-d) Reconstructed graph embeddings for the four approaches respectively: Tutte embedding, spring relaxation, total topological distance matrix, and peripheral topological distance matrix. (e-f) Voronoi image reconstructions corresponding respectively to the graphs from a-d.

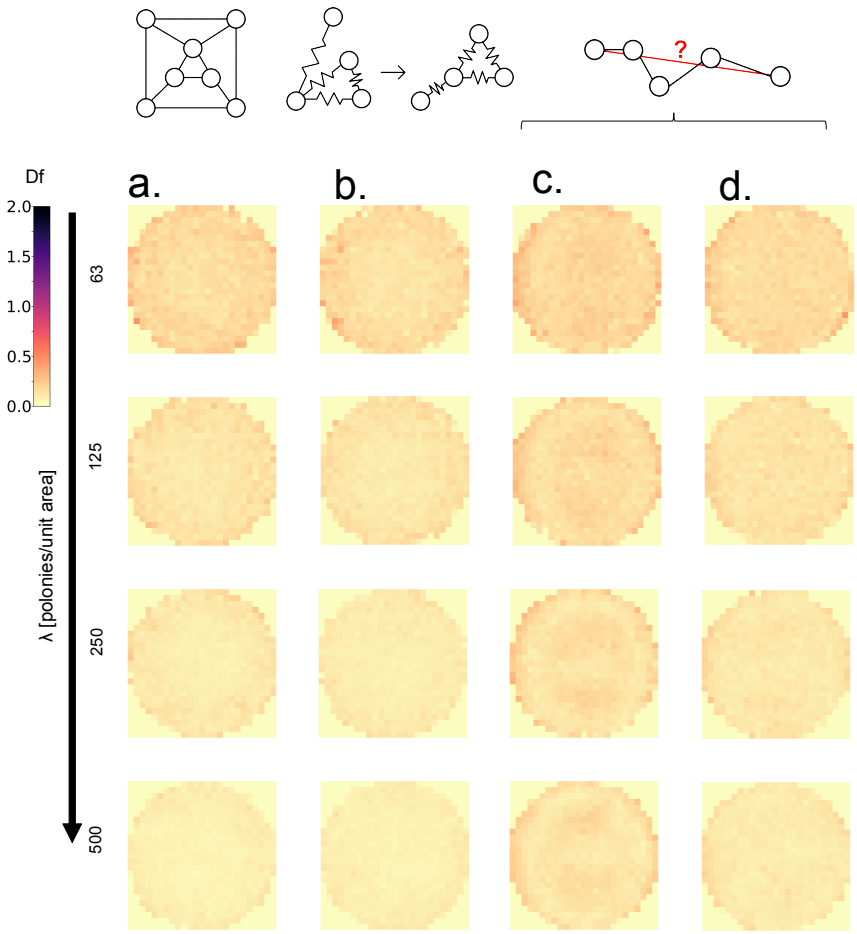

Fig. S7. Average distortion visualized as 2D histograms for varying polony densities λ showing progressive improvement in distortion with larger λ. Radial systematic distortions can be distinguished from local noise due to averaging effect. Distortions appear most severe near the edge. Color bar indicates scale of distortion, with the unit circle diameter 2.0 being the maximum. Distortion histograms are shown for (a) the Tutte embedding, (b) spring relaxation, (c) total topological distance matrix, and (d) peripheral topological distance matrix reconstruction approaches. ρ = 500 sites per polony for all, and each diagram represents n = 5000/λ independent simulation results.

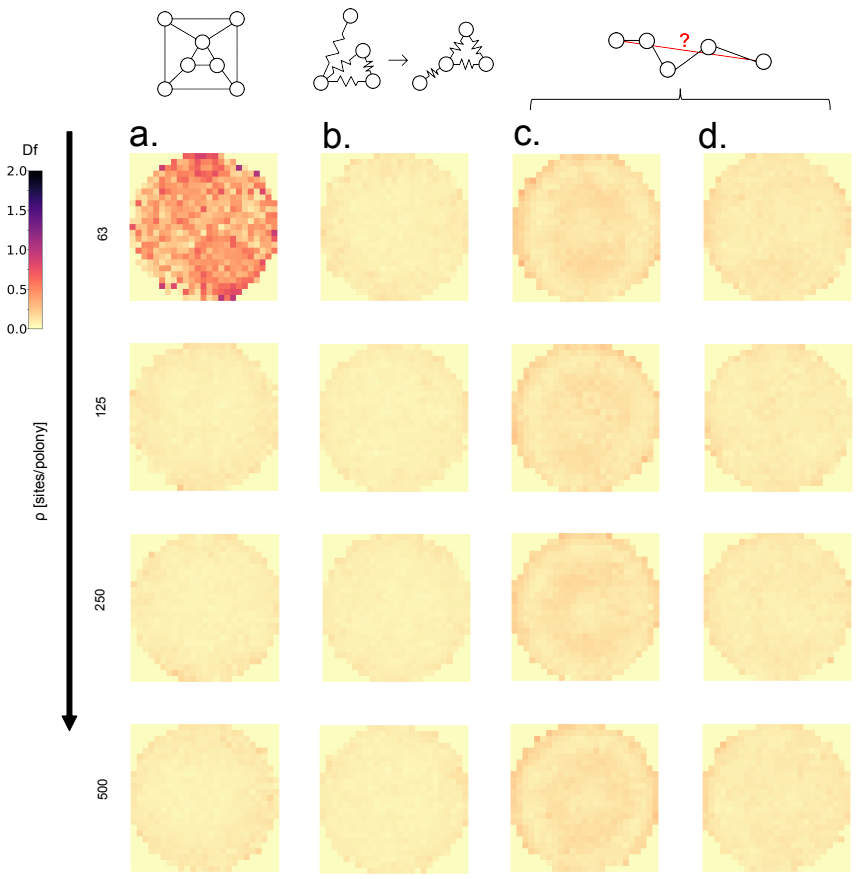

Fig. S8. Average distortion visualized as 2D histograms for varying relative site densities ρ. As ρ is increased, relatively negligible reduction in distortion is seen. However for critically low values, graphs begin to have defects affecting reconstruction. This critical point was observed for Tutte embedding graphs at ρ = 63 sites per polony, whereas the other approaches remained robust at this density. Distortion histograms are shown for (a) the Tutte embedding, (b) spring relaxation, (c) total topological distance matrix, and (d) peripheral topological distance matrix reconstruction approaches. λ = 500 polonies/unit area, and each diagram represents 10 simulation results.

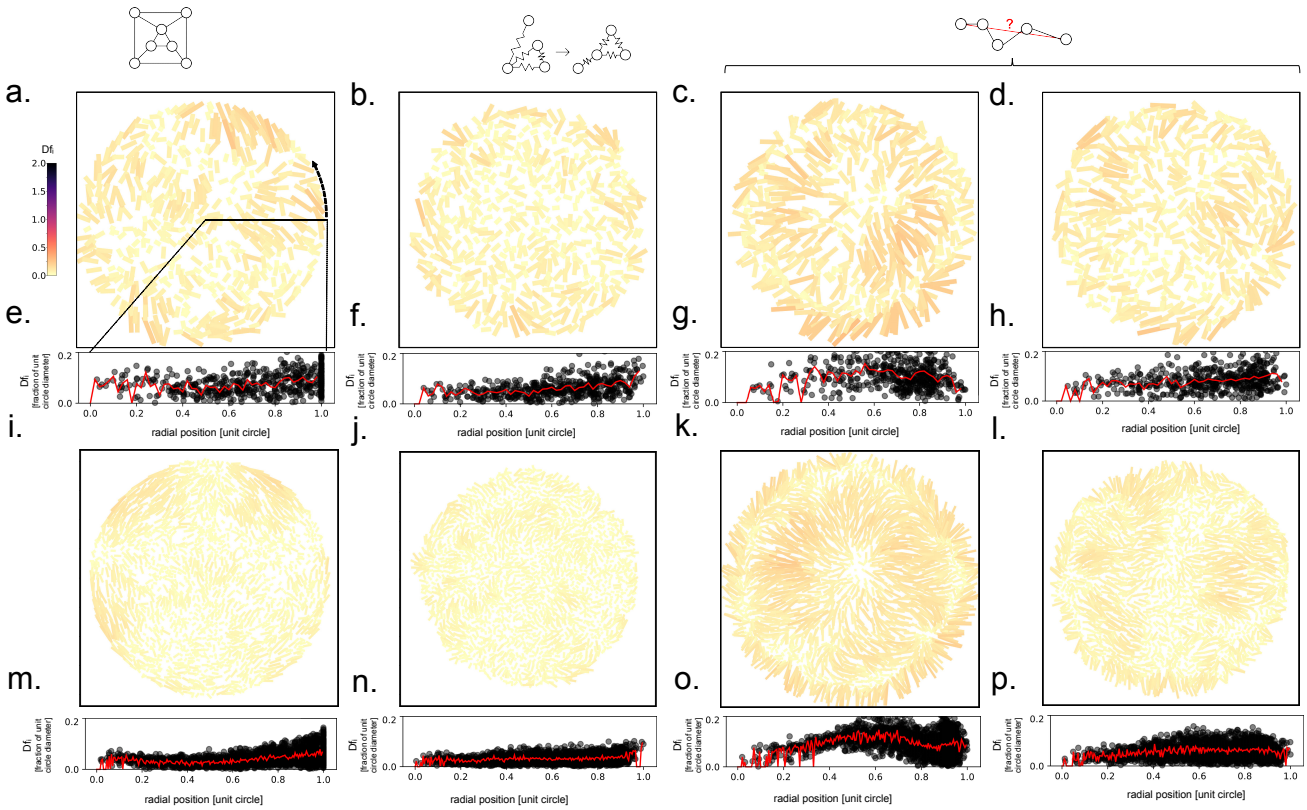

Fig. S9. (a-b) Distortion maps, visualized for single instance reconstructions for λ = 500 polonies/unit area for the four reconstruction approaches, respectively, Tutte embedding, spring relaxation, total topological distance matrix, and peripheral topological distance matrix. Lines show individual displacements Df_i connecting *a priori* seed molecule locations and the *a posteriori* reconstructions following alignment to obtain $min(\sum_{i=1}^{N} Df_i)$. Colormap indicates line length with unit circle diameter 2.0 being the maximum. (e-h) Radial profiles of distortion Df_i corresponding to the single instance distortion maps from a-d. (i-l) Distortion maps for single instance reconstructions for λ = 2000 polonies/unit area for the four respective reconstruction approaches. (m-p) Radial profiles of distortion Df_i corresponding to the single instance distortion maps from a-d.

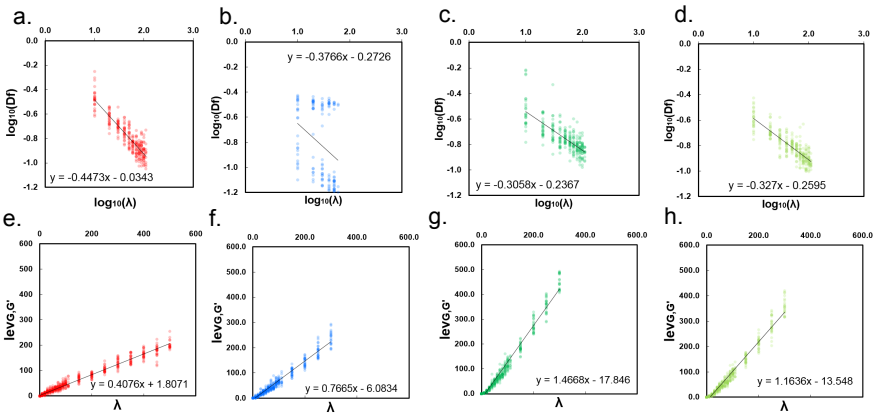

Fig. S10. Quantification of reconstruction quality as a function of polony density λ for the four reconstruction approaches respectively, Tutte embedding, spring relaxation, total topological distance matrix, and peripheral topological distance matrix. (a-d) Log-log plots of average distortion Df (ρ = 500 sites/polony) versus polony density λ. Slope values to the fits indicate scaling constant, with Tutte embedding ($Df \propto \lambda^{-0.447}$) outperforming the other techniques, i.e. the fastest improvement in distortion upon progressive addition of more polonies per unit area. Poorly performing simulations formed a secondary population in the spring relaxation case (b), due to the tendency for singly-connected vertices to diverge away from their adjacent neighbors. (e-h) Linear plots of the *a posteriori* Levenshtein distance error metric ($lev_{G,G'}$) or number of edits required to create identical topology between the original untethered graph and the graph derived from a post-reconstruction Delaunay triangulation. The Tutte embedding approach outperformed the other techniques followed by spring relaxation in $lev_{G,G'}$ scaling, i.e. the error grows slower upon progressive addition of polonies per unit area. Each condition (λ value) was simulated independently for n = 25 times, and each point represents a single simulation.

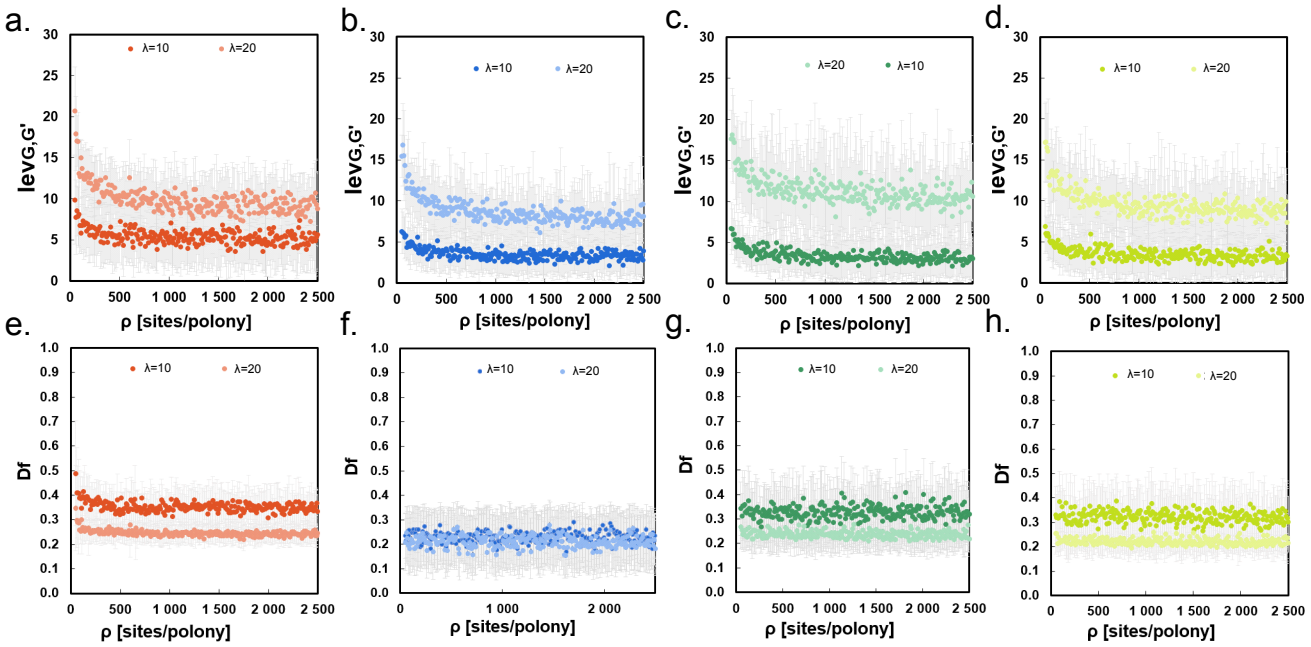

Fig. S11. Quantification of reconstruction quality as a function of relative site density ρ for the four reconstruction approaches respectively, Tutte embedding, spring relaxation, total topological distance matrix, and peripheral topological distance matrix. (a-d) Plots of the *a posteriori* Levenshtein distance metric ($lev_{G,G'}$) or number of edits required to create identical topology between the original untethered graph and the graph derived from a post-reconstruction Delaunay triangulation. Each are plotted for two values of λ = 10 and 20 polonies per unit area. Error bars represent standard deviations, n = 25 independent simulations per point. All approaches display a transient high error regime at low values of ρ which are more prominent for the greater λ value. (e-h) Average distortion Df plotted versus ρ for two values of λ = 10 and 20 polonies per unit area. Error bars represent standard deviations, n = 25 independent simulations per point. Transient high error regimes manifested less prominently in distortion compared to $lev_{G,G'}$.

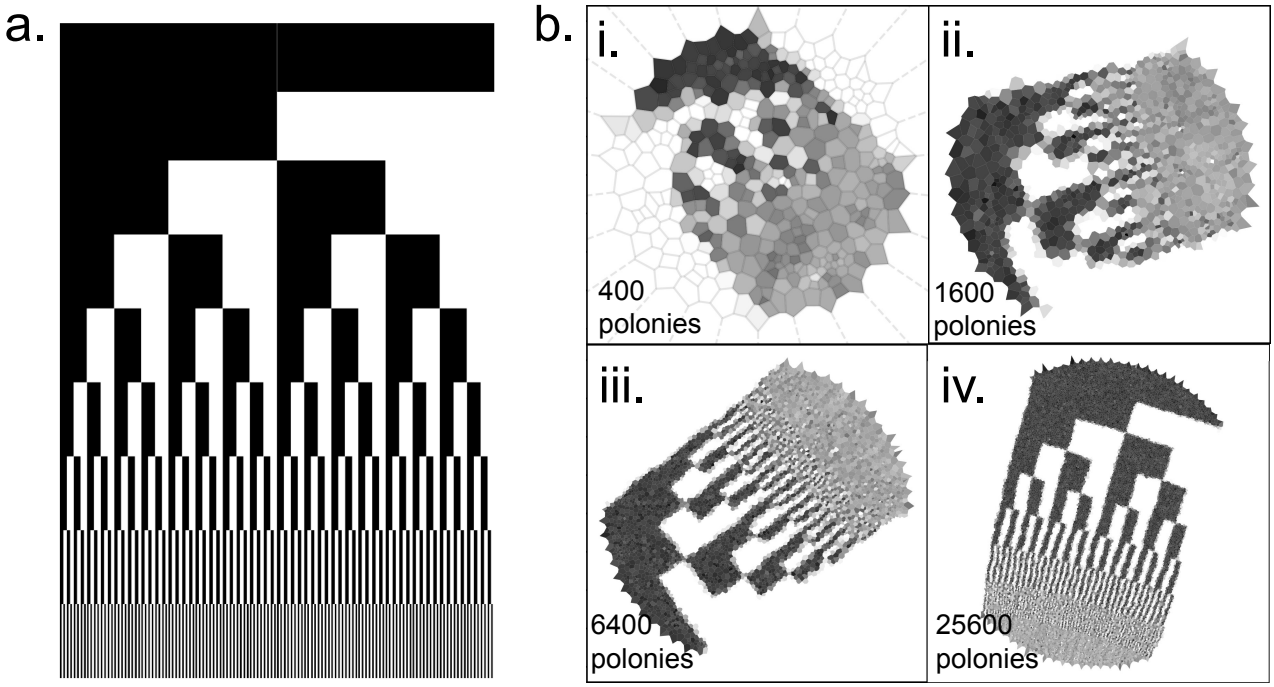

Fig. S12. Visualization of resolution limit and its dependence on polony density λ (polonies per unit area). (a) Image of black and white comb to be reconstructed. (b) Voronoi images with facets colored according to probability sampled from the original image, reconstructed with the Tutte embedding approach. Boxes i-vi show reconstructions with progressively quadrupled values of λ to approximately halve the minimum size of distinguishable features in the image each increment.

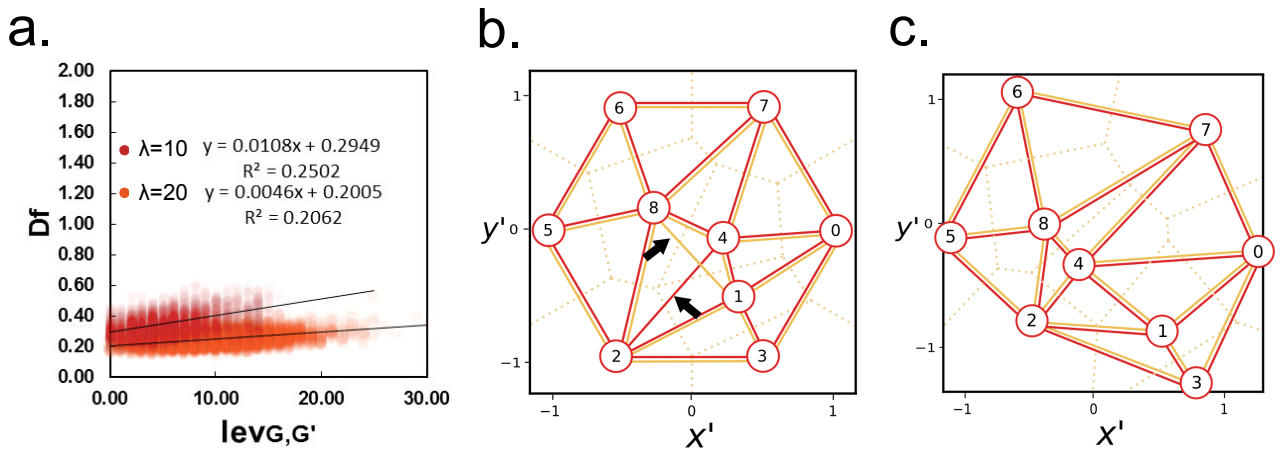

Fig. S13. Towards optimal exploitation of available information. (a) Plot of average distortion Df versus $lev_{G,G'}$ for two values of λ = 10 and 20 polonies per unit area and a fixed ρ = 500 sites per polony for the Tutte embedding approach. The figure indicates a weak but positive correlation between distortion and Levenshtein distance. Although Df requires omniscient knowledge of the *a priori* distribution, $lev_{G,G'}$ on the other hand is available *a posteriori* requiring only the untethered graph and a proposed reconstruction. Proportionality between the two indicates that $lev_{G,G'}$ could serve as a noisy proxy for Df as a quality metric. (b) An example Tutte reconstruction (red solid lines) compared with re-computed Voronoi tessellation (orange dotted lines) and re-computed Delaunay triangulation (orange solid lines). Edge discrepancy denoted with black arrows. (c) Iteratively adjusted embedding (red solid lines) compared with re-computed Voronoi tessellation (orange dotted lines) and re-computed Delaunay triangulation (orange solid lines).

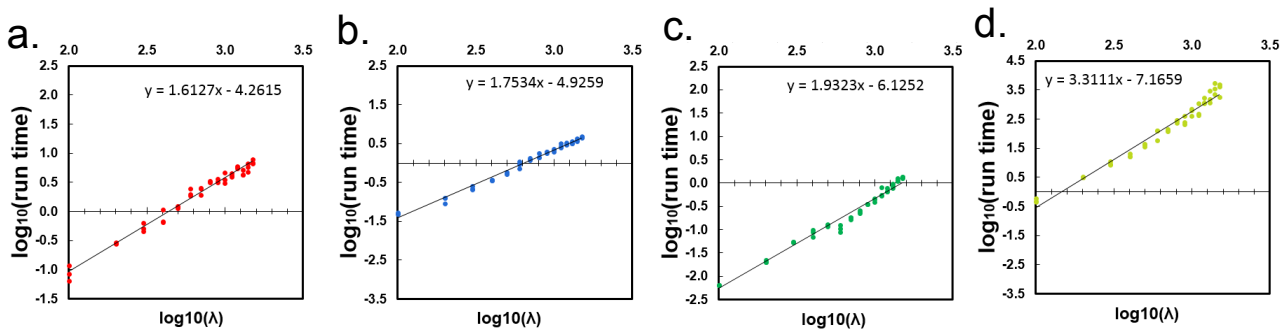

Fig. S14. Estimated algorithmic complexity based on run times as a function of λ (shown as log-log plots) for the four reconstruction techniques. All algorithms exhibit apparent polynomial time typical case complexity (worst case performance was not considered). (a) Tutte embedding approach scales according to $\approx O(\lambda^{1.61})$ and was the best performing of the four approaches followed closely by (b) spring relaxation which scales according to $\approx O(\lambda^{1.75})$. (c) This was followed by the total topological distance matrix method that scales according to $\approx O(\lambda^{1.93})$, and the worst-performing was the peripheral topological distance matrix method scaling according to $\approx O(\lambda^{3.31})$. Each condition ($\lambda$) value was simulated independently n = 3 times for a fixed ρ = 500 sites per polony.

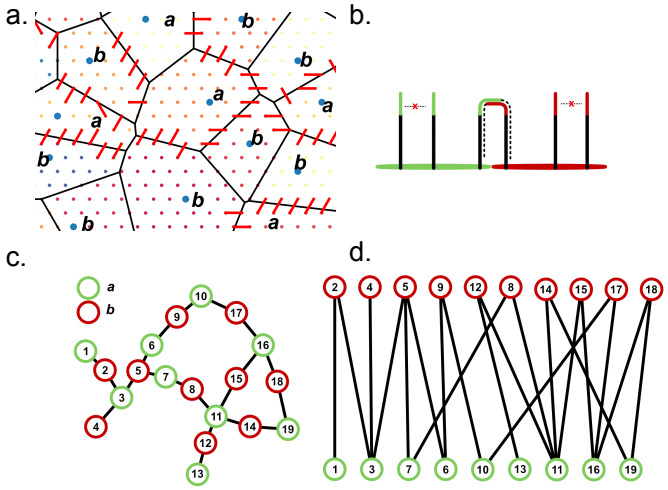

Fig. S15. Concept for bipartite network formation. (a) Two types (a and b) polonies are seeded whose interactions (b) are restricted to cross-complementary 3' end overlap, prohibiting self-interactions. (c) The resulting network has gaps where self-adjacency events occur, however the overall structure of the network should be preserved in cross-pairing events. (d) The final network can be represented in bipartite format showing an absence of intra-species linkages. This approach means that all reads would deliver information about polony adjacency.

7. Gerling T, Dietz H (2019) Reversible covalent stabilization of stacking contacts in dna assemblies. *Angewandte Chemie* 131(9):2706–2710.
8. Peikon ID, et al. (2017) Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Research* 45(12):e115–e115.