frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 05 November 2018
doi: 10.3389/fninf.2018.00074

Neural Reconstruction Integrity: A Metric for Assessing the Connectivity Accuracy of Reconstructed Neural Networks

Elizabeth P. Reilly, Jeffrey S. Garretson, William R. Gray Roncal, Dean M. Kleissas†, Brock A. Wester, Mark A. Chevillet† and Matthew J. Roos*†*

Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States

OPEN ACCESS

Edited by:
Andrew P. Davison,
FRE3693 Unité de Neuroscience, Information et Complexité (UNIC), France

Reviewed by:
Tolga Cukur,
Bilkent University, Turkey
Juan Nunez-Iglesias,
Monash University, Australia

***Correspondence:**
Elizabeth P. Reilly
elizabeth.reilly@jhuapl.edu
Matthew J. Roos
matt.roos@binarycognition.com

†Present Address:
Dean M. Kleissas, Gigantum, Washington, DC, United States
Mark A. Chevillet, Facebook, Menlo Park, CA, United States
Matthew J. Roos, Binary Cognition, Baltimore, MD, United States

Received: *21 December 2017*
Accepted: *09 October 2018*
Published: *05 November 2018*

Citation:
Reilly EP, Garretson JS, Gray Roncal WR, Kleissas DM, Wester BA, Chevillet MA and Roos MJ (2018) Neural Reconstruction Integrity: A Metric for Assessing the Connectivity Accuracy of Reconstructed Neural Networks. Front. Neuroinform. 12:74. doi: 10.3389/fninf.2018.00074

Neuroscientists are actively pursuing high-precision maps, or *graphs* consisting of networks of neurons and connecting synapses in mammalian and non-mammalian brains. Such graphs, when coupled with physiological and behavioral data, are likely to facilitate greater understanding of how circuits in these networks give rise to complex information processing capabilities. Given that the automated or semi-automated methods required to achieve the acquisition of these graphs are still evolving, we developed a metric for measuring the performance of such methods by comparing their output with those generated by human annotators ("ground truth" data). Whereas classic metrics for comparing annotated neural tissue reconstructions generally do so at the voxel level, the metric proposed here measures the "integrity" of neurons based on the degree to which a collection of synaptic terminals belonging to a single neuron of the reconstruction can be matched to those of a single neuron in the ground truth data. The metric is largely insensitive to small errors in segmentation and more directly measures accuracy of the generated brain graph. It is our hope that use of the metric will facilitate the broader community's efforts to improve upon existing methods for acquiring brain graphs. Herein we describe the metric in detail, provide demonstrative examples of the intuitive scores it generates, and apply it to a synthesized neural network with simulated reconstruction errors. Demonstration code is available.

Keywords: connectome, computer vision, segmentation, brain graph, evaluation, Electron Microscopy, Neural Reconstruction Integrity

1. INTRODUCTION

Traditionally, reconstructions of neural tissue at the voxel level are obtained by imaging tissue slices, mosaicing and aligning these 2D digital slices to form a 3D volume of voxels, and labeling voxels with unique neuron and synapse identifiers (Saalfeld et al., 2012; Takemura et al., 2013; Lee et al., 2016). If neuron and synapse relationships are annotated as well (e.g., the post-synaptic portion of synapse *i* is found on neuron *j*) then a brain graph reconstruction can be derived from the annotated tissue reconstruction. Herein we use the term *annotate* to encompass both labeling of voxels and identifying neuron-synapse relationships.

Although trained individuals can generate annotated reconstructions with high accuracy, the labor involved cannot feasibly scale to the larger tissue volumes needed to provide informative

graphs. Based on the labor estimate from a recent reconstruction effort (Kasthuri et al., 2015), it would take roughly 30,000 people-years to manually annotate a 1 mm^3 volume. To annotate tissue reconstruction at such scales, researchers are developing automated or semi-automated methods (Helmstaedter et al., 2011; Funke et al., 2012, 2018; Nunez-Iglesias et al., 2013; Knowles-Barley et al., 2016; Lee et al., 2017; Januszewski et al., 2018) with varying degrees of success. To aid in the continuing development of these methods, a variety of metrics have been developed to measure the accuracy of semi-automated reconstructions as compared to "ground truth"[1] reconstructions that are manually generated. Classic reconstruction metrics such as the Rand Index (Rand, 1971), and variations thereof operate at the voxel level—penalizing reconstructions for which all voxels of a given object do not have a corresponding object in the ground truth data with a one-to-one voxel match.

While neuronal morphology almost certainly plays a role in neural processing (e.g., dendritic integration and compartmental processing) it is likely that a graph representation composed solely of vertices (representing whole neurons or reconstructed portions) and directed edges (representing directed synapses) is nonetheless sufficient to allow for a substantial increase in our understanding of brain networks and the manner in which they process information. Richer insight can be obtained by layering attributes as reconstructions improve in fidelity. As such, there are disadvantages to limiting oneself to voxel-level reconstruction metrics given that many voxel-level errors (e.g., minor neuron segmentation errors) do not result in erroneous brain graph connections. Additionally, there are reconstruction techniques that do not operate on images (Marblestone et al., 2014) and thus cannot be fairly compared with image based techniques using voxel-level measures. We present the Neural Reconstruction Integrity (NRI) metric, which is designed to be sensitive to aspects of a reconstruction that relate to the underlying brain graph, while being insensitive to those that do not. This method allows for a direct assessment of graph connections, which may be performed even when annotations are not available or not created, as with emerging sequencing methods (Marblestone et al., 2014).

1.1. Terminology

A *brain graph* refers to an attributed graph, $G = (V, E, A)$ where V is a set of vertices representing neurons, E is a set of edges representing (directed) synapses, and A is a set of edge attributes. A directed synapse from neuron $u \in V$ to neuron $v \in V$ is denoted (u, v). One common edge attribute is location in Euclidean space, for instance the centroid of the synapse. It is also possible for the graph to have vertex attributes, however this is not necessary or relevant to the proceeding discussion. In the following, we assume direction of synapses is known, which results in a directed graph.

A (ground truth) brain graph may be generated, for example, through a combination of automated reconstruction algorithms and annotator proofreading over a ground truth image volume (Plaza, 2014). As such, we sometimes refer to a brain graph neuron as a *ground truth neuron*[2] and a brain graph synapse as a *ground truth synapse*. Throughout this paper, we call a single ground truth neuron G_i and it is an element of V. When calculating the Neural Reconstruction Integrity metric (to be discussed at length later), we will consider the pre-synaptic and post-synaptic terminals separately. In particular, this means that the set of all edges with post-synaptic terminals associated with G_i is the set $\{(u, v) \in E : v = G_i\}$ whereas the set of all edges with pre-synaptic terminals associated with G_i is the set $\{(u, v) \in E : u = G_i\}$.

A *graph reconstruction* of a brain graph is, likewise, a graph consisting of vertices and edges, denoted $S = (V', E')$. The vertices represent reconstructed neuron fragments and may be denoted individually by S_i for $i = 1, \ldots, |V'|$. Note that, as an example, a ground truth neuron in the original brain graph may correspond to two vertices in the reconstructed graph if the employed neuron segmentation algorithm split the ground truth neuron in two. In other words, for an imperfect graph reconstruction, the original brain graph and the reconstructed graph are not aligned, or there is not an identity mapping between the vertex sets of the graphs. The edges of the graph reconstruction correspond to detected synapses where the corresponding attribute indicates the estimated centroid of the synapse.

Graph connectivity, or simply connectivity, refers to the neuron-synapse-neuron relationships within the graph. In other words, when we try to evaluate connectivity accuracy, we are interested in evaluating how well the neuron to neuron relationships are identified, which is related to how well paths through the graph are reconstructed.

A *neural reconstruction* can refer to several things including an image reconstruction (labeled images) or as a brain graph reconstruction as described above. Throughout this paper, we use neural reconstruction, or simply reconstruction, to refer to a brain graph reconstruction, as we are focused on evaluating the connectivity of neurons via synapses.

Unless otherwise indicated, the term *local* is used to refer to a single-neuron focused analysis or to a small subset of neurons within a larger volume. The term *global* refers to a network level or full volume analysis. In other words, a global metric is one calculated over several (possibly connected) neurons found in the same volume.

2. MATERIALS AND METHODS

2.1. Evaluation Criteria

The primary function of the NRI metric is to evaluate the degree to which an annotated reconstruction contains a brain graph that is an accurate reflection of the true brain graph. In large part this implies an insensitivity to neuron segmentation errors that do not

[1]Given that even expert human annotators do not always agree as to the proper labeling of a voxel or object, *gold standard* may serve as better terminology than *ground truth*. However, we use *ground truth* since that is the term commonly used in machine learning literature. Errors in manual annotations are commented upon further in the Discussion section.

[2]Throughout this article we use the term *neuron* generically, with recognition that elements in the ground truth are likely to be fragments of neurons rather than whole neurons, and elements in the reconstruction may be neuron fragments, merged neurons, merged neuron fragments, or even something non-neuronal altogether. Use of terms such as neuron fragment or neuron element are sometimes used to draw attention to this fact.

impact the brain graph. However, additional metric qualities are desirable.

- *Can operate on relatively small volumes of ground truth data:* One of the largest challenges of evaluating the accuracy of a reconstruction is that little ground truth data is available due to the extensive manual labor needed to generate it. Typical graph similarity metrics are removed from consideration since the volume of ground truth data will be much smaller than that generated by semi-automated methods. As a result, the evaluation metric should not strictly be a graph connectivity metric, but rather a proxy metric that measures reconstruction aspects critical for representing an accurate graph.
- *Applicable at various levels of granularity:* The metric should be flexible enough to evaluate reconstructions at various levels of granularity including single neurons, a small number of neurons or neuron fragments, or large, densely-annotated volumes. This allows one to compute the metric on a variety of types of ground truth data (e.g., sparsely annotated or densely annotated). In addition it allows one to evaluate the fidelity of spatially restricted regions throughout a reconstruction volume as well as identify whether inaccuracies are uniformly scattered across the volume or if they are concentrated at a few poorly reconstructed neurons. Global evaluation (a single metric score computed from the annotation intersection of the reconstruction volume and the ground truth volume) would allow one to measure overall improvement of a reconstruction method across reconstruction iterations or compare between reconstruction methodologies.
- *Provides locally independent scores:* An intuitive requirement is that if an entire neuron is "ground truthed" (manually annotated) and scored by the metric, this score should not change if additional neurons are subsequently ground truthed and the metric is then reapplied to the original neuron. Similarly, if the metric is applied to a geometrically local region, the score should not change if a *spatially disjoint* region of the volume is subsequently ground truthed and the original region is re-scored. We highlight this requirement because we found that alternative metrics based on information theory failed to fulfill this criterion.
- *Scales well to larger reconstruction and ground truth volumes:* Computation of the metric should be feasible even as the size of reconstruction and ground truth volume grow over time. Both are expected to grow substantially in coming years thanks to improvements in data acquisition technologies and targeted efforts such as the Intelligence Advanced Research Projects Activity (IARPA) MICrONS program[3]. Based on expected output under that program, an evaluation metric should be capable of being computed on reconstruction volumes containing billions of synapses and hundreds of thousands of neurons, at a minimum.
- *Provide intuitive scores:* Ideally scores should fall in a limited range such as $[0, 1]$ and be intuitively commensurate with reconstruction errors.

[3]https://www.iarpa.gov/index.php/research-programs/microns

2.2. Previous Work

As our goal here is to assess the accuracy of a reconstruction as it pertains to the brain graph, metrics that only assess neuron segmentation are not sufficiently informative. For example, the error-free path length (Helmstaedter et al., 2011) measures the frequency of errors made during manual skeleton tracing. It is defined as the total length of neuron skeleton divided by the number of errors made during tracing. The connectivity of a neuron is not considered in this measure, simply how well the skeleton of a neuron is reconstructed.

Several existing methods of evaluation assess the voxel-level similarity of a reconstruction volume and a ground truth volume. For example, the Rand Index (Rand, 1971), Adjusted Rand Index (Hubert and Arabie, 1985), and Warping Index (Jain et al., 2010) are often utilized as image segmentation error measures. The Rand Index applied to annotated images is defined as the proportion of pairs of voxels that are paired in the same segment in both ground truth and the reconstruction. If both neurons and synapses are annotated, the Rand Index can correlate with brain graph accuracy in some cases. However, this scoring method can frequently give results that are poor characterizations of the accuracy of the reconstructed brain graph. For example, large groups of voxels may be mislabeled yet connectivity is unaffected (e.g., mislabeling many voxels at the edge of a large diameter synapse-free process). Conversely, only small groups of voxels may be mislabeled yet connectivity is substantially disrupted (e.g., voxels across dendritic spines are mislabeled, resulting in orphaned synapses on spine heads). It is possible to adapt the Rand index to handle point synapses rather than voxels. Even so, the Rand index includes true negatives which can result in an optimistic evaluation when true negatives dominate. Jain et al. (2011) makes note of this relationship, as they dismantled the Rand index and used precision and recall to measure voxel-based reconstructions, ignoring true negatives.

A more recently adopted voxel-level metric is the variation of information (Nunez-Iglesias et al., 2013; Plaza, 2014). Variation of information (VI) is an information theoretic measure defined as

$$VI(S, G) = H(G|S) + H(S|G) \tag{1}$$

where S is a reconstruction, G is ground truth, and H is the entropy function. It is possible to apply variation of information to abstracted neuron-synapse relationship information (the same information utilized by the NRI) rather than directly to voxel information, as suggested in Plaza (2014). In that case, the variation of information when applied to a fully annotated (both reconstruction and ground truth) neural network has a number of desirable properties. However, there is not a simple, well-behaved way to define *VI* for a single neuron. The key dilemma is that the $H(G|S)$ term cannot naturally be broken down into elements that are relevant to a single ground truth neuron while still providing locally independent scores (see section 2.1). In other words, if ground truth data is provided for additional portions of the volume, then VI calculated for one of the original ground truth neurons will likely change, even if the ground truth neuron were wholly contained in the smaller, original volume.

Arganda-Carreras et al. (2015) define the Rand F-score, which is based on probability distributions. It is closely related to the Rand Index and normalized between 0 and 1. They also introduce an information theoretic F-score, which is based on mutual information, is closely related to Variation of Information, and is normalized between 0 and 1. The authors explore the relationship of these two F-score variations of Rand and VI as applied to boundary maps (transformed to voxel-level image segmentations) and show that they are highly correlated on real data.

Another approach that is similar in spirit to NRI is a line graph-based Graph f_1 score (Gray Roncal et al., 2015). This metric also evaluates connectivity by focusing on true positive, false positive, and false negative pathways connecting synapses. However, this metric was applied only to dense full volumes and undirected graphs and performance on error sub-types was not systematically evaluated.

More recently, the tolerant edit distance (TED) was proposed as a segmentation evaluation metric aimed at assessing topological correctness (Funke et al., 2017). The TED was used in the 2016 Medical Image Computing and Computer Assisted Intervention (MICCAI) challenge on Circuit Reconstruction from Electron Microscopy (CREMI)[4]. The TED is calculated at the image level, yet aims to capture topological errors, specifically splits and merges. Calculation of the TED requires solving an integer linear program (ILP), which selects the relabeling of one segmentation to minimize the number of splits and merges with respect to another segmentation. By selecting a reasonable tolerance threshold, the TED can ensure that "tolerable" errors, or those which don't affect the topology of the circuit, are ignored in the error calculation. One potential issue with the TED is that the proposed ILP may not be computationally tractable, though this often is not the case in practice. And while the TED's tolerance of segmentation errors is a desirable quality with regard to a metric that characterizes brain graph accuracy, the TED metric does not measure connectivity and thus cannot serve in this capacity independent of additional metrics.

2.3. Neural Reconstruction Integrity

2.3.1. Definition

We propose a new reconstruction metric called the Neural Reconstruction Integrity (NRI) metric. The NRI is a single neuron metric, which can be extended to a local network (a subset of neurons from the network, or a geometrically restricted region) or a global network metric. For a given ground truth neuron, we consider all synaptic terminals associated with the neuron. Presynaptic and postsynaptic terminals are treated independently—that is, only the presynaptic or postsynaptic "half" of a synapse is associated with a given neuron (except in the case of an autapse, in which both halves of the synapse would be associated with the same neuron). The NRI description below assumes that terminals in the reconstruction volume and the ground truth have already been matched. A proposed method for performing this matching is discussed in a subsequent section.

[4]https://cremi.org/

The NRI measures the extent to which *intracellular* paths between all possible pairings of ground truth synaptic terminals are preserved in the reconstruction. For a pair of terminals on a ground truth neuron, a true positive indicates those two synaptic terminals are both associated with a single neuron in the reconstructed volume—that is, an intracellular path is found between the terminals in the reconstruction. For instance, in **Figure 1**, post-synaptic terminals A″ and C″ are correctly associated with the same neuron of the reconstruction, which yields a true positive. However, B″ and C″ are not associated with the same neuron, yielding a false negative.

In graph theoretic terms, we find the set of edges (synapses) incident on a ground truth vertex (neuron), taking into consideration the directionality. For every pair of edges in this set, we check whether those edges are incident on the same vertex (neuron fragment) in the reconstruction, forming a true positive, or whether they are on different vertices, forming a false negative. We find false positives in pairs of edges that are incident on the same vertex but should not be. Note again that we do not require alignment of vertices between the ground truth graph and reconstructed graph since this may not even be possible due to splits and merges. Rather, we are interested in occasions that synapses are correctly associated on the same neuron. We consider this to be a key strength of the NRI metric, in that it can be interpreted as a measure of graph similarity based on edge clusterings, with no requirement for matching graphs via subjective pairing of ground truth neurons with reconstructed neurons.

The NRI is an f_1 score, which is the harmonic mean of precision and recall calculated on the true positive, false positive, and false negative paths as described above. For a given ground truth neuron, G_i,

$$NRI(G_i) = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

where precision and recall have the usual definitions involving true positive (TP) counts, false positive (FP) counts, and false negative (FN) counts, $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$. Notice that, using the definitions of precision and recall, the NRI can be rewritten as

$$NRI(G_i) = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3}$$

To obtain a local network or global NRI value, one calculates the total number of TPs, FPs, and FNs over the set of ground truth neurons under consideration and uses these values to calculate the f_1 score as usual.

Note that the global NRI value is strongly related to the line graph f_1 metric used in Gray Roncal et al. (2015). In some sense, the NRI can be viewed as an extension of the line graph f_1, which also counts TPs, FPs, and FNs of intracellular paths in a reconstruction. There are two key differences between the NRI and the line graph f_1 as defined and calculated in Gray Roncal et al. (2015). First, the NRI allows for evaluation at a variety of scales including single neurons, local networks, or global networks, allowing users to identify localized sources

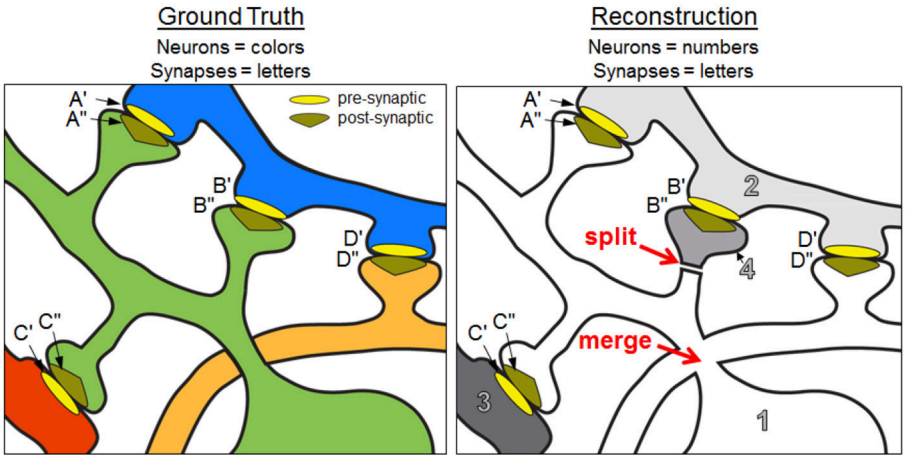

FIGURE 1 | Ground truth neurons and a reconstruction containing split and merge errors. Focusing on the green neural fragment, A″– C″ is a true positive path, B″– A″ and B″– C″ are false negative paths, and D″– A″ and D″– C″ are false positive paths. The NRI score of the green ground truth neuron is 0.333 (based on the neural fragments and synaptic terminals shown in the panels). See text for additional details.

of error within the overall reconstruction and achieve a snapshot performance of the entire network. The second key difference is that the NRI operates on directed graphs, or a reconstruction where synapses have direction. Accordingly, a neuron is penalized when one of its synapses is correctly identified in the reconstruction, but the direction is reversed—a penalty that would not arise in the line graph f_1. Despite these key differences we expect that, in many scenarios, the global NRI and the line graph f_1 would be highly correlated.

2.3.2. Examples

Consider **Figure 1** where a sample ground truth "neuron" (the green neuron) is reconstructed with a split error and a merge error. In particular, a spine head (neuron 4 in the reconstruction) is split from the dendritic shaft of the neuron so the post-synaptic terminal B″ no longer has an intracellular path to A″ or C″. This mistake yields two false negatives—one for the lost A″ to B″ path and one for the lost C″ to B″ path. Additionally, the orange neuron has been merged with the main body of the green neuron, resulting in new intracellular paths between D″ and the post-synaptic terminals A″ and C″. The merged neuron element is labeled as 1 in the reconstruction. This merge yields two false positives—one for the D″ to A″ path and one for the D″ to C″ path. The intracellular path between A″ and C″ is retained, resulting in one true positive. Using equation 2, we obtain an NRI score of 0.333.

The NRI is degraded when neuron split, neuron merge, synapse insertion, and synapse deletion errors occur. Synapse insertions increase the number of false positives while synapse deletions increase the number of false negatives. Additionally, if the synapse direction is reversed, the NRI decreases due to additional false positives *and* additional false negatives. For example, in **Figure 1**, if the presynaptic and postsynaptic terminals of synapse A were reversed so A′ was associated with neuron 1 and A″ was associated with neuron 2, then the NRI values of both the green and blue ground truth neurons decrease. With respect to the green neuron, not only is the intracellular path between C″ and A″ absent (false negative), but a new path between C″ and A′ is introduced (false positive).

2.3.3. Intuitive Scores

Here we highlight the intuitive relationship between reconstruction errors and the scores generated by the NRI metric. In each example scenario in **Table 1** it is assumed that all neurons have an equal number of synaptic terminals associated with them and that splits occur proportionately with regard to these terminals. We give global NRI scores (which are equal to single neuron scores in scenarios involving only one neuron) as well as precision (P) and recall (R). Note that because NRI is a scalar metric, its value does not indicate which types of reconstruction errors may have dominated in the event of a poor score. However, low precision scores are solely due to neuron merges and synapse insertions, whereas low recall scores are solely due to neuron splits and synapse deletions.

2.4. Implementation of NRI

Computation of the NRI requires three steps: (1) pairing synapses in the ground truth with those in the reconstruction based on proximity, (2) summing the total number of matching synapses for every possible pair of ground truth neuron and reconstruction neuron, and assembling these sums into a *count table*, and (3) using entries in the count table to determine the total number of true positive, false positive, and false negative pairs. Demonstration code is available at https://github.com/aplbrain/connectome_nri.

2.4.1. Synapse Alignment Using Centroids

The first step is to determine which synapse(s) in the reconstruction correspond to synapses in the ground truth by synapse assignment, for which we propose using the Hungarian-Munkres algorithm (Kuhn, 1955, 1956; Munkres, 1957). In

TABLE 1 | NRI scores and the precision and recall components for various scenarios.

Scenario	P	R	Global NRI
A neuron is split into two pieces with equal number of synapses	1.00	0.50	0.67
A neuron is split into three pieces with equal number of synapses	1.00	0.33	0.50
Two whole neurons are merged	0.50	1.00	0.67
Three whole neurons are merged	0.33	1.00	0.50
One neuron in a network of 10 neurons is split into 9 pieces and each piece is merged with one of the other 9 neurons	0.82	0.91	0.86
In a network of neurons, 20% of synapses on each neuron are deleted	1.00	0.64	0.78

Scores are intuitively commensurate with the magnitudes and types of reconstruction errors.

general, assignment can be handled in a variety of ways depending on the format of existing data such as synapse centroids or labeled voxels.

In the following we assume that the information necessary for computing NRI has been extracted and stored in two data files—one for the ground truth data and one for the reconstruction. Each file contains a list of synapses with associated neurons and locations. For a particular synapse the file contains an ID for the presynaptic neuron, an ID for the postsynaptic neuron, and an (x, y, z) coordinate representing the centroid of the synapse. There is no guarantee, and in fact it is unlikely, that the IDs or (x, y, z) coordinates will correspond perfectly between the two lists due to reconstruction errors. By applying the Hungarian-Munkres algorithm to synapse centroids, we reconcile the difference in synapse identifiers. Note that it is not necessary to perform any neuron alignment, or any explicit pairing of ground truth neurons and reconstructed neurons.

Assigning synapses in the reconstruction to those in the ground truth can be nuanced, particularly if we consider volumetric synapse representations (labeled voxels). For example, if the voxels of a reconstructed synapse overlap with half of those of a ground truth synapse, and also overlap with an equal number of voxels outside of the ground truth synapse, it is somewhat subjective as to whether or not the reconstructed synapse should be assigned to the ground truth synapse. However, the aim of the NRI metric is to measure characteristics important for representing brain graph connectivity rather than specific voxels or detailed synapse morphology. Thus, we propose the use of synapse centroids, which eliminates judgment calls based on the amount of voxel overlap. To allow for unassigned synapses (accommodating erroneous synapse deletions or insertions in the reconstruction), the Hungarian-Munkres algorithm can be modified to prevent assignment when distance between centroids is unrealistically high (e.g., $> 300\,nm$).

2.4.2. Count Table Calculation

Once synapse assignment is complete, it is possible to generate the count table (a matrix). In the count table, each row corresponds to a ground truth neuron and each column corresponds to a reconstructed neuron. An entry in the table, c_{ij}, corresponds to the number of *matched synaptic terminals* between ground truth neuron G_i and reconstructed neuron segment S_j. Matching synapse terminals are those for which both (1) the reconstruction synapse of neuron S_j has been assigned to the ground truth synapse of neuron G_i, and (2) the polarity of the terminals are the same (presynaptic or postsynaptic). Thus, if a terminal is presynaptic on G_i in the ground truth and postsynaptic on S_j in the reconstruction, then G_i and S_j do not share that terminal even though the synapses are assigned to each other. Note that if N reconstruction synapses are assigned to ground truth synapses, then there will be a total of $2N$ matching synaptic terminals in the count table (excluding those of the insertion row and deletion column—see below). This applies for synaptic junctions with one pre-synaptic and one post-synaptic process, which is the case for the vast majority of known connections in mammalian cortex, but not for organisms such as *drosophila*. Polysynaptic junctions will generate additional count table entries.

TABLE 2 | The count table for the ground truth and reconstruction depicted in **Figure 1**.

	del	**1**	**2**	**3**	**4**
ins	0	0	0	0	0
Green	0	2	0	0	1
Red	0	0	0	1	0
Blue	0	0	3	0	0
Orange	0	1	0	0	0

The del column and ins row are for counts of deleted and inserted synaptic terminals, respectively.

The count table corresponding to **Figure 1** is shown in **Table 2**. Examination of the count table immediately reveals useful information. For instance, the "green neuron" was split into two elements in the reconstruction while "neuron 1" of the reconstruction is a merge of two ground truth neurons.

Additionally, the count table has a row corresponding to inserted synapses (*ins*), or those found in the reconstruction and not the ground truth. It also contains a column for deleted synapses (*del*), or those found in the ground truth and not the reconstruction.

2.4.3. Calculating NRI From the Count Table

Once the count table is established, it is possible to calculate the NRI.

Let C be the count table for a local network of the ground truth brain graph and the associated portions of the reconstruction.

The 0*th* row refers to synapse/terminal insertions and 0^{th} column refers to synapse/terminal deletions while all other rows and columns indicate ground truth and reconstruction neurons, respectively. There are I total ground truth neurons and J total corresponding reconstructed neurons (those that share at least one synapse with at least one ground truth neuron). Neurons (or other objects such as glia) that share no synapse correspondences are ignored when computing NRI, as they do not impact our graph. If c_{ij} denotes the i,j-entry of the count table, then the total number of true positives, false negatives, and false positives across the volume can be computed using the equations below.
True positives:

$$TP = \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{c_{ij}}{2} \tag{4}$$

Note that the outer summation is over the ground truth neuron index, i, thus the number of true positives for a single ground truth neuron is simply the inner summation over j for a given i.
False negatives:

$$FN = \sum_{i=1}^{I} \left[\binom{c_{i0}}{2} + \sum_{j=0}^{J-1} \sum_{j'=j+1}^{J} c_{ij} c_{ij'} \right] \tag{5}$$

Notice that the false negative total includes contributions from the synapses in the deletions column (column 0) in two forms—once with all synapses matched to those in the *ground truth* neuron and again by pairing all possible combinations in the deleted column. This ensures that the sum of the true positives and false negatives is equal to the total number of synapse pairs on the ground truth neuron. As for true positives, the number of false positives for a single ground truth neuron is simply the value of the term inside the outer summation, for a given neuron i.
False positives:

$$FP = \sum_{j=1}^{J} \left[\binom{c_{0j}}{2} + \sum_{i=0}^{I-1} \sum_{i'=i+1}^{I} c_{ij} c_{i'j} \right] \tag{6}$$

Computation of the total number of false positives is essentially identical to that for the false negative total, except computed in the other direction across the count table (effectively, computed on the transpose of the count table). Contributions from the insertions row (row 0) play a similar role to those from the deletions column under the false negatives computation—being counted for incorrect pairing once with all synapses matches in the *reconstructed* neuron and counted again for incorrect pairing in all possible combinations with each other.

Determining the number of false positives for a single ground truth neuron is open to interpretation, as there is ambiguity with regard to false positives that arise due to synapses being inserted on merged neurons. In addition, if two neurons are merged, the false positives created by the pairing of their synapses should be distributed between the neurons. In the latter case, we chose to attribute half the false positives to one neuron, and half to the other. Regarding insertions, false positives due to pairs of inserted synapses are not attributed to a ground truth neuron (although false positives between an insertion and synapses found on a ground truth neuron *are* attributed to that neuron) but they are added to the total count of network false positives. Thus,

$$FP = \sum_{i=0}^{I} FP(i) \tag{7}$$

where FP is the total count of network false positives, $FP(i)$ is the number of false positives attributed to individual ground truth neurons and (for $i = 0$) those due to pairs of inserted synapses, and

$$FP(i) = \begin{cases} \sum_{j=1}^{J} \binom{c_{0j}}{2}, & \text{if } i = 0 \\ \sum_{j=1}^{J} c_{ij} c_{0j} + \frac{1}{2} \sum_{j=1}^{J} \sum_{\substack{i'=1 \\ i' \neq i}}^{I} c_{ij} c_{i'j}, & \text{otherwise} \end{cases} \tag{8}$$

Once the total number of true positives, false positives, and false negatives have been tallied (for individual neurons or for the entire network), the final step is to use the calculated values in equation 3 for a local network NRI value.

As a concrete example, consider **Figure 1** and the corresponding count table in **Table 2**. The number of true positives for the green ground truth neuron is $\binom{2}{2} + \binom{1}{2} = 1$, or the number of *pairs* of green neuron terminals that are also found in the reconstruction[5]. This is calculated by examining the row in **Table 2** corresponding to the green neuron. The number of false negatives for the green neuron is $2 \cdot 1 = 2$, or the number of pairs of terminals incorrectly split across neuron fragments in the reconstruction, also calculated by examining the green neuron row. Finally, a false positive count may be obtained by looking at any given column. For instance, the number of false positives associated with the green ground truth neuron is $(2 \cdot 3) + (2 \cdot 1) = 8$, which is then divided by two to prevent false positives from being double counted when they are summed over the entire network.

2.5. Adapted Alternative Metrics

Although the NRI metric operates on matched synaptic terminals rather than voxels, it is otherwise closely related to the Rand Index in that it utilizes TP, FP, and FN values to compute a final score. (The number of True Negatives (TN) is used in computing the Rand Index, but not the NRI—a distinction we discuss further in the Results section.) Similarly, one can conceive of an adapted version of VI that is computed from the same count table as that utilized by the NRI. We define terminal-based

[5]Where $\binom{n}{2}$ indicates n-choose-2, or the number of all possible pairs of elements from a set of n elements.

adaptations of these alternative metrics for comparison with the NRI.

$$TP = \sum_{i=0}^{I}\sum_{j=0}^{J}\binom{c_{ij}}{2} \tag{9}$$

$$FN = \sum_{i=0}^{I}\sum_{j=0}^{J-1}\sum_{j'=j+1}^{J} c_{ij}c_{ij'} \tag{10}$$

$$FP = \sum_{j=0}^{J}\sum_{i=0}^{I-1}\sum_{i'=i+1}^{I} c_{ij}c_{i'j} \tag{11}$$

$$TN = \sum_{i=0}^{I-1}\sum_{j=0}^{J-1}\sum_{i'=i+1}^{I}\sum_{j'=j+1}^{J} c_{ij}c_{i'j'} \tag{12}$$

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

2.5.1. Normalized Variation of Information

As described above for the adapted Rand index, when computing the adapted VI, the insertion row and deletion column of the count table are simply treated as if they are additional neurons in the ground truth and reconstruction networks, respectively. In addition, the VI score is normalized by $H(G,S)$ to provide a normalized VI (NVI) score that ranges from 0 to 1. From the count table, the NVI is computed as follows.

$$p_{ij}^{(g,s)} = c_{ij} / \sum_{i'=0}^{I}\sum_{j'=0}^{J} c_{i'j'} \quad \text{for all } i,j \tag{14}$$

$$p_i^{(g)} = \sum_{j=0}^{J} p_{ij}^{(g,s)} \tag{15}$$

$$p_j^{(s)} = \sum_{i=0}^{I} p_{ij}^{(g,s)} \tag{16}$$

$$H(G|S) = -\sum_{i=0}^{I}\sum_{j=0}^{J} p_{ij}^{(g,s)} \log \frac{p_{ij}^{(g,s)}}{p_j^{(s)}} \tag{17}$$

$$H(S|G) = -\sum_{i=0}^{I}\sum_{j=0}^{J} p_{ij}^{(g,s)} \log \frac{p_{ij}^{(g,s)}}{p_i^{(g)}} \tag{18}$$

$$H(G,S) = -\sum_{i=0}^{I}\sum_{j=0}^{J} p_{ij}^{(g,s)} \log(p_{ij}^{(g,s)}) \tag{19}$$

$$NVI = \frac{H(G|S) + H(S|G)}{H(G,S)} \tag{20}$$

where $p_{ij}^{(g,s)}$ is the joint probability of a matched terminal being found on the ith ground truth neuron and the jth reconstruction neuron, and $p_i^{(g)}$ and $p_j^{(s)}$ are marginal matched terminal distributions for the ground truth and reconstruction neurons, respectively.

2.6. Simulated Data

To test the NRI metric behavior we would ideally apply it to a large 3D volume for which ground truth data existed, as well as semi-automated reconstructions generated over a range of methods and parameters. When compared to the volume of raw data currently being collected, most available ground truth datasets tend to be small (hundreds of neurons) and sparse (few connections between neurons), and composed primarily of small fragments of neurons rather than large fragments or whole neurons (Takemura et al., 2013; Lee et al., 2016). We therefore chose to synthesize a neural network with modestly realistic anatomical properties, and introduce errors into the network ("perturb" the network) to simulate reconstruction errors (resulting in imperfect reconstructions). This approach also allowed us to independently examine the effect of individual types of errors on the NRI scores, at graded perturbation levels.

To generate cortical networks with large numbers of neurons, we turned to NeuGen 2.0, a product developed at the University of Heidelberg, for generation of neurons and neural networks (Eberhard et al., 2006). NeuGen is an open source Java program that synthesizes neurons by using a probabilistic model of the growth of neuronal processes—e.g., turning and branching. Processes are composed of numerous short, cylindrical segments. Synapse generation is based on Peter's Rule (distance between processes), modified to prevent synapse clustering (excessively dense synapse formation in localized process regions). Neurons were modeled after those in the rodent somatosensory barrel cortex as specified by the default NeuGen parameters. Our synthesized networks consisted of 872 complete neurons (312 L2/3 pyramidal neurons, 62 L4 stellate neurons, 62 L4 star pyramidal neurons, 218 L5A pyramidal neurons, and 218 L5B pyramidal neurons) and over one million synapses—approximately 2,320 synaptic terminals per neuron, with somata confined in a volume of $x = y = 79\mu$m and $z = 1,300\,\mu$m. Computational memory and processing limitations prevented us from generating a denser network. Although neuron density of the synthesized network is only about 1/10th that of real cortical tissue, we consider the networks to be sufficiently large and complex to serve as a proxy for real data in testing of the NRI metric. We generated five such networks, using different seed values for the underlying random number generators.

Current reconstruction methods generally introduce four types of reconstruction errors, with the error rates for each type often traded-off based on choice of algorithm parameters. For example, synapse detection algorithms often have a tradeoff between synapse precision and recall, leading to added and/or deleted synapses in the final reconstruction. Neuron segmentation algorithms may fail to differentiate membrane boundaries in poor quality images, resulting in merged neurons. Yet if parameters are tuned to minimize false merges, the algorithm may identify nonexistent boundaries at thin portions of a neuron resulting in a neuron split (e.g., splitting of dendritic spines from the shaft). To simulate the introduction of these errors into a reconstruction we built basic perturbation models for the generation of each type of error. Models are summarized in **Table 3**.

TABLE 3 | Descriptions of perturbation models used to produce imperfect graph reconstructions from a synthesized ground truth network.

Error type	Perturbation model description
Synapse deletion	A specified percentage of synapses is randomly selected from the set of all existing synapses and deleted.
Synapse insertion	For each possible pair of cylindrical process segments (from different neurons), insert a synapse with probability p where p is p_{max} for inter-process distance less than d_1, p is 0 for distance greater than d_2, and p follows a linear decreasing curve in (d_1, d_2).
Neuron split	For each cylindrical process segment, split the neuron at the segment with probability p where p is p_{max} for process diameter less than d_1, p is 0 for diameter greater than d_2, and p follows a linear decreasing curve in (d_1, d_2).
Neuron merge	For each possible pair of cylindrical process segments (from different neurons), merge the neurons at the segments with probability p where p is p_{max} for inter-process distance less than d_1, p is 0 for distance greater than d_2, and p follows a linear decreasing curve in (d_1, d_2).

It is possible to run each perturbation model sequentially to generate all types of errors in a single reconstruction. However, in the following analysis, we generated reconstructions with only one type of error in each reconstruction, as this allowed direct observation of how the type of error affects neuron and network NRI scores.

2.7. Real Data

Due to the limited size of most real network reconstructions with high-quality annotations, we were motivated to use synthetically created networks for metric testing. Nevertheless, application of the NRI metric to a small, real data set might provide confirmation of testing results on synthetic data, and/or expose conditions and outcomes not revealed by the synthetic data set. Therefore, we additionally assessed the metric by applying the perturbation model to a real network of 201 neurons in mouse visual cortex, taken from manually annotated EM data (Lee et al., 2016). This "core" network has no leaf nodes—that is, each neuron is connected to two or more other neurons.

3. RESULTS

3.1. Applying Metrics to Simulated Data

In this section, we empirically demonstrate relationships between error types and local (single neuron) NRI scores and provide explanations of why these relationships exist. The results in this section indicate that the NRI metric is well-behaved, scalable, and amenable to interpretation. For each error type—synapse deletion, synapse insertion, neuron split, and neuron merge—the perturbation model is applied to the ground truth networks described in Section 2.6 with several different perturbation parameter sets, intended to create imperfect reconstructed graphs of decreasing accuracy (at the network level). For example, in the case of synapse deletion, the percentage of synapses that are randomly deleted from the ground truth network is increased across individual simulations, resulting in reconstructed networks with different levels of synapse degradation. Given a ground truth network and an imperfectly reconstructed network, the global NRI is calculated for the entire reconstructed network and the local NRI is calculated for each ground truth neuron. Across the error types, we expect greater perturbation to lead to smaller NRI values. This is the case for both local NRI (although scores vary from neuron to neuron) and global NRI.

We additionally use the adapted versions of VI and Rand Index to compute global scores for the reconstructed networks. These alternative metrics do not have defined local scores, and are thus not compared with NRI at the scale of individual neurons.

3.1.1. NRI Scores for Synapse Deletions and Insertions

First, we consider synapse deletions. As described in **Table 3**, a fixed percentage of synapses are randomly chosen from across the entire volume and deleted. Thus, most ground truth neurons will be impacted roughly to the same degree (with some variance about a mean). When a single synapse is deleted, the number of true positives decreases and an equal number of false negatives is introduced. The result is a lower recall score and a lower local NRI score. The effect of decreased TPs and increased FNs is readily seen by studying Equation (3). A synapse deletion only impacts the local NRI scores of the ground truth neurons with which the synapse is associated (presynaptic and postsynaptic). The NRI decreases more for ground truth neurons that lose more synapses (as a fraction of total number of synapses associated with those neurons). This is evident in **Figure 2A** where the local NRI score is smaller for ground truth neurons that lose a greater fraction of their overall synapses. Additionally, **Figure 2A** shows that the network level or global NRI score also suffers when deletion rate is high. For example, the dark blue markers represent individual neurons from a single reconstruction in which the deletion rate was high. Both the network and neuron NRI scores are low in this case.

Next, we consider synapse insertions. Under the perturbation model, synapses are inserted probabilistically based on the distance between neuron membranes (more precisely, the distance between the cylindrical segments of which the neuronal processes are composed). Naturally, some neurons will be significantly more impacted by this error model than others. When a single synapse is inserted, several false positives are introduced where the number of false positives depends on how many synapses are associated with the original ground truth neuron. False positives decrease the precision term and thus the total (local or global) NRI value. Again, a synapse insertion effects the local NRI values of only the two neurons on which the synapse is incident (presynaptic and postsynaptic). One measure of the extent to which a ground truth neuron has been impacted by insertions is the fraction of the reconstructed neuron's synapses that are not associated with those of the ground truth neuron. This is the perturbation metric used in **Figure 2D**. Neurons that experience a larger number of synapse insertions have lower NRI values, as seen in the figure. Notice that, because

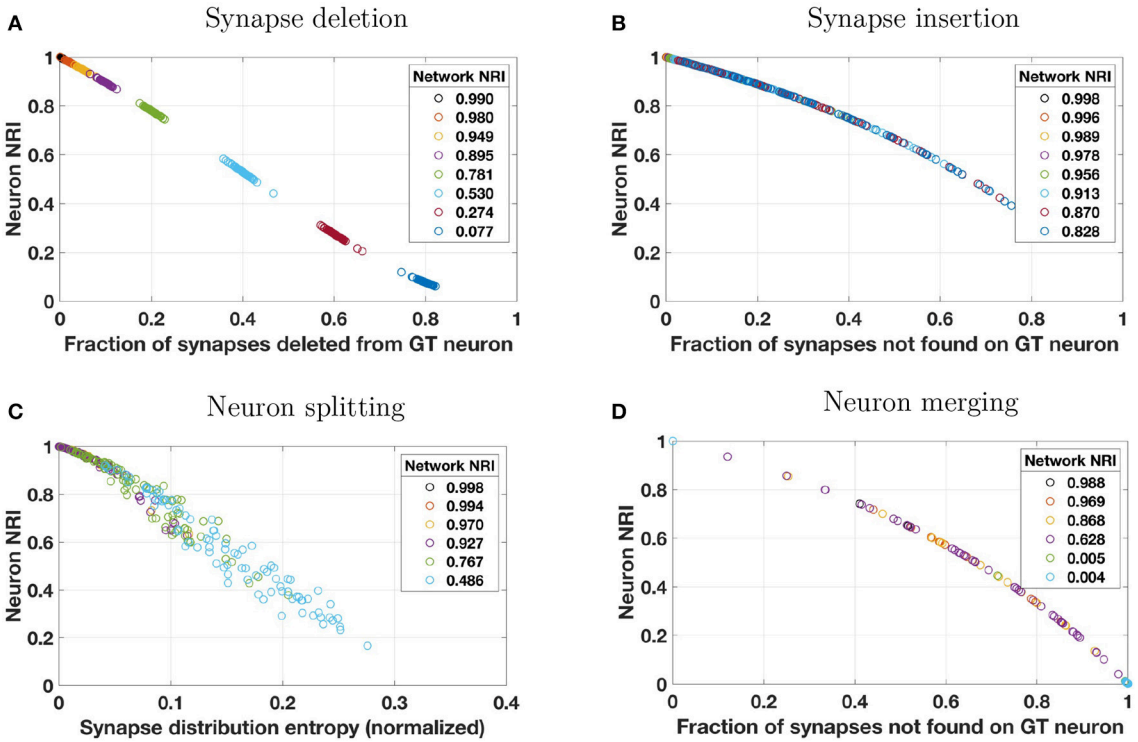

FIGURE 2 | Reconstruction errors were simulated by applying one of four perturbation models to a synthetically generated ground truth (GT) network. Perturbation models (see **Table 3**) introduced errors by **(A)** deleting a percentage of synapses from the overall network, **(B)** probabilistically inserting synapses where two neuron membranes are closely apposed, **(C)** probabilistically splitting neurons where process diameter is small, or **(D)** probabilistically merging neurons where two neuron membranes are closely apposed. These plots show how local NRI scores of individual neurons vary as a result of the introduction of these errors. Several perturbation metrics were used to compare perturbation magnitude to NRI scores. For synapse deletions, neuron NRI scores are compared to the fraction of synapses that were deleted from a given GT neuron. For synapse insertions and neuron merges, NRI is compared to the fraction of synapses *not* found on the GT neuron. In the case of neuron merges, this means that if neuron A is merged with neurons B and C in the reconstruction, then the perturbation score for neuron A is $\frac{n_B+n_C}{n_A+n_B+n_C}$ where n_A, n_B, and n_C are the number of synapses associated with neurons A, B, and C, respectively. For neuron splits, neuron NRI is compared to the entropy of the synapse distribution across the split pieces of the GT neuron (normalized by the total number of synapses). The color of each neuron's data point indicates the global network in which the neuron resided, and the NRI score for that global network is indicated in the plot's legend. For example, for synapse deletions in plot A, the data points colored dark blue at the bottom right of the plot are neurons from a single perturbed network whose network NRI score is 0.077. Individual neuron NRI scores are close to the network NRI score in this particular case. To enhance visibility, results from only 10% of the neurons (uniformly chosen across neuron size) are plotted. Overlap of markers of different colors is indicative of a broad range of neuron NRI scores for a single reconstructed network.

this perturbation model will greatly impact a handful of neurons and leave others virtually untouched (due to the fact that the probability of insertion depends on the density of processes in the synthetic network, which is higher at the center of the volume and lower at the edges), **Figure 2B** does not show the same separation between reconstructed networks as **Figure 2A** does. Global NRI values are not as heavily impacted and every reconstructed network has some neurons with low deletion and high NRI.

3.1.2. NRI Scores for Neuron Splits and Merges

Segmentation errors made during reconstruction can result in neuron splits and neuron merges. First, we consider neuron splits, which are made probabilistically based on process diameter (see **Table 3**). As with synapse insertions, the probabilistic model used will result in some neurons that are greatly affected by multiple splits and other neurons that are rarely or never split. A single neuron split, say into pieces A and B, will introduce several false negatives between all pairs of synapses where one synapse is associated with piece A and the other synapse is associated with piece B. Such an error only effects the NRI of the split neuron and the effect is immediately seen through inspection of Equation (3). **Figure 2C** shows that greater splitting results in lower local NRI value. Because neurons in a network are not uniformly impacted, there is no clear local NRI separation between neurons from low perturbation networks and those from high perturbation networks.

Finally, we consider neuron merges, which are made probabilistically when two neurons (processes) fall within a certain distance of each other. Notice that, when this model is applied, whole neurons are merged together whenever a merge is indicated. Thus, each ground truth neuron is a subset of a reconstructed neuron. As for synapse insertions, we measure the extent to which a ground truth neuron has been impacted by merges as the fraction of the reconstructed neuron's synapses that are not associated with those of the ground truth neuron.

This is the perturbation metric used in **Figure 2D**. Once again, the nature of the neuron merge model is that some neurons may be involved in several merges and others may be involved in a small number, possibly none. Thus there is no clear separation in the NRI scores of high perturbation network neurons and low perturbation network neurons. Merging two ground truth neurons, say *A* and *B*, into one reconstructed neuron introduces a false positive for each synapse-synapse pair where one synapse is associated with neuron *A* and the other is associated with neuron *B* in the ground truth data. The effect of additional false positives can readily be seen upon examination of Equation (3). **Figure 2D** verifies that ground truth neurons subject to a great deal of merging also tend to have small local NRI scores.

3.1.3. NRI Neuron Score Distributions

The underlying construction of the NRI metric indicates that errors impacting larger neurons could have an outsized impact on the global NRI score, since the number of synapse *pairs* associated with a neuron is approximately proportional to the square of the number of synapses. However, although anecdotal and specific to our simulations, we did not observe a dominating impact of large neurons on the global NRI scores (**Figure S1**). For modestly perturbed networks (global NRI near 0.9), the average of a network's local NRI scores was found to be within about 3% of global NRI scores, suggesting a relatively balanced contribution from individual neurons.

3.1.4. Alternative Network Scores

Network scores based on the NRI, NVI, and Rand Index metrics are shown in **Figure 3**. While each metric provides scores that trend lower with an increasing number of reconstruction errors, there are distinct differences between the metrics. Most notably, the Rand Index gives higher scores than the other two metrics, exhibiting less sensitivity to errors. This is particularly evident for splitting errors, for which the Rand Index gives scores nearly equal to 1.0, even when splitting errors are extensive. This is due to the inclusion of TNs in the computation of the Rand Index, and the effect has been noted by others. Researchers have noted (Jain et al., 2011) that the Rand Index applied to voxels creates a "classification task [that] is highly imbalanced, because the vast majority of voxel pairs belong to different ground truth clusters. Hence even a completely trivial segmentation in which every voxel is its own cluster can achieve fairly low Rand error." Similarly for the terminal-based Rand Index, even if each terminal was assigned to a unique neuron (maximum split errors), the number of FNs will be dwarfed by the number of TNs and thus the Rand Index score will remain relatively high. The NVI provides network scores closer to those of the NRI. Like the Rand Index, the NVI shows lower sensitivity to errors (except for synaptic insertion errors, to which it is more sensitive), however this lack of sensitivity is not as drastic as for the Rand Index. Thus, depending on the desired metric sensitivity, the NVI may be a suitable metric for measuring *network* reconstruction accuracy.

3.2. Applying the NRI Metric to Real Data

For limited testing on real data, the perturbation models were applied to a manually-annotated network of 201 neurons from the mouse visual cortex (Lee et al., 2016) and subsequently scored by the NRI metric. Because our ground truth data does not include information on diameters of neuronal processes, the splitting perturbation model was not applied, as it requires this information to determine split probability.

Results shown in **Figure 4** are in accord with those of the synthetic data in that single neuron NRI scores largely lie along a dominant trend curve that is monotonic. Unlike the synthetic data, however, some scores clearly deviate from the dominant trend. This occurs when there are very few synapses on the neuron or pair of neurons at which a reconstruction error is made. For example, under the synaptic deletion model (**Figure 4A**), there are numerous neurons for which the fraction of synapses deleted from the ground truth neuron is 0.5. While most have an NRI score at or slightly below 0.4, there are some that are notably lower, including the extreme case of an NRI score of 0. In this extreme case, the neuron had only two synapses, one of which was deleted. Thus the neuron has zero true positives and one false positive, resulting in an NRI score of 0 based on Equation (3). In general, the real neurons may have an NRI score that is the ratio of two very small numbers and thus may deviate from the dominant trend. In contrast, neurons in the synthetically generated networks have many more synapses per neuron, so even when there are many reconstruction errors, a neuron NRI score is never a ratio of two small numbers. It should be noted that the "real" neurons do not really have such a small number of synapses—rather, only a small number of synapses were annotated.

4. DISCUSSION

4.1. Results

Results from simulations utilizing both synthetic and real data indicate that the NRI has several of the desired qualities of a metric for assessing reconstructions with regard to the brain graph accuracy. For individual types of reconstruction errors, scores are intuitively commensurate with the magnitude of errors, with scores ranging from 0 to 1. Although not shown directly in the simulations (but see **Table 1**), when applied to reconstructions that contain multiple types of errors, observation of the precision and recall components of the NRI score lend additional insight into the types of errors contained in the reconstruction. Finally, NRI computation was performed on a modern personal computer with run times on the order of seconds. Although the synthetic data sets were of modest size compared to that expected of real data sets in coming years, NRI computation on larger data sets will be feasible by utilizing the methods outlined in section 2.4 for synapse matching, and by leveraging more powerful computing hardware.

We briefly address two concerns with the NRI metric. First, when synaptic insertion errors are present in the reconstruction, the global FP count cannot be broken down into a set of factors comprised solely of individual neuron FP counts, due to the FP(0) term in Equation (8). Nonetheless, the vast majority of FPs can be assigned to individual neurons, and the FP(i) terms will sum to the global FP asymptotically as the number of insertion errors goes to zero.

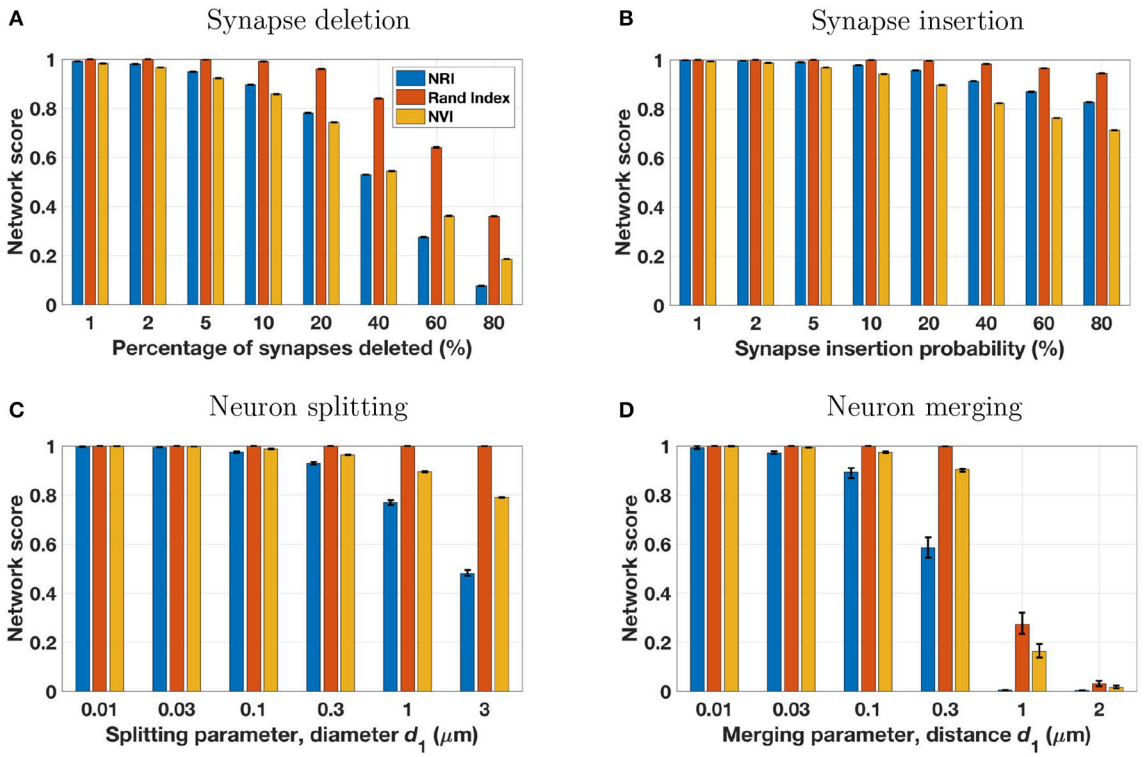

FIGURE 3 | The four perturbation models (see **Table 3** and **Figure 2**) were applied to five different ground truth networks, using a range of model parameters to adjust the number of reconstruction errors introduced. Each set of three bars in the above panels represents a single perturbation model and set of model parameters, applied to the five ground truth networks with results scored by three metrics. Bar height is the mean score of the five reconstructions for the specified metric and model parameter, with error bars extending from the lowest score to the highest score. In many cases the error range is very tight, owing to the similarity of the five ground truth networks and the uniform and consistent manner in which perturbation models introduce simulated reconstruction errors.

Second, because the metric considers all possible pairs of synaptic terminals except true negatives, one might approximate the effective weighting of single neurons in the global score as proportional to the square of the number of synapses on the neurons (a proxy for neuron size). Subjectively, we did not see evidence of this in results from the synthetic data (**Figure S1**). Regardless, whether or not such weighting would be problematic depends partially on the goal of scoring with the NRI metric. From the standpoint of a researcher analyzing neural connectivity patterns or inferring brain function based on the graph, it may be justified to give substantially greater weighting to larger neurons when assessing reconstruction accuracy. Additionally, the issue may be moot for some real reconstructions, as most if not all of the neurons have a relatively equal number of synapses (e.g., 5,000–10,000 in cortex) and thus weighting will not vary substantively across neurons.

4.2. Ground Truth Data

We discuss here some aspects of real ground truth data that should be considered when applying the NRI metric. Obtaining ground truth data through the manual sampling (annotating) of an image volume typically takes one of two forms—densely annotating a geometrically confined region (e.g., a small cube within the larger volume) or sparsely annotating large portions of a few neurons and their processes, perhaps along with a subset of their synaptic partners. In either case, we must remain aware that there is vastly more information in a large semi-automated reconstruction than in the ground truth data, and some aspects of the reconstruction may in fact be a more accurate depiction of the real brain graph than that depicted by the ground truth data.

As a specific example, consider a branching process for which ground truth data exists for a pair of branches but not for the branching point (i.e., the branching point is outside of the manually annotated region). In this case, the ground truth data would label these processes as unique neuron fragments. However, if the larger reconstruction data captures the branching point, the two branches as well as the branching point would be correctly labeled as a unique neural fragment. If the NRI were computed on these data naively, the reconstruction would be unjustly penalized with many false positives since from the perspective of the ground truth data, the two branches were erroneously merged. Thus, a preprocessing step is needed in which the reconstruction is cropped to match the confined region of the ground truth data, and neuron fragments are relabeled based on connected components (i.e., generating two new identifiers for branches that do not have adjacent voxels in the cropped volume) such that cropped reconstruction labeling is equivalent to that which would have been obtained had the entire

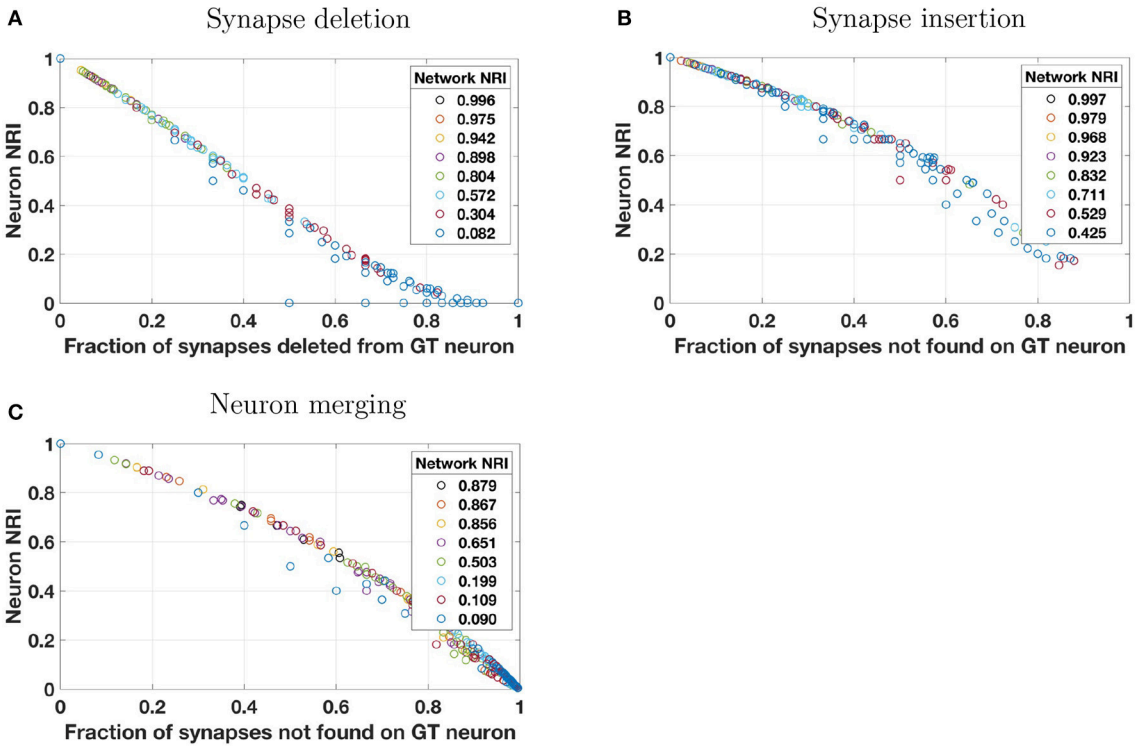

FIGURE 4 | Simulated reconstruction errors were introduced into a real network of 201 neurons in mouse visual cortex, taken from manually annotated EM data (Lee et al., 2016). This "core" network has no leaf nodes—that is, each neuron is connected to two or more other neurons. Perturbation models and plotting style are identical to that of **Figure 2** except that the splitting model was not applied because neuron process diameter (which is needed by the splitting model to determine split probability) is not present in the ground truth annotations. Neuron scores that deviate from the predominantly monotonic trend are those of neurons with few synapses, as described in the text in section 3.2.

reconstruction been composed only of the confined ground truth region.

An additional problem arises when sparsely annotated ground truth data is used. In that case it is more likely that manual annotation errors will arise in the form of dendritic spine splits and associated orphaned synapses on spine heads, because all pixels are not assigned and so small details are more easily missed. As mentioned in the introduction, ground truth should actually be treated as "gold standard" data, that, despite being used for assessing reconstruction quality, may itself have some errors. One mitigating approach to the aforementioned problem is to revise the manner in which ground truth data is collected. For example, all synapses in the volume could first be annotated, and then traced back to a dendritic shaft, thereby reducing the likelihood of missing synapses. Or as a compromise, the same approach could be taken but synapses would be annotated only within a fixed diameter range about a ground truth dendritic process, with the assumption that synapses outside this range could not belong to the dendrite. Finally, a modification to the NRI metric would make it insensitive to such errors, as described below.

4.3. Future Extensions

In this manuscript, we defined an NRI operating point as the harmonic mean of precision and recall (e.g., f_1). For graph inference tasks, it might be more favorable to choose a different β value in f_{beta}, which has the effect of weighting the contribution of false positive and false negative paths asymmetrically. Another extension would be to consider different methods of computing a global NRI score, such as weighting each neuron's contributions equally rather than weighted by the number of paths. Many (brain)-graphs are produced without polarity information; NRI can be easily extended to undirected paths if desired.

4.4. A Modified, Segmentation-Only NRI

Rigorous procedures are necessary to ensure that synapses are not missed when manually generating *sparse* ground truth annotations (e.g., missed detection of dendritic spine shafts results in a missed synapse). One approach to relaxing manual annotation accuracy requirements in this regard is to use a segmentation-only version of the NRI in conjunction with other metrics. If the NRI is computed using only matched synapses (that is, unpaired synapses representing synapse deletions and synapse insertions are not included in the count table) then missed synapse errors in the "ground truth" annotation will not result in unjust penalization of reconstructions that do not make these errors.

While this might appear to result in a metric that is insensitive to some errors in the reconstruction, this is only true if the associated synapses are deleted from the reconstruction as well. In reality, if the modified NRI is coupled with a synapse detection

metric [as with the TED metric (Funke et al., 2017) in the 2016 MICCAI CREMI challenge[6]] and the score of the synapse detection metric is high, then segmentation quality will still be an important component of the NRI score.

5. CONCLUSION

We present an NRI metric for assessment of a reconstructed volume of neural tissue that emphasizes network connectivity. Our results indicate that the metric serves this purpose well based on several desirable qualities including applicability to both dense and sparsely annotated ground truth volumes, and applicability to single neurons, local regions, and global networks. Additionally the metric produces an interpretable score that falls within [0, 1] and is computationally feasible even at scales much larger than that of currently available data sets. We highlight NRI in the context of high-resolution brain graphs, but this metric applies broadly to graphs estimated using a variety of methods and at a variety of scales. Indeed, it is potentially relevant for other problem domains where path finding is a critical objective (e.g., road detection, autonomy).

The metric has yet to be tested on a large volume of real ground truth data with a real reconstruction pipeline. In addition to confirming the utility of the metric, such an effort is likely to help refine strategies for manually annotating ground truth data and may ultimately facilitate researchers' efforts toward creating automated or semi-automated reconstruction methods leading to high quality, large scale brain graphs.

AUTHOR CONTRIBUTIONS

ER conceived of the NRI. ER and MR formalized the NRI and developed the equations for its calculation. WG, DK, BW, and MC contributed technical insight on NRI utility, and on assessment of NRI as a metric. ER, JG, and MR wrote code to implement the NRI and analyze NRI on simulated data. JG and MR generated synthetic data and simulated errors for analysis of NRI. ER and MR wrote the manuscript and ran the final experiments, with inputs from all authors.

[6]https://cremi.org/

FUNDING

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17032700004-005 under the MICrONS program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation therein.

ACKNOWLEDGMENTS

The authors wish to thank Stephen Plaza, Carey Priebe, Joshua Vogelstein, and Vince Lyzinski for discussions about brain graph reconstruction evaluation. Stephen Plaza, in particular, provided valuable insight into the processes by which EM-derived connectomes are generated, and the practical realities in developing and utilizing metrics for reconstruction evaluation. We also thank Wei-Chung Lee and Clay Reid for the use of their manually annotated reconstruction of mouse cortex.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2018.00074/full#supplementary-material

REFERENCES

Arganda-Carreras, I., Turaga, S. C., Berger, D. R., Cireşan, D., Giusti, A., Gambardella, L. M., et al. (2015). Crowdsourcing the creation of image segmentation algorithms for connectomics. *Front Neuroanatomy* 9:142. doi: 10.3389/fnana.2015.00142

Eberhard, J. P., Wanner, A., and Wittum, G. (2006). Neugen: a tool for the generation of realistic morphology of cortical neurons and neural networks in 3d. *Neurocomputing* 70, 327–342. doi: 10.1016/j.neucom.2006.01.028

Funke, J., Andres, B., Hamprecht, F. A., Cardona, A., and Cook, M. (2012). "Efficient automatic 3D-reconstruction of branching neurons from EM data" in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI: IEEE), 1004–1011.

Funke, J., Klein, J., Moreno-Noguer, F., Cardona, A., and Cook, M. (2017). Ted: a tolerant edit distance for segmentation evaluation. *Methods* 115, 119–127. doi: 10.1016/j.ymeth.2016.12.013

Funke, J., Tschopp, F. D., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., et al. (2018). Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Trans Patt. Anal. Mach. Intell.* doi: 10.1109/TPAMI.2018.2835450. [Epub ahead of print].

Gray Roncal, W., Kleissas, D. M., Vogelstein, J. T., Manavalan, P., Lillaney, K., Pekala, M., et al. (2015). An automated images-to-graphs framework for high resolution connectomics. *Front. Neuroinform.* 9:20. doi: 10.3389/fninf.2015.00020

Helmstaedter, M., Briggman, K. L., and Denk, W. (2011). High-accuracy neurite reconstruction for high-throughput neuroanatomy. *Nat. Neurosci.* 14, 1081–1088. doi: 10.1038/nn.2868

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J Classificat.* 2, 193–218. doi: 10.1007/BF01908075

Jain, V., Bollmann, B., Richardson, M., Berger, D. R., Helmstaedter, M. N., Briggman, K. L., et al. (2010). "Boundary learning by optimization with topological constraints" in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA: IEEE), 2488–2495. doi: 10.1109/CVPR.2010.5539950

Jain, V., Turaga, S. C., Briggman, K., Helmstaedter, M. N., Denk, W., and Seung, H. S. (2011). "Learning to agglomerate superpixel hierarchies" in *Advances in Neural Information Processing Systems* (Granada), 648–656.

Januszewski, M., Kornfeld, J., Li, P., Blakely, T., Lindsey, L., Maitin-Shepard, J., et al. (2018). High-precision automated reconstruction of neurons with flood-filling networks. *Nat. Methods* 15, 605–610. doi: 10.1038/s41592-018-0049-4

Kasthuri, N., Hayworth, K. J., Berger, D. R., Schalek, R. L., Conchello, J. A., Knowles-Barley, S., et al. (2015). Saturated reconstruction of a volume of neocortex. *Cell* 162, 648–661. doi: 10.1016/j.cell.2015.06.054

Knowles-Barley, S., Kaynig, V., Jones, T. R., Wilson, A., Morgan, J., Lee, D., et al. (2016). Rhoananet pipeline: dense automatic neural annotation. *arXiv:1611.06973* [preprint].

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Res. Logist. Q.* 2, 83–97. doi: 10.1002/nav.3800020109

Kuhn, H. W. (1956). Variants of the hungarian method for assignment problems. *Naval Res. Logist. Q.* 3, 253–258. doi: 10.1002/nav.3800030404

Lee, K., Zung, J., Li, P., Jain, V., and Seung, H. S. (2017). Superhuman accuracy on the snemi3d connectomics challenge. arXiv:1706.00120.

Lee, W.-C. A., Bonin, V., Reed, M., Graham, B. J., Hood, G., Glattfelder, K., et al. (2016). Anatomy and function of an excitatory network in the visual cortex. *Nature* 532, 370–374. doi: 10.1038/nature17192

Marblestone, A. H., Daugharthy, E. R., Kalhor, R., Peikon, I. D., Kebschull, J. M., Shipman, S. L., et al. (2014). Rosetta brains: a strategy for molecularly-annotated connectomics. arXiv:1404.5103.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math.* 5, 32–38. doi: 10.1137/0105003

Nunez-Iglesias, J., Kennedy, R., Parag, T., Shi, J., and Chklovskii, D. B. (2013). Machine learning of hierarchical clustering to segment 2D and 3D images. *PLoS ONE* 8:e71715. doi: 10.1371/journal.pone.0071715

Plaza, S.M. (2014). Focused proofreading: efficiently extracting connectomes from segmented EM images. *arXiv:1409.1199* [preprint].

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66, 846–850. doi: 10.1080/01621459.1971.10482356

Saalfeld, S., Fetter, R., Cardona, A., and Tomancak, P. (2012). Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nature* 9, 717–720. doi: 10.1038/nmeth.2072

Takemura, S., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S., Rivlin, P. K., et al. (2013). A visual motion detection circuit suggested by Drosophila connectomics. *Nature* 500, 175–181. doi: 10.1038/nature12450

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.