A Graph Coarsening Algorithm for Compressing Representations of Single-Cell Data with Clinical or Experimental Attributes

Chi-Jane Chen[†], Emma Crawford[ʌ], and Natalie Stanley[‡]

Department of Computer Science and Computational Medicine Program
The University of North Carolina at Chapel Hill,
Chapel Hill, NC, 27599, USA
{[†]*chijane@cs.unc.edu*,[ʌ] *emmabc@email.unc.edu*,[‡] *natalies@cs.unc.edu*}

Graph-based algorithms have become essential in the analysis of single-cell data for numerous tasks, such as automated cell-phenotyping and identifying cellular correlates of experimental perturbations or disease states. In large multi-patient, multi-sample single-cell datasets, the analysis of cell-cell similarity graphs representations of these data becomes computationally prohibitive. Here, we introduce *cytocoarsening*, a novel graph-coarsening algorithm that significantly reduces the size of single-cell graph representations, which can then be used as input to downstream bioinformatics algorithms for improved computational efficiency. Uniquely, cytocoarsening considers both phenotypical similarity of cells and similarity of cells' associated clinical or experimental attributes in order to more readily identify condition-specific cell populations. The resulting coarse graph representations were evaluated based on both their structural correctness and the capacity of downstream algorithms to uncover the same biological conclusions as if the full graph had been used. Cytocoarsening is provided as open source code at `https://github.com/ChenCookie/cytocoarsening`.

Keywords: Graph Coarsening; Single-Cell Bioinformatics; Cytometry

1. Introduction

Advancements in a range of single-cell technologies, such as flow and mass cytometry and single-cell RNA sequencing, have become essential in uncovering and understanding cellular heterogeneity in a range of translational applications.[1–3] These immune profiling techniques have proven to be particularly essential in unraveling immunological heterogeneity through the simultaneous measurement of 20-45 protein markers in each cell.[4] This simultaneous measurement enables both phenotypic (e.g. cellular identity) and functional characterization of cells.[5] Despite effective identification and characterization of immune cell-types, a current challenge is to accurately link these immune cells to external *attributes* of interest, such as clinical labels or experimental perturbations.[6–9] For example, it is common in translational applications to profile blood samples from patients *across* clinical phenotypes or disease states in order to identify the driving, stratifying cell-types.[6,10] Blood samples are also often perturbed through stimulation,[11] and cellular correlates are identified by observing functional responses to the stimulation. Moreover, to efficiently link cellular heterogeneity to clinical or experimental attributes, automated bioinformatics methods have become critical in analysis.

Many of the bioinformatics algorithms for such tasks operate on a graph representation of the single-cell data.[7–9] In these graphs, nodes are cells, and edges between a pair of cells imply that they are sufficiently similar across measured features (for example, the aforementioned protein markers). The task at hand is to use the graph structure to identify cells that are prototypical of particular external attributes, such as clinical or experimental labels. MELD[7] accomplishes this by modeling the external attributes as a signal on the graph and computing a score for each cell reflecting its probability of association with each condition. To exemplify another approach, Milo[8] and CNA[9] seek to identify critical *cellular neighborhoods*, or groups of phenotypically-similar cells enriched across attributes.

Practically, it is challenging to apply these bioinformatics algorithms to the extremely large graph representations of multi-patient, multi-sample cohorts with millions of cells. Although the large graph size would make computations on it prohibitive, the graph inherently involves redundant information, since we have multiple cellular instances from a single population encoding the same biological information. To reduce the graph size, then, we merge redundant cells into *coarse nodes* or *super nodes*, leveraging existing graph-coarsening strategies[12,13] and adapting them to consider biologically relevant external attributes. The rich literature of existing graph-coarsening methods[13–18] tend to optimize for merges of nodes that maintain critical structural and spectral properties for the original graph, but do not consider these node attributes.

Baselines. As an example of a graph-coarsening approach, Loukas *et al.* proposed a family of *local variation* algorithms to simplify and reduce the size of the original graph.[14] These algorithms begin with a family of *coarsening candidate sets*: subsets of nodes that are known to be highly related based on the graph structure. The two main approaches discussed are edge-based variation (LV-E) or node-based variation (LV-N). Using LV-E, the candidate sets are exactly the edge pairs of the graph. In contrast, the candidate sets in LV-N are formed by grouping each node with its immediate neighborhood. In Ref.14, Loukas *et al.* compared these variation-based methods to other graph coarsening methods, including heavy-edge matching (HEM),[15] algebraic distance (AD),[16] and affinity (AFF).[17] The local variation methods outperformed these methods in spectral approximation, and all of the methods (with the exception of AFF, which is slower) scale quasi-linearly in the number of edges in the graph. Briefly, HEM seeks to coarsen the graph such that the principal eigenvalues and eigenspaces of the coarsened graph Laplacian are close to those of the original graph Laplacian. Instead of considering spectral properties, the AD and AFF methods identify nodes to merge by considering the connectedness of both individual nodes and node neighborhoods.

With existing coarsening approaches focusing primarily on preserving overall graph structure or underlying spectral properties, we seek to adapt the methods to additionally take into account external attributes of the cells, such as clinical state or experimental perturbation status. Our method will therefore merge individual nodes (representing cells) into coarse nodes according to both cellular phenotype and associated attributes (see overview figure, Fig 1). This gives us a graph of reduced size to use as input for downstream bioinformatics algorithms, and it facilitates simpler identification of cells that are related both in phenotype and in clinical or experimental attribute.

2. Methods

Notation and problem formulation. We consider a multi-sample single-cell dataset with p profiled samples, denoted as $\{\mathbf{X}_i\}_{i=1}^{p}$. Here, each $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ represents the d protein or gene expression measurements for each of the n_i cells measured in sample i. We also assume that each cell has an *attribute label* (such as experimental label or disease state), encoded in the vector $\mathbf{x}$. A graph representation of all of these cells would render further computation expensive and time-consuming. Thus, we seek a graph representation of the $N = \sum_{i=1}^{p} n_i$ cells that has $N' << N$ nodes while still representing the biologically relevant information that would be present in the full graph. To accomplish this, we introduce the *cytocoarsening* algorithm. In this section, we outline the general steps of the algorithm; pseudocode is provided in Algorithm 1.

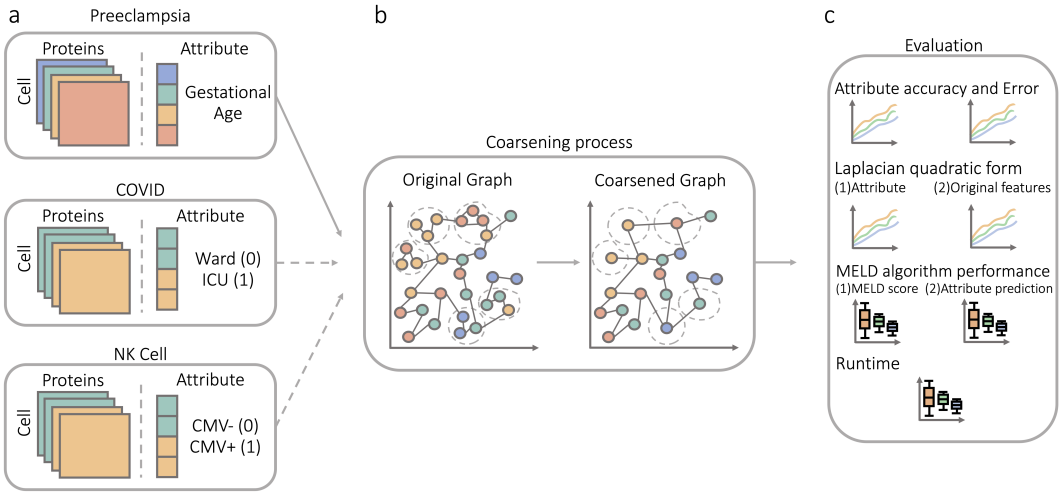

Fig. 1. **Overview.** Given a multi-sample single-cell dataset with clinical attributes (**a**), the cytocoarsening algorithm creates a *coarse* graph representation of all cells (**b**). The coarse graph representation takes into account phenotypic similarity of cells (edges) and the clinical attributes (colors). (**c**) Quantitative evaluation metrics were developed to assess the quality of the coarse graph representation and its effectiveness as input to downstream graph-based bioinformatics algorithms.

Graph representation of single-cell data. The algorithm begins by constructing a joint graph representation $\mathcal{G}$ of all profiled cells across samples. Given a data matrix of cells $\times$ measured features defined as $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_p]$ (where $|$ denotes vertical concatenation), each cell is connected to its K nearest neighbors according to Euclidean distance in the measured feature space via scikit-learn's `kneighbors_graph` function[19] (`KNN()` in Algorithm 1). To actually carry out computations with this graph, we will use the adjacency matrix $\mathbf{A}$, which has all the edge weights of the graph encoded in its off-diagonal entries and zeros on the diagonal. We will also use the graph Laplacian $\mathbf{L}$, which is exactly the negative of this matrix but with a diagonal instead defined as $L_{i,i} = \sum_{j=1}^{N} A_{i,j}$.

Algorithm 1 Cytocoarsening

Inputs: feature matrix $\mathbf{X}$, attribute vector $\mathbf{x}$, number of passes P, number of KNN neighbors K, cutoff parameter α
Output: coarsened graph $\mathcal{G}'$
for $1 : P$ **do** ▷ P coarsening passes
  $\mathcal{G}$=`KNN`$(\mathbf{X}, K)$ ▷ Creates K-nearest neighbor graph from feature matrix
  C =`Get.K.Neighborhoods`$(\mathcal{G})$ ▷ Identifies coarsening candidates
  I^C = `Get.Index.Sets`(C) ▷ Gets indices of nodes in each candidate set
  $T = |C|/4$ ▷ Defines max number of coarse nodes
  for $C_j \in C$ **do**
    $c_j^d = \max_{j,k \in I_i^C} \{||\mathbf{X}_{j,:} - \mathbf{X}_{k,:}||_2\}$ ▷ Calculates distance cost
    $c_j^q = \mathbf{x}_{C_i}{}' {}^T \mathbf{L}_{C_i} \mathbf{x}_{C_i}{}'$ ▷ Calculates attribute cost
  end for
  $\{T^q, T^d\}$ = `Set.Thresholds`$(\mathbf{c}^q, \mathbf{c}^d, \alpha)$ ▷ Finds α^{th} percentile of each cost vector
  C^L =`Nodes.To.Coarsen`$(C, \mathbf{c}^q, \mathbf{c}^d)$ ▷ Finds lowest-cost coarsening candidates
  $\{S, I^S\}$ = `Form.Super.Nodes`$(C^L, V(\mathcal{G}))$ ▷ Creates coarse graph node list
  for $i = 1, \ldots, |C^L|$ **do**
    $\tilde{S}_i$ = `Find.Representative`(C_i^L) ▷ Locates optimal super node representative
  end for
  $\mathcal{G}'$ = `Make.Graph`(S) ▷ Creates coarse graph with node set S
  $\{\mathbf{X}, \mathbf{x}\}$ = `Update.Xs`$(\mathbf{X}, \mathbf{x}, I^S)$ ▷ Updates for next pass
end for

Establishing and ranking coarsening candidates. The KNN graph is used to define the *coarsening candidate node sets* as each node and its K nearest neighbors; the candidate sets are stored in the list C with corresponding index set list I^C, i.e. $I_j^C = \{i | v_i \in C_j\}$ (KNN enumeration → `get.K.Neighborhoods`(), indices of nodes within coarsening candidate → `get.Index.Sets`() in Algorithm 1). To decide which candidate sets to coarsen, we define two different cost functions: distance in feature space ($\mathbf{c}^d$) and graph-level attribute variation ($\mathbf{c}^q$).

Distance cost ($\mathbf{c}^d$). The distance cost reflects the overall phenotypical similarity between cells in a coarsening candidate to ensure that highly similar nodes are likely to be aggregated. We define c_i^d, the distance cost of the i^{th} coarsening candidate, as the maximum euclidean distance of all cells within a coarsening candidate:

$$c_i^d = \max_{j,k \in I_i^C} \{||\mathbf{X}_{j,:} - \mathbf{X}_{k,:}||_2\}. \tag{1}$$

Attribute cost ($\mathbf{c}^q$). The attribute cost measures the overall variation of the attributes of cells within a coarsening candidate, so that we can prioritize merges of cells with similar attributes. Given a coarsening candidate, C_i, we can extract its sub-adjacency matrix, $\mathbf{A}_{C_i}$ via $\mathbf{A}_{C_i} = \mathbf{A}(I_i^C, I_i^C)$ and compute its corresponding Laplacian matrix, $\mathbf{L}_{C_i}$. We further let $\mathbf{x}_{C_i} = \mathbf{x}(I_i^C)$ be the corresponding subvector of attributes for the coarsening candidate set.

Then the attribute cost c_i^q for coarsening candidate C_i is computed as

$$c_i^q = \mathbf{x}_{C_i}^T \mathbf{L}_{C_i} \mathbf{x}_{C_i} \tag{2}$$

Joint cost (r). We use a joint ranking criteria to rank coarsening candidates according to their phenotypic between-cell similarity ($\mathbf{c}^d$) and attribute consistency ($\mathbf{c}^q$) by simply taking the log of their geometric mean:

$$r_i = 1/2(\log_2 c_i^q + \log_2 c_i^d). \tag{3}$$

The 30 coarsening candidates with the lowest joint cost are then considered for further evaluation.

Evaluating coarsening candidates. A coarsening candidate C will be added to the coarsening list C^L (i.e. selected to be aggregated) if all of the following are true: 1) less than T coarsening candidates have been chosen, 2) both costs $\mathbf{c}^q$ and $\mathbf{c}^d$ are below some percentile thresholds T^q and T^d (see `Set.Thresholds()` in Algorithm 1) to make sure both two costs are sufficiently low, 3) none of the nodes in C are already represented in the coarsening list. Our method will stop trying to find more coarsening candidates to merge if all coarsening candidates remaining have a cost larger than c_{max}, a global constant. If some nodes in the candidate are already present in C^L, then those nodes are removed from the set and the costs are recomputed for this smaller candidate set. In the cases where only one node remains or there are no edges between the remaining candidate nodes, we assign both costs the value of c_{max} in order to remove that set from consideration (see function `Nodes.To.Coarsen`() in Algorithm 1). Once the coarse node sets have been decided, we form the node set for the coarse graph S (with corresponding index set I^S) by taking the union of the coarse nodes with all the individual nodes from the original graph (see function `Form.Super.Nodes`() in Algorithm 1).

Defining super node representatives. Once we know which sets of nodes to merge, we find the original node in each set that is most representative of the group by considering two factors: phenotypical similarity and attribute similarity. Consider the i^{th} super node in the following discussion. For phenotypical similarity, we find the mean point of the nodes in feature space $\mu_i = \frac{1}{|S_i|} \sum_{j \in I_i^S} \mathbf{X}_{j,:}$, and then we calculate the euclidean distance from μ_i to each node in the set. Weights are assigned so that nodes closer to μ_i are more highly weighted. For attribute similarity, we sort the attribute labels by the number of their occurrences in S_i and weight the nodes so that nodes with frequently-occurring attribute values are more highly weighted. To combine these two weights, we normalize them individually and add them together. The representative node is then chosen as the one with the maximum aggregate weight. We will denote the representative node for the i^{th} super node as $\tilde{S}_i$, with original graph index $I^{\tilde{S}_i}$ (see function `Find.Representative`() in Algorithm 1).

Updating edge list. An edge is defined between a pair of nodes S_i and S_j in the coarse graph if, in the original graph, there was at least one edge between any of the nodes in S_i and S_j. (`Make.Graph`() function in Algorithm 1).

The above outlines one pass of the algorithm. To coarsen further, we update the feature matrix $\mathbf{X}_{new} = \mathbf{X}(I^{\tilde{S}}, I^{\tilde{S}})$ and the attribute vector $\mathbf{x}_{new} = \mathbf{x}(I^{\tilde{S}})$ (see function `Update.Xs()` in Algorithm 1).

3. Results

To explore the effects of graph coarsening on biological information, we applied our cytocoarsening algorithm to three publicly available mass cytometry (e.g. CyTOF) datasets. First, the *preeclampsia dataset*[20] profiles blood samples collected 9.7 millions cells from 45 women throughout their pregnancies (33 features measured per cell). The clinical attribute of interest for this dataset was cell gestational age, which ranged from 8 to 28 weeks. Next, the *covid dataset*[21] contains 6.5 million cells collected from 49 total patients (23 features measured per cell). The patients ranged in severity with 6 healthy patients, 23 patients having mild cases of COVID, and 20 experiencing severe responses and were under ICU care. Due to the imbalance in the number of patients for each severity level, we only considered cells from 22 mild patients (one sample had less than 1,000 cells and was thus not considered) and 20 patients that had severe (ICU) COVID. The attribute of interest was disease severity (mild or severe). Finally, the *NK-cell dataset*[22] contains 261 thousand cells collected from 20 total patients (29 measured features per cell). Cytomegalovirus (CMV) status was the attribute of interest, with nine patients being positive for Cytomegalovirus (CMV) and 11 being negative for CMV.

We performed several experiments (Fig. 1c, additional experiments in Supplementary Information [a]) on cytocoarsening and existing coarsening methods (LV-E, LV-N, HEM, AD, and AFF[14]) to quantify their effectiveness in preserving structural and attribute information and in acting as input to downstream graph-based bioinformatics tasks. All experiments were repeated 30 times, sampling a new subset of cells from each sample. Cytocoarsening was run on all datasets with $P = 10$ passes, thresholds $T^d = 26$ and $T^q = 26$, and the max number of coarse nodes as $T = \frac{1}{4}|C|$, where $|C|$ denotes the number of elements (coarsening candidates) of C.

Accuracy and error of attributes in coarse nodes We defined accuracy and error metrics (Fig. 2a and 2b) to evaluate the consistency of attribute values for cells assigned to a coarse node. For all of the "non super node" cells within a coarse node (e.g. those cells that were not chosen to be the representative), we predicted their attributes to be the same as that of the super node representative. The error and accuracy metrics between the true and inferred attribute labels of cells are defined as

$$\text{Error} = \frac{1}{N}\left(\sum_{i=1}^{N'}\sum_{j\in I_i^S}|x_j - x_i'|\right) \tag{4}$$

$$\text{Accuracy} = \frac{1}{N}\left(\sum_{i=1}^{N'}\sum_{j\in I_i^S}\rho(x_j, x_i')\right) \tag{5}$$

[a] `https://github.com/ChenCookie/cytocoarsening/blob/main/Supplemental_Material.pdf`

where $\rho(x, y)$ returns 1 if x and y are equal and 0 otherwise.

Across datasets and coarsening ratios, Cytocoarsening exhibited superior performance, followed most closely by the variation neighborhood method. We note that the continuous attribute labels of cells in the preeclampsia dataset make the task more challenging than predicting binary attributes.

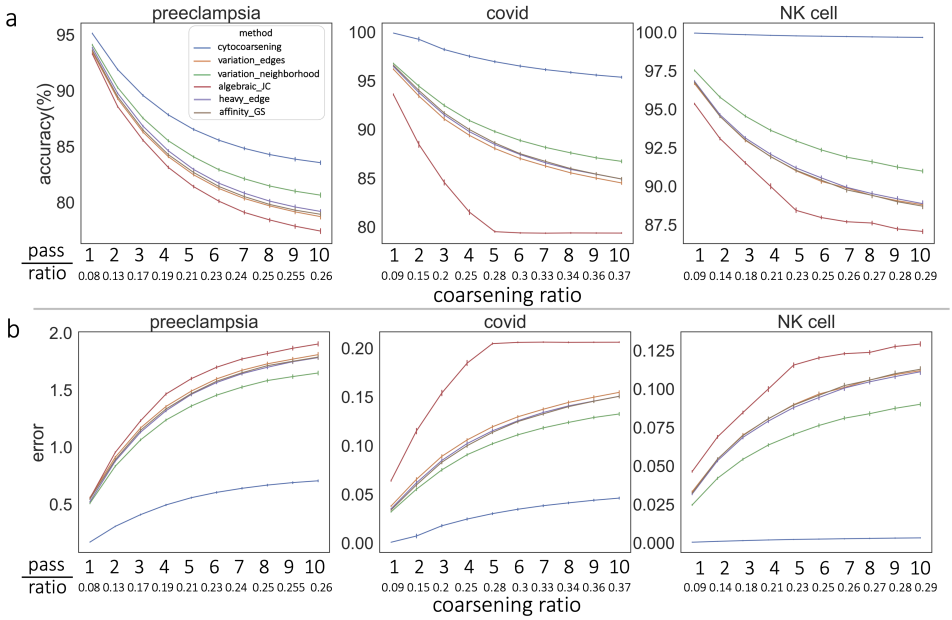

Fig. 2. **Attribute Consistency of Coarse Nodes.** Accuracy (**a**) and error (**b**) metrics were used to evaluate the similarity of attributes within each coarse node. Cytocoarsening (blue) excels in accuracy and error at maintaining consistent attributes within coarse nodes across datasets. For details about baselines, refer to "**Baselines**" in the introduction.

Quantifying attribute and original feature variation across the coarse graph Given the graph Laplacian $\mathbf{L}' = \mathbf{L}(I^{\tilde{S}}, I^{\tilde{S}})$ corresponding to the coarse graph $\mathcal{G}'$ and the coarse attribute vector $\mathbf{x}' = \mathbf{x}(I^{\tilde{S}})$, the normalized Laplacian quadratic form $\frac{1}{N'}\mathbf{x}'^{T}\mathbf{L}'\mathbf{x}'$ (where N' is the number of coarse graph nodes) summarizes the alignment between structure and attributes. Since the Laplacian quadratic form is small for vectors where neighboring nodes have similar vector entries, the quadratic form will be small if alignment is good (Fig. 3a). Similarly, we can quantify the overall variation in the features over $\mathcal{G}'$ (Fig. 3b) as $\frac{1}{N'}\text{trace}(\mathbf{X}'^{T}\mathbf{L}'\mathbf{X}')$, where $\mathbf{X}' = \mathbf{X}(:, I^{\tilde{S}})$ is the coarsened feature matrix.

A good coarsening strategy would produce low values for the Laplacian quadratic forms for both attributes and in the features used to construct the original graph, implying those vary smoothly over the graph. Results across the three datasets in Fig. 3 reveals cytocoarsening produces the lowest values for both attributes (a) and original features (b) for all coarsening ratios, suggesting the cytocoarsening faithfully encodes such information.

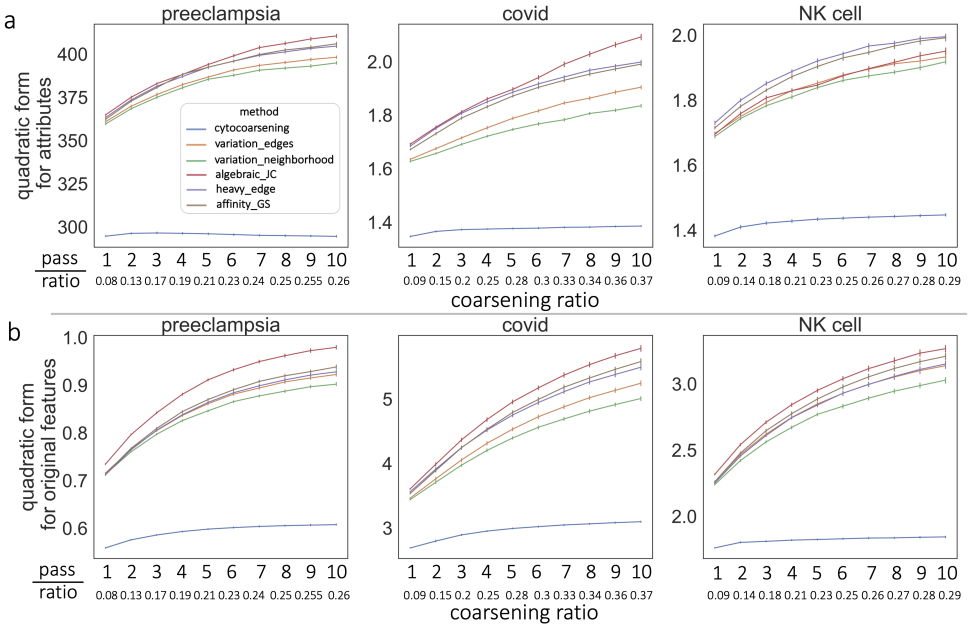

Fig. 3. **Evaluating Variation of Attributes and Original Features on $\mathcal{G}'$.** We used the Laplacian quadratic form on the coarse graph $\mathcal{G}'$ to quantify the variation of the attributes (**a**) and the original features (**b**) over $\mathcal{G}'$ as a function of the extent of graph coarsening (horizontal axis). Cytocoarsening (blue) achieves by far the lowest values for both attributes (**a**) and original features (**b**) across coarsening ratios.

Coarse graphs can be used as input to MELD. To see that we would reach the same biological conclusions by analyzing $\mathcal{G}$ and $\mathcal{G}'$, we used both of these graphs as inputs to MELD[7] and compared the results. Given binary attribute values $\{0, 1\}$, MELD returns a list M, where M_j is the probability that node v_j has an attribute value of 1. We therefore binarized the returned MELD score for a node as 1 if the for node j, $M_j > 0.5$ and assigned it a 0 otherwise. Let $\mathbf{m}^{coarse}$ denote the vector of coarse graph MELD scores. We assigned all nodes within a super node S_j to have the same MELD score as the super node representative. Notationally, then, we have $m_i^{\text{coarse}} = M_j$ whenever node v_i is in the j^{th} super node. Let $\mathbf{m}^{\text{orig}}$ denote the vector of MELD scores of the original graph. We then defined two measures to quantify the similarity and correctness of the MELD results obtained for $\mathcal{G}$ and $\mathcal{G}'$: first, Acc_{MELD} for accuracy. The accuracy metric quantifies the correctness of the MELD score in the coarse graph, defined as

$$Acc_{\text{MELD}} = \frac{1}{N}\left(\sum_{i=1}^{N} \rho(m_i^{\text{orig}}, m_i^{\text{coarse}})\right). \tag{6}$$

Here, $\rho(x, y)$ returns 1 if x and y are equal and 0 otherwise. The results shown in Fig. 4b show that cytocoarsening has the highest MELD score correctness in the coarse graph when setting the smoothness parameter to the default of $\beta = 1$. We note that the attributes for preeclampsia dataset were dichotimized into early and late pregnancy. Although the other methods achieved accuracies above 0.9, cytocoarsening consistently achieved the highest results across datasets with both discrete and continuous attributes. Next, we computed $Corr_{\text{MELD}}$, which is the

Pearson correlation [b] between MELD scores of the coarsened graph and those of the original graph (Fig. 4a).

A high correlation implies high concordance between the MELD scores using $\mathcal{G}'$ as input and those obtained using $\mathcal{G}$, i.e. no critical biologically-meaningful information was lost by reducing the size of the graph. All coarsening methods achieved a reasonable $Corr_{\text{MELD}}$ in all three datasets (Fig. 4a), with cytocoarsening excelling and followed most closely by LV-N.

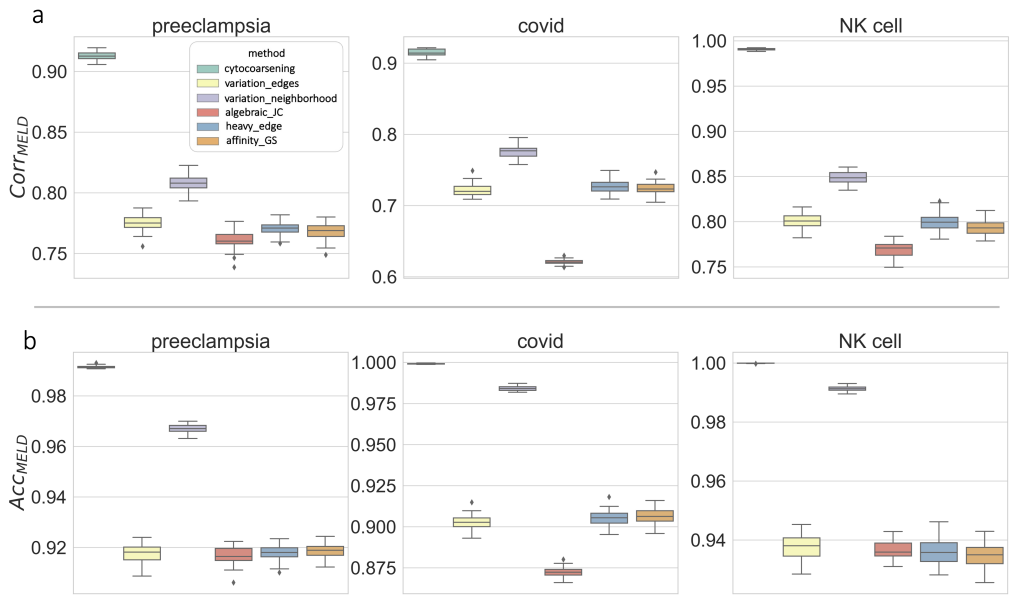

Fig. 4. **Quality of MELD Using $\mathcal{G}'$ as Input**. We computed metrics to evaluate the correlation (**a**) and the overall accuracy (**b**) between MELD results obtained on $\mathcal{G}$ and $\mathcal{G}'$ for six different coarsening methods and three datasets. Results suggest that cytocoarsening, followed by LV-N, produce coarse graph representations that are adequate inputs to MELD.

Sensitivity of MELD parameters in coarse graph representations. MELD has a critical parameter, β, which controls the smoothness or consistency of MELD scores across the graph. To study performance as a function of β, we varied β when computing MELD scores on both the original graph $\mathcal{G}$ and the coarse graph $\mathcal{G}'$ (we denote the parameter in each case as denoted β and β', respectively). We note that due to MELD's expensive runtime, all experiments used only 200 cells per sample. The resulting $Corr_{\text{MELD}}$ scores (averaged over 30 trials) are visualized in the heatmap in Fig. 5. Cytocoarsening achieved the highest scores (denoted by stars) across datasets and combinations of β and β' in 29 of the 48 comparisons (e.g. heatmap grids). The LV-N and LV-E methods are second and third in performance with a total of 12 and 11 best scores, respectively, and they perform more optimally for high values of β and β'.

[b]https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html

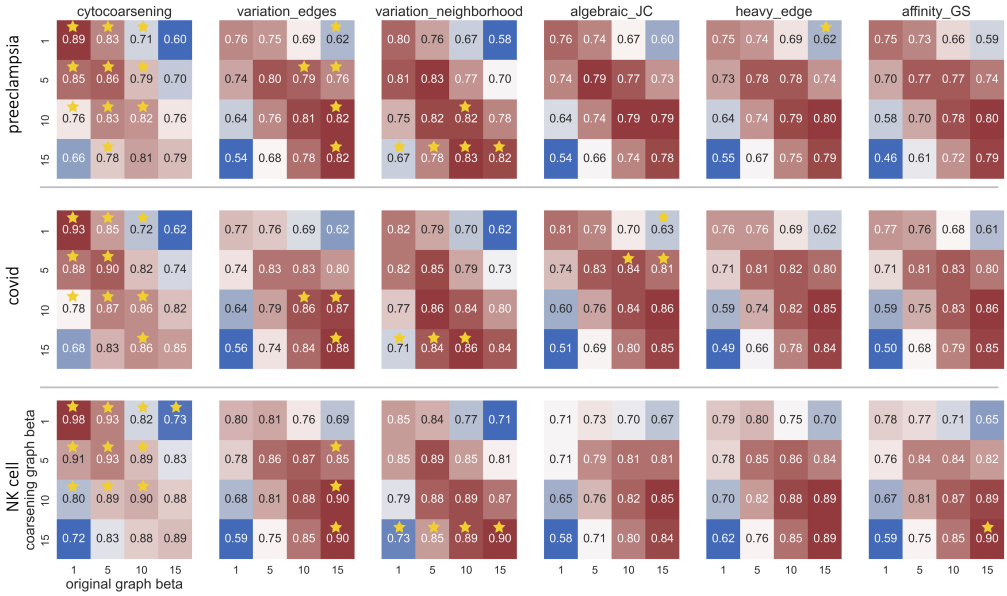

Fig. 5. **Sensitivity of MELD Results to β Parameter.** We evaluated the effect of various combinations for values of MELD's smoothing parameter, β across datasets coarsening methods. Each heatmap grid reflects the $Corr_{\text{MELD}}$ obtained using $\mathcal{G}$ (horizontal axis) and $\mathcal{G}'$ (vertical axis) for a particular dataset, coarsening algorithm and combination of β parameters. A starred grid entry implies that, for that particular combination of β, β', and dataset, the starred algorithm achieved the highest $Corr_{\text{MELD}}$ score; this is frequently achieved by cytocoarsening.

Runtime and scalability. We compared the scalability of cytocoarsening to all other coarsening methods[c] using 1000 subselected cells from each sample. (Fig. 6). To objectively compare our multipass cytocoarsening method to existing coarsening methods, which are only one pass, we also ran cytocoarsening with a single pass. Our results show that AFF has by far the longest runtime across three datasets. Although cytocoarsening is not the fastest method, the runtime only differs slightly from the other four methods. The preeclampsia dataset is the largest in terms of patient samples and measured features and hence took the most time. In contrast, the NK cell dataset is significantly smaller and took half the time (Fig. 6).

4. Discussion

The cytocoarsening algorithm compresses graphs of single-cells by adapting standard graph coarsening approaches to accommodate the associated clinical or experimental cellular attributes. While existing graph coarsening approaches are optimized to create a compressed graph representation with strong *structural* similarity to the original graph, our approach uses new cost functions and a joint ranking strategy to incorporate biologically meaningful cellular information into the coarsening process. We defined several quantitative evaluation strategies to evaluate cytocoarsening and the other existing coarsening approaches on their capacity to preserve more than just structural properties of the original graph. Using three

[c]https://github.com/loukasa/graph-coarsening

CyTOF datasets, we showed that, in comparison to other methods, the cytocoarsening method excels in grouping together cells that are both related in phenotype and in disease state or experimental condition.

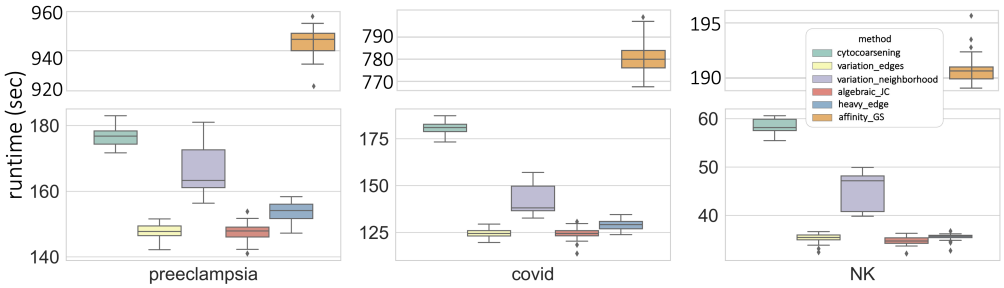

Fig. 6. **Run-Time Evaluations.** Evaluating run-time of all coarsening approaches across datasets, using 1000 cells per profiled sample. Cytocoarsening has similar run-times to the other coarsening strategies, while offering increased performance in encoding attribute information.

Cytocoarsening is a methodological innovation towards adapting primarily structure-preserving coarsening algorithms to single-cell data with associated clinical or experimental attributes, with the aim to compress the input graph for downstream graph-based bioinformatics algorithms. However, to further increase the utility of cytocoarsening in analyzing modern multi-sample flow and mass cytometry datasets, we can modify the initial graph-construction phase for improved scalability. An area of future work is to build coarse graph representations for each sample in *parallel*, and then merge there graphs in a principled manner. Further, additional work can explore how to optimize the coarsening ratio for a particular graph. In summary, Cytocoarsening facilitates more rapid identification of phenotypically-similar cells that are likely associated with a clinical or experimental condition.

References

1. E. A. Ganio, N. Stanley, V. Lindberg-Larsen, J. Einhaus, A. S. Tsai, F. Verdonk, A. Culos, S. Ghaemi, K. K. Rumer, I. A. Stelzer *et al.*, Preferential inhibition of adaptive immune system dynamics by glucocorticoids in patients after acute surgical trauma, *Nature communications* **11**, 1 (2020).
2. A. S. Tsai, K. Berry, M. M. Beneyto, D. Gaudilliere, E. A. Ganio, A. Culos, M. S. Ghaemi, B. Choisy, K. Djebali, J. F. Einhaus *et al.*, A year-long immune profile of the systemic response in acute stroke survivors, *Brain* **142**, 978 (2019).
3. V. L. Tawfik, N. A. Huck, Q. J. Baca, E. A. Ganio, E. S. Haight, A. Culos, S. Ghaemi, T. Phongpreecha, M. S. Angst, J. D. Clark *et al.*, Systematic immunophenotyping reveals sex-specific responses after painful injury in mice, *Frontiers in immunology* **11**, p. 1652 (2020).
4. T. Liechti, L. M. Weber, T. M. Ashhurst, N. Stanley, M. Prlic, S. Van Gassen and F. Mair, An updated guide for the perplexed: cytometry in the high-dimensional era, *Nature Immunology* **22**, 1190 (2021).
5. M. H. Spitzer and G. P. Nolan, Mass cytometry: single cells, many features, *Cell* **165**, 780 (2016).
6. N. Stanley, I. A. Stelzer, A. S. Tsai, R. Fallahzadeh, E. Ganio, M. Becker, T. Phongpreecha,

H. Nassar, S. Ghaemi, I. Maric *et al.*, Vopo leverages cellular heterogeneity for predictive modeling of single-cell data, *Nature communications* **11**, 1 (2020).
7. D. B. Burkhardt, J. S. Stanley, A. Tong, A. L. Perdigoto, S. A. Gigante, K. C. Herold, G. Wolf, A. J. Giraldez, D. van Dijk and S. Krishnaswamy, Quantifying the effect of experimental perturbations at single-cell resolution, *Nature biotechnology* **39**, 619 (2021).
8. E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan and J. C. Marioni, Differential abundance testing on single-cell data using k-nearest neighbor graphs, *Nature Biotechnology* **40**, 245 (2022).
9. Y. A. Reshef, L. Rumker, J. B. Kang, A. Nathan, I. Korsunsky, S. Asgari, M. B. Murray, D. Moody and S. Raychaudhuri, Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics, *Nature Biotechnology* **40**, 355 (2022).
10. E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan and J. C. Marioni, Differential abundance testing on single-cell data using k-nearest neighbor graphs, *Nature Biotechnology* **40**, 245 (2022).
11. Z. B. Bjornson-Hooper, G. K. Fragiadakis, M. H. Spitzer, H. Chen, D. Madhireddy, K. Hu, K. Lundsten, D. R. McIlwain and G. P. Nolan, A comprehensive atlas of immunological differences between humans, mice, and non-human primates, *Frontiers in immunology* **13** (2022).
12. Y. Jin, A. Loukas and J. JaJa, Graph coarsening with preserved spectral properties, in *International Conference on Artificial Intelligence and Statistics*, 2020.
13. N. Stanley, R. Kwitt, M. Niethammer and P. J. Mucha, Compressing networks with super nodes, *Scientific reports* **8**, 1 (2018).
14. A. Loukas, Graph reduction with spectral and cut guarantees, *Journal of Machine Learning Research* **20**, 1 (2019).
15. A. Loukas and P. Vandergheynst, Spectrally approximating large graphs with smaller graphs, in *International Conference on Machine Learning*, 2018.
16. D. Ron, I. Safro and A. Brandt, Relaxation-based coarsening and multiscale graph organization, *Multiscale Modeling & Simulation* **9**, 407 (2011).
17. O. E. Livne and A. Brandt, Lean algebraic multigrid (lamg): Fast graph laplacian linear solver, *SIAM Journal on Scientific Computing* **34**, B499 (2012).
18. Y. Jin, A. Loukas and J. JaJa, Graph coarsening with preserved spectral properties, in *International Conference on Artificial Intelligence and Statistics*, 2020.
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
20. X. Han, M. S. Ghaemi, K. Ando, L. S. Peterson, E. A. Ganio, A. S. Tsai, D. K. Gaudilliere, I. A. Stelzer, J. Einhaus, B. Bertrand *et al.*, Differential dynamics of the maternal immune system in healthy pregnancy and preeclampsia, *Frontiers in immunology* , p. 1305 (2019).
21. L. Vanderbeke, P. Van Mol, Y. Van Herck, F. De Smet, S. Humblet-Baron, K. Martinod, A. Antoranz, I. Arijs, B. Boeckx, F. Bosisio *et al.*, Monocyte-driven atypical cytokine storm and aberrant neutrophil activation as key mediators of covid-19 disease severity, *Nature communications* **12**, 1 (2021).
22. E. Arvaniti and M. Claassen, Sensitive detection of rare disease-associated cell subsets via representation learning, *Nature communications* **8**, 1 (2017).