AWARD NUMBER: W81XWH-20-1-0301

TITLE: Identification of Novel Features to Assess Risk and Improve Therapeutic Decision Making for Prostate Cancer Through a Novel High-Parameter Imaging System

PRINCIPAL INVESTIGATOR: Patrick M. Reeves, PhD

CONTRACTING ORGANIZATION: Massachusetts General Hospital, Boston, MA

REPORT DATE: May 2022

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
May 2022	Annual	01May2021-30Apr2022

4. TITLE AND SUBTITLE

Identification of Novel Features to Assess Risk and Improve Therapeutic Decision Making for Prostate Cancer Through a Novel High-Parameter Imaging System

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-20-1-0301

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Patrick M. Reeves, PhD

E-Mail: pmreeves@partners.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts General Hospital
Boston, Massachusetts 02114-2554

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick. Marvland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

About 15% of men will be diagnosed with prostate cancer and many will face difficult treatment decisions. This project's goal is to develop and apply a novel imaging strategy that increases the information and potential clinical utility from prostate biopsy samples. Combining two high-dimensional imaging methods: antibody-based imaging mass cytometry (IMC), and mass spectroscopy imaging (MSI) that are layered together, the goal is to generate exceptionally detailed maps of the cells, molecules, and structures from tumor samples. These layered multi-modal images can be used to identify markers that improve assessments of risk and may ultimately inform clinical decisions. In the first year of the project, despite challenges resulting from the COVID19 pandemic, we are nearing completion of method development and optimization studies, generated a tissue microarray containing (TMA) more than 80 prostate cancer biopsy samples, and have nearly completed imaging of the TMA by MSI. Work on applying the IMC platform to the TMA is ongoing. However, using data generated to date, significant progress was made in developing a novel computational workflow to integrate and analyze high-parameter imaging data. The

15. SUBJECT TERMS

Prostate cancer, imaging, biopsy, prognostic, tissue microarray, mass spectroscopy, imaging mass cytometry, computational analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	Unclassified	72	USAMRDC
Unclassified	Unclassified	Unclassified			19b. TELEPHONE NUMBER *(include area code)*

TABLE OF CONTENTS

	Page
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	14
5. Changes/Problems	15
6. Products	17
7. Participants & Other Collaborating Organizations	17
8. Special Reporting Requirements	19
9. Appendices	20

1.**INTRODUCTION:**

The objective of this research program is to identify novel targets within prostate biopsy tissues that can inform treatment decisions and improve outcomes for patients with prostate cancer. To accomplish this, we aim to enhance the information obtained from prostate cancer biopsy samples using a multi-modal high-parameter imaging approach to identify specific cell populations, proteins, and other tissue features that can inform diagnostic and therapeutic decision making. The premise is that application of advanced imaging methods and computational tools will increase the information and clinical value derived from biopsy tissues. The imaging method combines gold-standard H&E histopathology information with data from two mass spectrometry-based methods. The two methods operate in tandem to provide spatial data for >40 antibodies and a label-free assessment of proteins, lipids, and metabolites. The resulting data can be layered together to provide a highly detailed map and characterization of each biopsy sample. The resulting information can be interrogated for cell populations, proteins, spatial relationships, and other analytes to identify statistically significant associations with clinical data. These analyses can identify signatures that enhance the capacity to discriminate tissues based on prostate cancer grade, risk of progression, and likelihood of recurrence. As anticipated, novel developments for both technical methodologies and computational tools have been required to generate high-quality data and statistically sound analyses of high-parameter imaging data. The first-year project period required careful management and recalibration of efforts to combat complications arising from work stoppages and supply chain constraints resulting from the COVID19 pandemic. In the second project year, supply chain and lab shortage issues have persisted. Despite these challenges, we have been able to advance the studies, submitted a manuscript to Nature Methods, enhanced the capacity of the imaging platform, and recruited staff to support the work. We are enthusiastic about the momentum of the studies and the prospects for meeting the scientific objectives of the grant within budget.

2. KEYWORDS: Prostate, cancer, biopsy, grade, recurrence, progression, prognostic, tissue microarray, imaging mass cytometry, mass spectroscopy imaging

3.**ACCOMPLISHMENTS:**

o **What were the major goals of the project?**

Major Task 2: Specific Aim 1: ***Identify alterations in the quantity and distribution of cellular populations and biomolecules that distinguish between PC tumor grades***.

Subtask 4: Apply validated methodologies to conduct imaging mass cytometry from PC TMA containing a range of tumor grades.

Completed remaining tasks in month 17 (November 2021), consistent with the 3-month delay anticipated in the 2020 progress report.

Subtask 5: Apply validated methodologies to conduct mass spectroscopy imaging of PC TMA containing a range of tumor grades.

Work for Subtask 5 was completed in month 15 (September 2021), a short delay after the initial goal of month 14 (August 2021).

Subtask 6: Determine features that distinguish tumor grades from both mass spectrometry imaging and imaging mass cytometry. Use computational image and clinical correlation analysis of imaging data from both imaging

modalities.

Primary analysis of the data was completed in month 17 (November 2021), 3 months delayed after the initial goal of month 14 (August 2021). Analysis is ongoing for previous and new data generated to address reviewer's comments and enable re-submission to journal.

Subtask 7: Manuscripts & Reporting

A manuscript was submitted in month 18 (December 2021). A second manuscript is pending following acceptance of the first.

Major Task 3: Specific Aim 2: Analyze biopsy tissues to determine those alterations in the abundance and spatial organization of cellular populations and biomolecules that correlate to risk for progression.

Subtask 1: Preparation of tissue microarray (TMA) from prostate cancer (PC) sample with corresponding de-identified up to date clinical data

A TMA was generated by Dr. Wu and was available on schedule.

Subtask 2: Optimization of immunolabeling of PC tissue sections and analysis by imaging mass cytometry

Technical considerations for the optimization of PC tissue section and analysis were completed by end of month 21 (March 2022), three months after the initial target date (December 2021). The optimization included refinements to the antigen retrieval process, antibody panel design, incorporation of additional features to facilitate data analysis. Development of the data analysis and computational tools focused on image processing.

Subtask 3: Conduct mass spectroscopy imaging to generate images from PC tissue collection arranged as a Tissue Microarray (TMA).

Imaging of the TMA was conducted in month 18 (December 2021) in agreement with the schedule. However, a second imaging was required in month 20 (February 2022).

Subtask 4: Computational image and clinical correlation analysis of imaging data from both mass spectrometry imaging and imaging mass cytometry

Computational and statistical analysis of images is ongoing and we expect completion in month 25 (July 2022), five months after the initial target of month 20 (February 2022).

Subtask 5: Manuscripts & Reporting

A manuscript detailing the findings from Aim 2 is pending, with a target of month 27 (September 2022).

Milestone #2: Demonstration of validated imaging method to analyze banked human PC samples, complete image acquisition, and assemble manuscript to convey results

Major Task 4: Specific Aim 3: **Conduct analysis of FFPE patient samples to identify changes in the abundance and spatial organization of both cell populations and biomolecules that correlate to risk of recurrence following prostatectomy.**

Subtask 4: Generate tissue microarray (TMA) containing biopsy cores from prostatectomy tissues. Subsequently update deidentified clinical data and patient outcomes.

In month 18, Dr. Wu made available a TMA containing tissue 80 biopsies appropriate for the objectives of Specific Aim 3. Clinical information is available for these cases and preparation and imaging mass cytometry analysis of these sections by is currently anticipated by end of month 27 (September 2022).

Subtask 2: Conduct imaging mass cytometry on TMA prepared from prostatectomy sample set

Imaging of the TMA for Aim 3 while efforts focus on ongoing work in the preceding Major Tasks. We anticipate completion by the end of month 27 (September 2022), since sections will be cut for both imaging methods simultaneously.

What was accomplished under these goals?

Major Task 2: Specific Aim 1: ***Identify alterations in the quantity and distribution of cellular populations and biomolecules that distinguish between PC tumor grades***.

During the study period the remaining subtasks for Major Task 2 related to the development of methodologies to image prostate biopsy samples and analyze the resulting data were accomplished. The technical and computational methods developed were assembled into a manuscript that was submitted to Nature Methods and sent for external review. Though the overall comments were positive, the reviewers had technical questions and asked that a larger data set be presented within the manuscript. We are now working to meet those requests by including more of the data generated as part of Specific Aim 1. The goal is to prioritize this high-impact publication to illustrate the overall utility of the method. A second manuscript that explores the data in more detail and focuses more narrowly on application to prostate cancer patients will follow.

Subtask 4: Apply validated methodologies to conduct imaging mass cytometry from PC TMA containing a range of tumor grades.

As noted above, Subtask 4 was completed in month 17. A panel containing 40 antibodies (Table 1) was designed, assembled, and validated. Importantly, several key clinically relevant antibodies (e.g. NKX3.1, ERG, cytokeratin 34βE12) were specifically developed for this project. These antibodies are used clinically, either individually or in pairwise combinations, to support histopathological assessments of prostate biopsies. Their inclusion provides a benchmark for trained pathologists to assess data generated using our multi-modal imaging approach. Generating these antibodies required in house conjugation and validation, since they were not available from commercial vendors. Following antigen retrieval, the antibody panel was applied to the TMA shown in Figure 1. The imaging mass cytometry of the TMA required approximately 7 days of continuous acquisition to complete.

Metal Label	Antibody Target
141Pr	aSMA
142Nd	EGFR
143Nd	p53
144Nd	EpCAM
145Nd	T-bet
146Nd	CD16
147Sm	CD163
148Nd	P63
149Sm	CD15
150Nd	CD38
151Eu	LAMP1
152Sm	CD45
153Eu	CD44
154Sm	CD38
155Gd	IDO
156Gd	CD4
158Gd	E-Cadherin
159Tb	CD68
160Gd	CCR6
161Dy	CD20
162Dy	CD8a
163Dy	CCR7
164Dy	c-Myc
165Ho	PD-1
166Er	NKX3.1
167Er	ERG
168Er	Ki-67
169Tm	p504s (AMACR)
170Er	CD3
171Yb	Keratin
172Yb	34βE12
173Yb	CD45RO
174Yb	Keratin 8/18
175Lu	CD25
176Yb	pTen
194Pt	ERG
195Pt	NaK ATPase
196Pt	ZO-1
198Pt	Vimentin

Table 1. 40-Antibody Panel
Metal conjugated antibodies for labeling prostate tissues

Subtask 5: Apply validated methodologies to conduct mass spectrometry imaging of PC TMA containing a range of tumor grades.

Working with ImaBiotech to accomplish the mass spectrometry imaging, Subtask 5 was completed in month 15. The initial step was to conduct method development studies with the goal to optimize the conditions for processing and imaging of slides containing the PC TMA. Through the course of the development process several conditions were evaluated including: two distinct antigen retrieval conditions (Citrate pH 6 or Tris-EDTA pH 9), various antigen digest conditions, and matrix were applied (summary Table 2). Unfortunately, testing of these conditions yielded two major concerns. First the integrity of the tissue was significantly impacted, with loss of tissue structure or completely loss of the tissue observed for numerous cores (Figure 1). Second, for those tissue cores that remained, the quality of the data – both in terms of spatial resolution and signal-to-noise was not robust. This was a significant concern, given that the capacity to align data obtained from adjacent tissue sections imaged using different methods is central to the execution of this study. This performance loss was at odds with our previous experience and data when working with ImaBiotech. In discussion with ImaBiotech, we learned that several changes to their process and COVID19 induced challenges to accessing instrumentation likely contributed. To mitigate the impact of these changes, we undertook additional testing and troubleshooting steps as well as tested a second mass spectrometry imaging method (FTICR) which does not require specially coated slides and is less prone to such technical challenges.

Digestion at 37°C	Matrix	Recrystallization
2 hrs / 4 hrs	CHCA	No
2 hrs / 4 hrs	CHCA	Yes
2 hrs / 4 hrs	DHB	No
2 hrs / 4 hrs	DHB	Yes
2 hrs	DHB Ammonium Sulfate	No
2 hrs	DHB Ammonium Sulfate	Yes
Overnight	CHCA/Aniline	No

Table 2. Tissue Processing Conditions for Mass Spectrometry Imaging Following incubation at 95C in either pH6 or pH9 buffer, tissue sections were processed according to the matrix above and imaged.

Following further troubleshooting and consultation with the scientists at ImaBiotech, we sought the expert advice of Erin Seeley, PhD. Dr. Seeley is Director of the Mass Spectrometry Imaging Facility at the University of Texas, Austin (which came online in early 2021). Dr. Seeley has extensive academic and industry experience conducing mass spectrometry imaging experiments on tissue samples. Working with Dr. Seeley, we have been able overcome the previous technical challenges and were able to process tissues without loss of tissue integrity and obtain data with good signal to noise with 30μM and 20μM resolution and signal loss at 10μM (Figure 2).

Importantly, the core that Dr. Seeley directs was established expressly for the purpose of facilitating academic implementation of mass spectrometry imaging technologies, from sample preparation to data analysis. One of the newly hired team members here has previously worked with Dr. Seeley and possesses both the technical and computational skills needed to improve the quality and speed of this component of the project.

Subtask 6: Determine features that distinguish tumor grades from both mass spectrometry imaging and imaging mass cytometry. Use computational image and clinical correlation analysis of imaging data from both imaging modalities

The initial analysis of imaging mass cytometry and mass spectrometry imaging data from prostate tissue was central to the development of our novel analysis pipeline "MIAAIM" (manuscript discussed below). MIAAIM (Multi-omics Image Alignment and Analysis by Information Manifolds) analysis pipeline where data from multiple imaging modalities can be aligned, processed, combined to create an overlayed image set allowing for simultaneous cross-modal interrogation of data. One of the major aspects of MIAAIM is the capacity to predict the cellular composition between samples and tissue types.

To do this, high-dimensional data are used to build a model where cell phenotypes are matched to mass spectrometry imaging data, and structured as a data manifold. he resulting model(s) or manifold(s), which contain both MSI and IMC data, can be used to predict the composition of other tissues where only one of the two data types is available (e.g. MSI but not IMC data). These predictions are based on the intersections between manifolds for the known data and the test sample.

This is useful because, mass spectrometry imaging data can be generated over a larger area more quickly than imaging mass cytometry. Therefore, it is possible to image a large number of samples or an entire prostatectomy sample using MSI and then make predictions of the cellular composition. These predictions can be used to guide selection for individual regions of interest, which are then imaged to verify the prediction and/or update the model to improve it. Here we show how information from diabetic foot ulcers (DFUs), which are well infiltrated by immune cells, can be used to predict poor infiltration into prostate tissue (Figure 3). These predictions were then validated using imaging mass cytometry (as show in Fig 3 e-h).

In ongoing analysis, refinements to MIAAIM based on reviewer feedback are being applied to the entire 84-core TMA is being used to build models of the four PC tumor grades. From these models, the defining features will be used to delineate "signatures" of each tumor grade based on the cellular composition and distribution of mass spectrometry analytes. This analysis is essential for the re-submission of the MIAAIM manuscript (see below) and we plan an addendum to this report to detail those findings and updated manuscript.

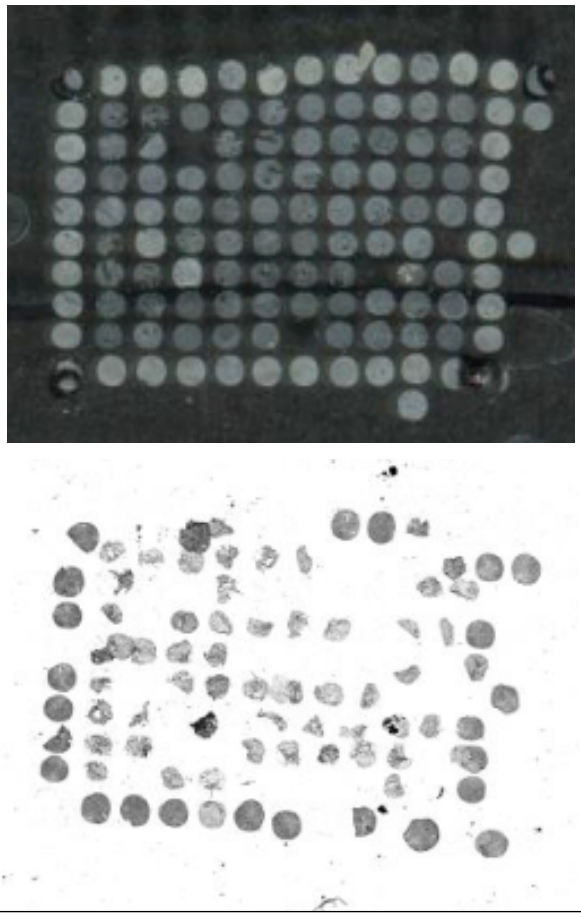


Figure 1. Impact of PC tissue processing. The upper panel shows a PC TMA prior to antigen retrieval and processing. The lower panel is the same TMA slide follow processing. Note the loss and deformation of cores.

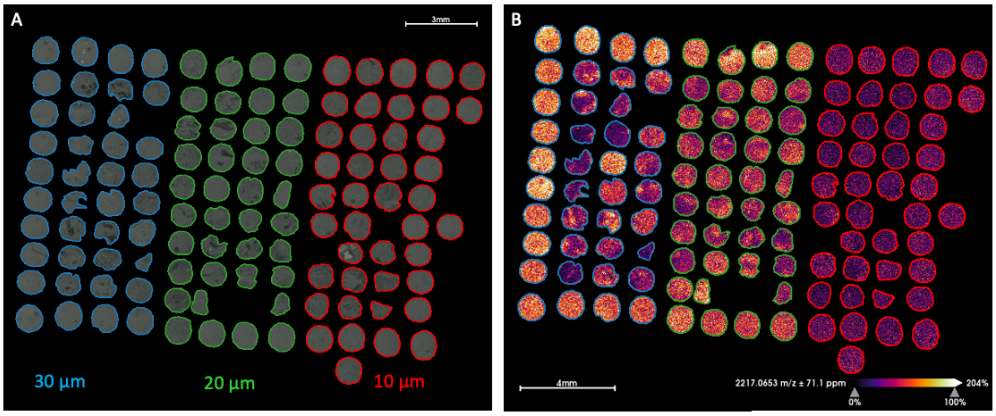


Figure 2. Mass Spectrometry Imaging of PC TMA. An adjacent section from the same TMA in Figure 1 was processed incorporating Dr. Seely's input. **A.** TMA was imaged at three resolutions (30, 20, or 10 µM2/pixel). **B.** Same TMA section, now showing the impact of resolution on signal intensity.

Subtask 7: Manuscripts & Reporting

A manuscript, titled "MIAAIM: Multi-omics image integration and tissue state mapping using topological data analysis and cobordism learning", which details the computational framework developed for multi-modal imaging analysis was prepared based on the approaches developed here. Data from analysis of both prostate and diabetic ulcer tissues were used to demonstrate the utility of MIAAIM in analysis of multi-modal data. The

manuscript submitted to Nature Methods and a pre-print is available (DOI: 10.1101/2021.12.20.472858). Following the reviewer's comments on the manuscript, we have undertaken additional work to address technical question, refine the text for clarity, and meet the request for more clinical significance by incorporating analysis of the 84-core PC TMA imaged in Subtasks 4 and 5.We anticipate re-submission to Nature Methods by July 15, 2022. A second manuscript, which provides a more detailed analysis of the clinical correlations and biological implications from the work in Major Task 3 will follow on from the Nature Methods manuscript.

Major Task 3: Specific Aim 2: Analyze biopsy tissues to determine those alterations in the abundance and spatial organization of cellular populations and biomolecules that correlate to risk for progression.

Subtask 1: Preparation of tissue microarray (TMA) from prostate cancer (PC) sample with corresponding de-identified up to date clinical data

Dr. Wu has identified and assembled TMAs for this Aim on schedule. Two TMAs were selected, one with 48 cores and a second with 80 cores. The rationale was to use the TMA with 48 cores to address any challenges with the mass spectrometry imaging as experienced in Aim 1 and then image the larger TMA.

Subtask 2: Optimization of immunolabeling of PC tissue sections and analysis by imaging mass cytometry

Based on learning points from the previous imaging experiments several modifications were made to the technical process of preparing and labeling the tissue. We developed an antigen retrieval process that is highly reproducible by utilizing a pressurized heating chamber that can hold desired temperature close to 95C without risk of boiling, and releases pressure in a controlled manner to preserve tissue integrity. We explored several antigen retrieval buffers and conditions ranging from pH6 to pH9, coupled with other details including the material of the

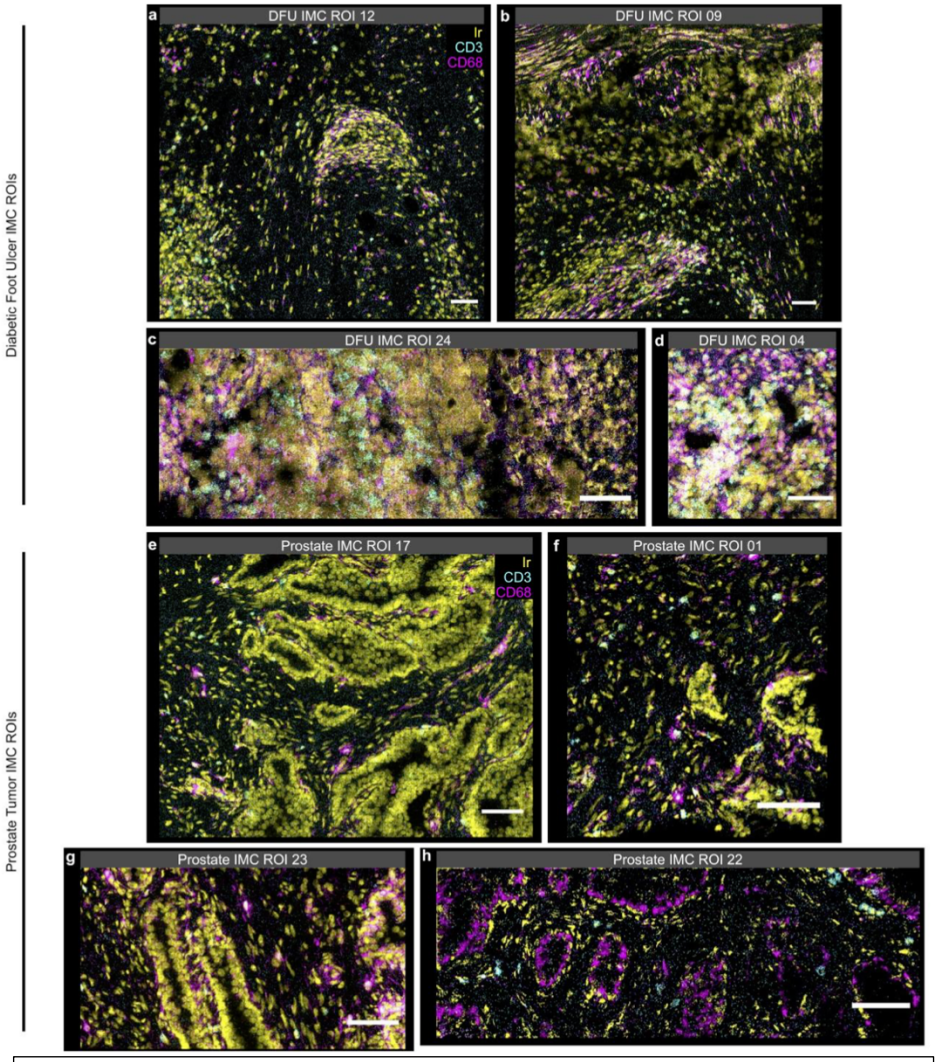


Figure 3. Sparse immune cell infiltration in prostate tumor biopsy IMC ROIs supports lack of boundary manifold connections between DFU and prostate In Panels **a-d.** DFU tissue biopsy IMC ROIs farthest from the center point of the ulcer that exhibit dense areas of immune cell populations (yellow, Iridium; cyan, CD3; magenta, CD68; scale bars = 80 μm). **e-h.** Prostate tumor tissue biopsy IMC ROIs showing sparse immune cell infiltration (yellow, Iridium; cyan, CD3; magenta, CD68; scale bars = 80 μm). * Contrast and intensity of channels in each image individually adjusted for visualization purposes.

coplin jar, volume for retrieval, and paraffin melting procedure all contributed to a more robust and reproducible method.

Beyond methodology improvements, the labeling of the tissue was also improved by the addition of several novel metal conjugated antibodies and a rapid ruthenium "stain." The ruthenium labeling combined with nuclear labeling can provide a pseudo-H&E image, which was important to meet critiques from colleagues in Pathology. The expanded panel now includes 45 antibodies in addition to ruthenium and DNA labels (Table 3).

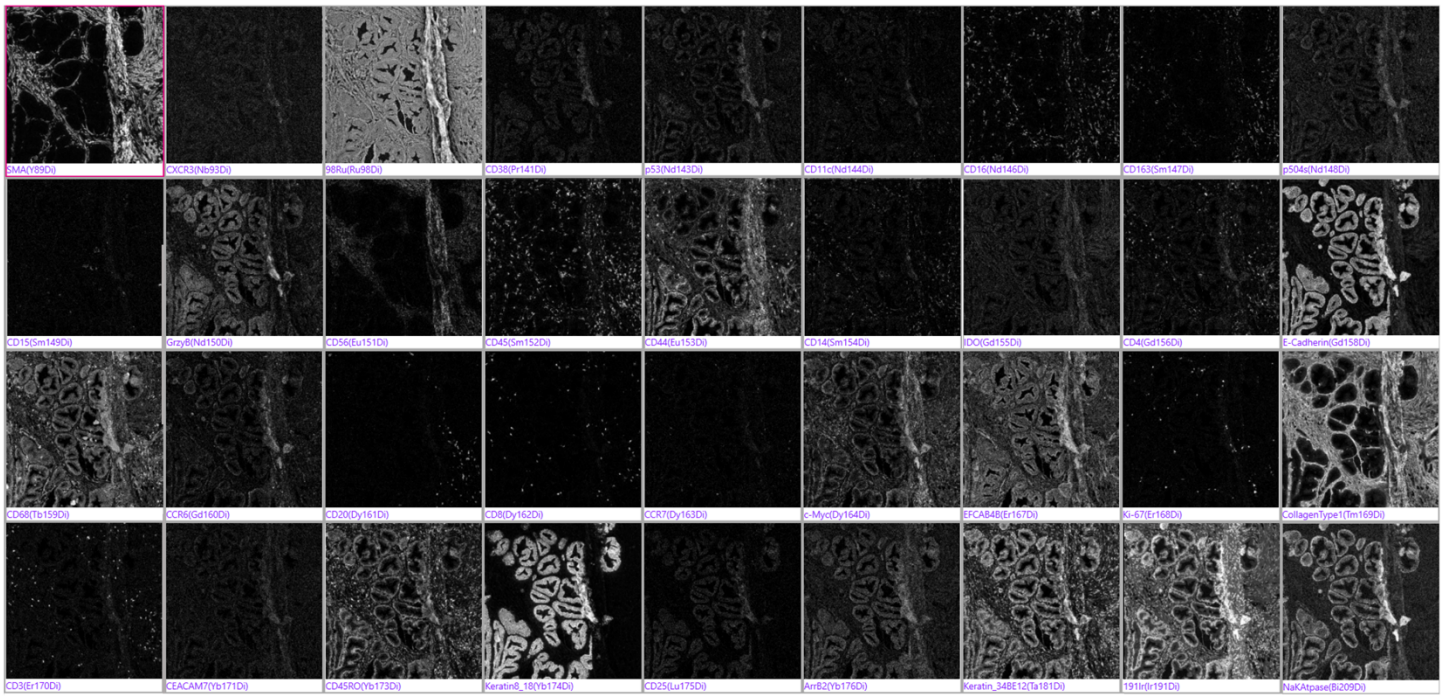


Figure 4 Imaging mass cytometry of prostate tumor. Representative image from imaging mass cytometry of a prostate cancer sample, with each panel showing an individual label, either for antibody, DNA, or tissue counterstain. For analysis of each PC biopsy core within the TMA, data from all channels is combined into one high-plex image.

Subtask 3: Conduct mass spectrometry imaging to generate images from PC tissue collection arranged as a Tissue Microarray (TMA).

Mass spectrometry imaging of the TMA was conducted in month 18 (December 2021) in agreement with the schedule. A second imaging was performed in month 20 (February 2022). However, the processing of the tissue by Imabiotech again resulted in degradation of the tissue architecture and complete loss of some samples - despite feedback from Dr. Seeley (Figure 5). As a result, the quality of the data from the subsequent imaging was poor, with decreased overall sensitivity and unfavorable signal to noise. At this stage we chose to work with Dr. Seeley directly, who was able to process the tissue in a satisfactory manner and data was obtained in March 2022 from the 48 core TMA. As we work to analyze those data, we are coordinating with Dr. Seeley to conduct imaging of the larger 96 core TMA.

Subtask 4: Computational image and clinical correlation analysis of imaging data from both mass spectrometry imaging and imaging mass cytometry

Computational and statistical analysis of images is ongoing and we expect completion in month 25 (July 2022), five months after the initial target of month 20 (February 2022).

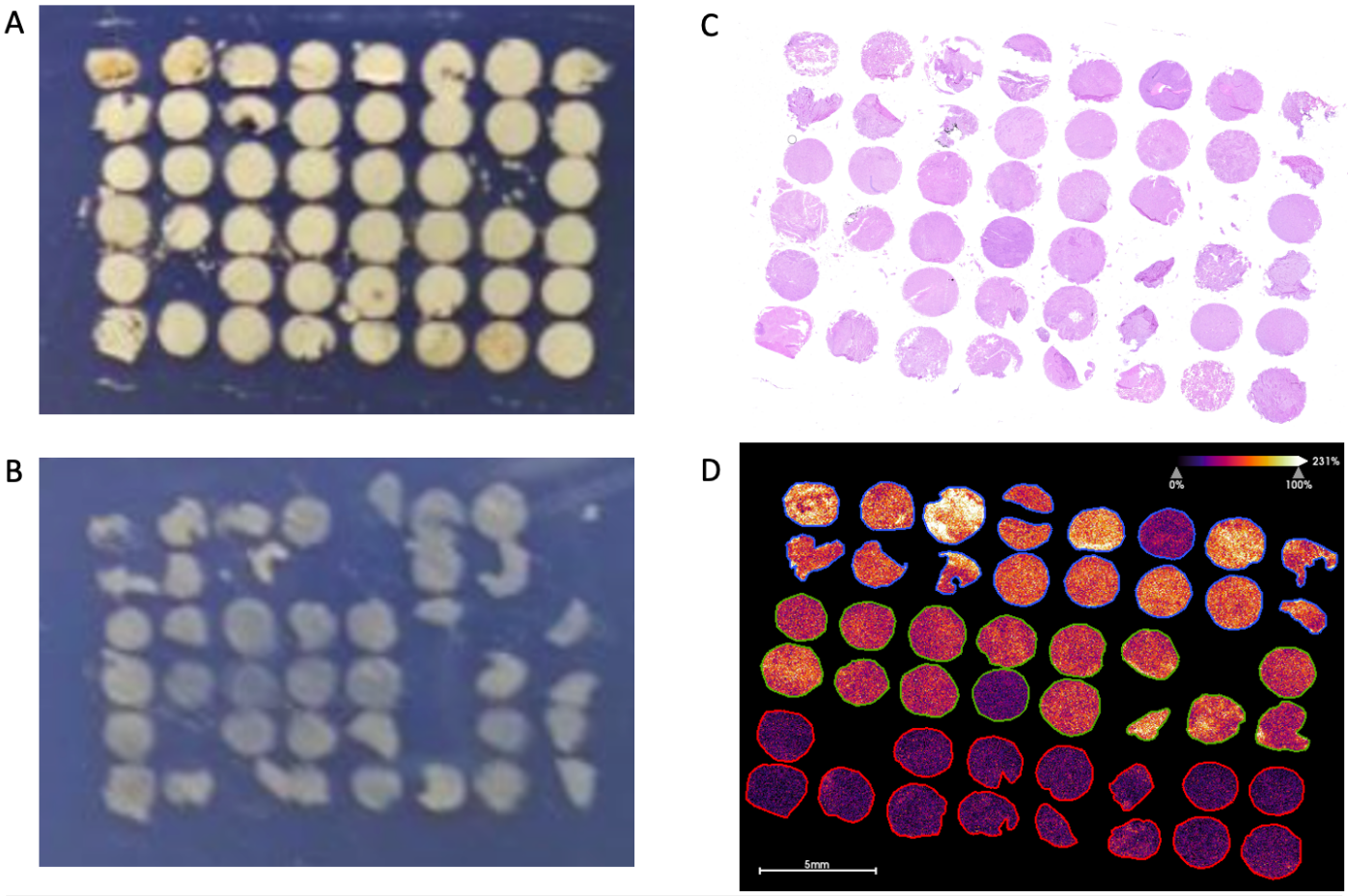


Figure 5 Optimization of mass spectrometry imaging of TMA A. TMA prior to antigen retrieval and digest. B. TMA following antigen retrieval and digest as performed by Imabiotech using the modified protocol developed in Aim 1 **C.** H&E of TMA after antigen retrieval and digest as performed by Dr. Seeley. **D.** Mass spectrometry imaging data obtained by Dr. Seeley, illustrating good spatial resolution

Subtask 5: Manuscripts & Reporting

A manuscript detailing the findings from Aim 2 is pending, with a target of month 27 (September 2022).

Other achievements

We had previously noted the use of our efforts to develop pseudo-H&E images generated from DNA and collagen signals. We have advanced that one step further using ruthenium as a "stain" to label all the cells in a tissue (Figure 6A), together with other markers, the addition of ruthenium labeling will allow us to create pseud0-H&E images that are highly similar to traditional H&E. This is important for engaging with pathologists and other colleagues who benefit from a familiar frame of reference when evaluating images.

Antibody labeling has also advanced due to the development of chemistries to enable conjugation of antibodies to isotopes that are not commercially available. Specifically, we have the capacity to conjugate antibodies of choice to 89Y, 113In, 115In, 157Gd, and 209Bi. This allows the team to implement antibody clones that improve the detection of proteins that can be difficult to label (e.g. CD56 and CD3, Figure 6B).

Though we continue to develop additional reagents to expand the number of antibody-isotope pairs available for imaging experiments, the enthusiasm for expanding the number of markers is restrained by the desire to continue implementing the current set of reagents to generate robust data to recoup time and accelerate the pace of the project. We have worked to enhance the reproducibility of the sample preparation by developing standardized protocols and adopting instrumentation that address or eliminate common sources of variance (e.g. antigen retrieval with a pressure-cooker which releases pressure at a fixed rate to prevent sudden boiling).

Metal Label	Antibody Target
89Y	SMA
102Ru	Ruthenium
114Cd	CD39
113In	CXCR3
115In	p504s (AMACR)
141Pr	CD38
142Nd	ERG
143Nd	p53
144Nd	CD11c
145Nd	P63
146Nd	CD16
147Sm	CD163
148Nd	p504s (AMACR)
149Sm	CD15
150Nd	GrzyB
151Eu	CD56
152Sm	CD45
153Eu	CD44
154Sm	CD14
155Gd	IDO
156Gd	CD4
157Gd	CD11b
158Gd	E-Cadherin
159Tb	CD68
160Gd	CCR6
161Dy	CD20
162Dy	CD8
163Dy	CCR7
164Dy	c-Myc
165Ho	CEACAM7
166Er	NKX3.1
167Er	EFCAB4B
168Er	Ki-67
169Tm	Collagen type 1
170Er	CD3
171Yb	MBD1
172Yb	MBD1
173Yb	CD45RO
174Yb	Keratin 8/18
175Lu	CD25
176Yb	ArrB2
192Ir	DNA
194Pt	PTEN
195Pt	Keratin
196Pt	VCL
198Pt	Keratin (34βE12)
209Bi	NaK ATPase

Table 3. 45 Antibody Panel

Working with Dr. Seeley, we have introduced efficiencies in the preparation of slides for mass spectrometry imaging. Another significant advantage of working with Dr. Seeley is her ability to prepare and image the samples in near real time. Previously working with ImaBiotech, the logisitics of shipping samples and coordinating imaging meant that samples could sit for several weeks prior to preparation and analysis. The prolonged storage time can adversely impact the sample integrity due to oxidation and degradation.

With respect to image analysis, we are working to evaluate recently published data pre-processing tools that can expedite analysis with MIAAIM. As other data analysis workflows for multiomics emerge, we continue to evaluate the performance of our workflow with respect to other tools. To date no algorithms or software packages are able to recapitulate, in whole or part, the analysis capacity of MIAAIM. <u>However, the priority is to implement MIAAIM as it stands to maximize focus and productivity for data analysis and publication generation</u>.

What opportunities for training and professional development has the project provided?

Nothing to report.

How were the results disseminated to communities of interest?

Nothing to report.

What do you plan to do during the next reporting period to accomplish the goals?

The overriding priority is to accelerate data generation for mass spectrometry imaging and data analysis for both imaging modalities. To date we have addressed a number of unanticipated technical hurdles that compounded supply chain and personnel challenges. Fortunately, these challenges have been costly in terms of time, though not with respect to budget. We remain sufficiently funded to complete the studies as described with the possibility to expand the number of samples imaged modestly due to efficiencies gained in sample preparation and reagent development. Even with increase spend rates on imaging, reagents, and personnel, we do anticipate the need for a no-cost extension to complete the studies. We will engage with the program official to discuss the details of a possible no-cost extension as we approach the mid-point of the final study period. The first priority is to re-submit the MIAAIM manuscript to

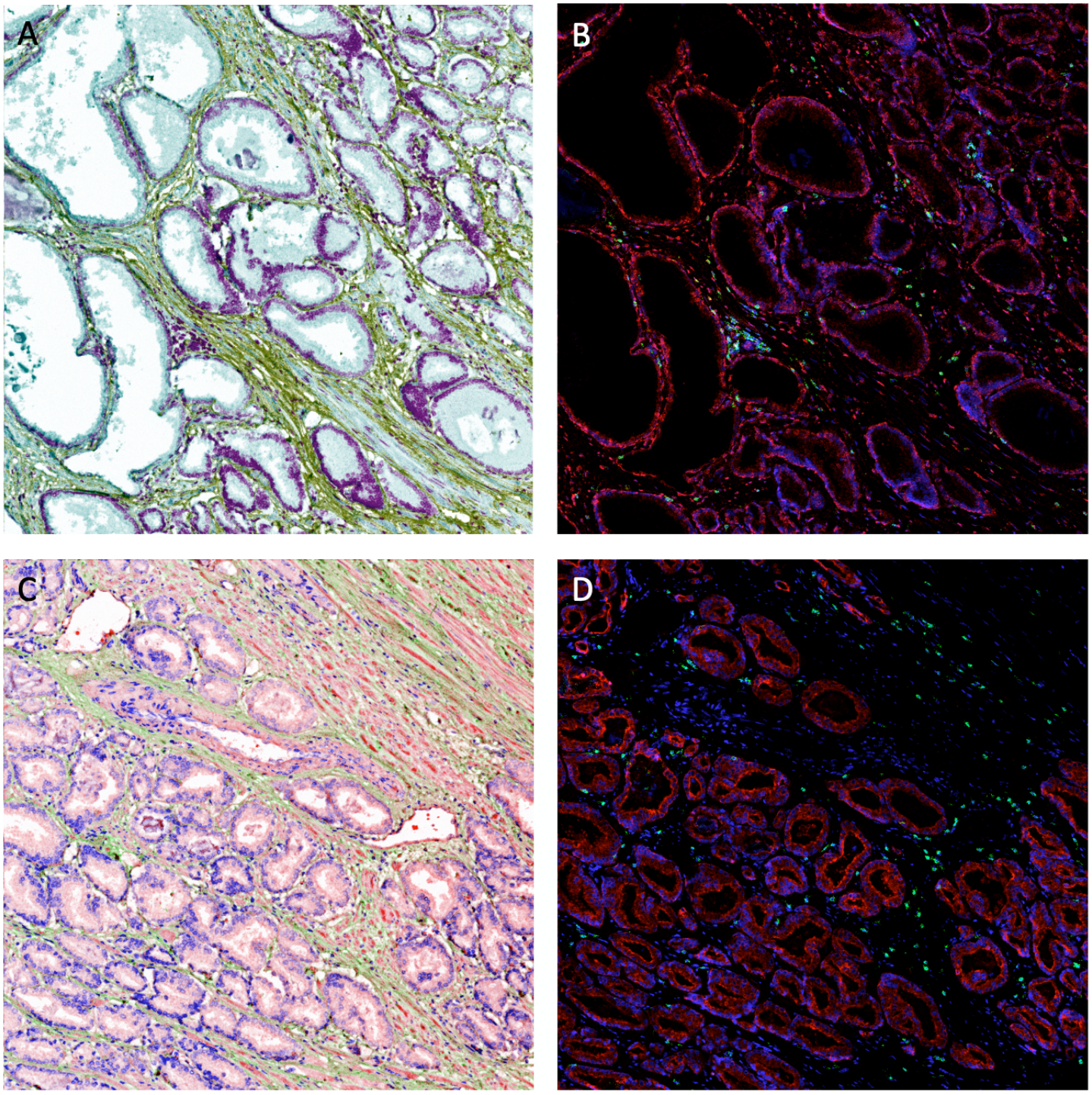


Figure 6. Imaging mass cytometry of prostate cancer tissue. A&C Image of prostate cancer section labeled with keratin (yellow), ruthenium (Cyan), to "counterstain" protein, and iridium labeled nuclei (blue) to enable visualization of the entire tissue akin to H&E staining. **B&D** Prostate section with labeled Keratin (Red), T cells (CD3 – Green), and nulcei (Iridium Blue).

Nature Methods and secure a high-impact publication from these studies. As noted above, we expect re-submission of the manuscript by July 15, 2022. The second priority is to close the gap on the completion of Aim 2, for which data is already generated and analysis is ongoing. Subsequent to completing data analysis, a manuscript will be prepared detailing the findings regarding signatures associated with tumor progression.

Subtask 4: Generate tissue microarray (TMA) containing biopsy cores from prostatectomy tissues. Subsequently update deidentified clinical data and patient outcomes.

In month 18, Dr. Wu made available a TMA containing tissue 80 biopsies appropriate for the objectives of Specific Aim 3. Clinical information is available for these cases is available will be updated in anticipation for imaging.

Subtask 2: Conduct imaging mass cytometry on TMA prepared from prostatectomy sample set

Imaging of the TMA for Aim 3 while efforts focus on ongoing work in the preceding Major Tasks. We anticipate completion by the end of month 27 (September 2022), as well as for the following mass spectrometry imaging task since sections for both methods will be cut and processed simultaneously.

We anticipate that the optimization and refinement of both the data acquisition and data analysis will create opportunities, budget permitting, to image additional TMAs to increase the statistical power and overall applicability of the findings.

4. IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

The work to date will both benefit the specific area of prostate cancer research as well as have broader applications to other areas of cancer research, imaging, and computational analysis.

The optimized methods for tissue preparation, including antigen retrieval, antibody labeling, enzymatic digestion, and matrix coating will aid future studies conducted as part of this project as well as work from other groups that may pursue similar research strategies. Similarly, details regarding methods to facilitate image acquisition by imaging mass cytometry and mass spectroscopy imaging will also be of value to the research community.

A major hurdle in the use of multi-modal imaging is the co-registration of images and data from each independent method. A significant portion of work during the first period focused on developing computational strategies that efficiently and robustly align high-dimensional imaging data. The analytical workflow currently under development enables the integration of multiple imaging methods to create one composite high-dimensional image. In addition, the analytical tools can be used to train predictive algorithms that can use one imaging method to reliably predict results from a second imaging method. The ability to make predictive associations is of particular value when not all imaging methods are not feasible to execute due to limitations in sample or instrumentation. The ongoing development of the workflow will also enable correlation of image features, groups of features, and ratios between features to clinical data. These correlations will be used to identify those components from imaging that are strongly linked to cancer grade scores and identify potential prognostic markers.

We anticipate that in the next project period a manuscript detailing the methods and computational tools developed will be prepared and submitted for publication. A second manuscript will specifically address the findings regarding the detailed analysis of prostate cancer biopsy and identification of features to enhance the discrimination of prostate cancer grade groups. Reporting of these findings will be of value to researchers that aim to improve their understanding of the biology of prostate cancer, and other cancers, or improve on prognostic tests to inform clinical practice.

What was the impact on other disciplines?

To date no other publications have cited the pre-print of the MIAAIM method paper that has been reviewed at Nature Methods. However, tracking data does show that researchers in mathematics, biology, and geography fields have followed or recommended the manuscript. Once published in a peer-reviewed journal, we anticipate that the study will have a more direct impact. Similarly, the publication currently in preparation which details several technical advances and as well as the biological insights related to prostate cancer, will inform analogous studies in other cancers.

What was the impact on technology transfer?

The patent filing for the MIAAIM has formed the basis for license agreement that was finalized in May of 2022. This license agreement will be used to form the basis of a start-up company that aims to adapt the technical and computational tools we have developed into clinical applicable commercial tests.

What was the impact on society beyond science and technology?

Nothing to report.

5. **CHANGES/PROBLEMS:**

Changes in approach and reasons for change

No changes in approach to report. We do anticipate a change in the mass spectrometry imaging vendor from ImaBiotech to the UT Austin Mass Spectrometry Imaging facility to overcome technical challenges, improve speed of data analysis, and reduce costs.

Actual or anticipated problems or delays and actions or plans to resolve them

Dr. Reeves had 1-month of parental leave to welcome a child. Although work continued during this time, Dr. Reeves' leave, which coincided with Mr. Hess' transition to graduate school, did transiently reduce project momentum. We do not foresee any additional leave for or team member departure in the coming period. Rather, two team members have joined, following extended candidate searches, who will accelerate the program.

The most significant problems and delays experienced during the report period result from supply chain, employee recruitment and instrumentation access constraints which can be attributed to the ongoing COVID19 pandemic. Though access to laboratory space is unimpeded within our institution, through the first half of the study period other institutions maintained restrictions on facility access that limited access to both the imaging mass cytometer and mass spectroscopy imager. Though improving, supply chain constraints for laboratory consumables, antibodies, and other reagents also resulting from COVID19 have also impacted execution of studies. The competition for qualified candidates, particularly those with computational or mathematical training was an unanticipated complication.

We have addressed these challenges in several ways. We have diversified our vendor pool and developed in house capacities to bypass reliance on the primary vendor of antibodies for imaging mass cytometry. To address technical and instrumentation access issues through ImaBiotech, we sought the expert advice of Dr. Erin Seeley. The facility Dr. Seeley directs at the University of Texas has recently opened to outside academic users and we anticipate conducting future research activities for mass spectrometry imaging with Dr. Seeley. Working with Dr. Seeley will improve both the speed data acquisition and quality of the data, without increasing budget obligations. In addition, building the collaborative relationship with Dr. Seeley will facilitate future studies from

our group on prostate cancer and provide opportunities to explore the latest instrumentation and advances in methodology, which is not currently possible with Imabiotech. Lastly, in October of 2021 we were able to obtain local access to an imaging mass cytometer, which significantly eases logistical challenges in conducting testing and analyses. As noted in the previous report, the significant investment in the data analysis pipeline has already facilitated the completion of a submitted manuscript and will ensure that ongoing and future analyses are completed in a timely manner.

We are committed to conducting and completing the studies in the most rigorous and timely manner possible. Every effort has been made to contain budget expenditures and maximize the overall success of the studies, even in the face of previous and ongoing challenges stemming from the COVID19 pandemic. We are happy to provide interim reports and updates to mitigate any concerns and address questions. Similarly, we anticipate the need for possible no-cost extension, the scope of which will rely on progress over the first half of the upcoming study period. We would propose evaluating a no-cost extension in more detail as we approach the mid-point of the final study period.

Changes that had a significant impact on expenditures

Ongoing issues related to the COVID19 pandemic and supply chain constraints led to reduced expenditures during the first portion of the second period. In addition, the technical challenges encountered with ImaBiotech delayed final billing for mass spectroscopy imaging services. The funds to pay for these services have been allocated internally, and the vendor submitted an invoice for the first half of the proposed costs. Similarly, charges for reagents and run time for use of the imaging mass cytometer were delayed. To mitigate

Lastly, recruiting scientific personnel with appropriate skills and availability is an issue facing many laboratories and scientific research groups – this project is no exception. Mr. Hess departed his full-time position on the team to start a PhD program. Fortunately, we were able to retain Mr. Hess on a part-time basis, enabling him to provide his expertise as needed. This was essential, as he is first author on the current Nature Methods manuscript and instrumental in onboarding two new team members. We were also able to recruit two additional technical staff to support the generation and analysis of the imaging data. This strategy of having dedicated technical staff for each of the imaging modes is made possible by the carryover from the reduced spending in year 1 and will accelerate the pace of work in the months ahead.

As discussed above, we have taken several steps to offset the impact of delays in acquiring reagents and instrument time, we have worked to develop relationships with a number of vendors, particularly for antibodies, and developed the capacity to conjugate antibodies in house. This important advance frees us from reliance on a single vendor for antibodies, which had caused delays previously. Also, the ability to consult with Dr. Seeley has greatly improved the quality and ease of preparing tissue for mass spectrometry imaging. We do anticipate direct engagement with Dr. Seeley as a means to both accelerate the pace of work and contain costs associated with mass spectrometry imaging. Reflective of the improved access to reagents, instrumentation, and on boarding of staff, the pace of expenditures increased significantly in the second half of year 2. Provided that remaining fund from the second year are carried over the remaining funds will be sufficient for completing work for year 2 & 3 tasks.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to report.

Significant changes in use or care of human subjects

Nothing to report.

Significant changes in use or care of vertebrate animals.

Nothing to report.

Significant changes in use of biohazards and/or select agents

Nothing to report.

6. **PRODUCTS:**

Publications, conference papers, and presentations

Journal publications.

As described above we have submitted a manuscript titled “MIAAIM: Multi-omics image integration and tissue state mapping using topological data analysis and cobordism learning”, to Nature Methods detailing the image analysis methodology and some of the initial biological analysis to Nature Methods. The reviewers’ and editor’s comments were generally supportive; however, to recommend publication it was asked that we address specific technical questions and expand the analysis to more clinical samples. To address this, we are now working to include the analysis from Major Task 2/Aim 1 which incorporates an 84-core TMA. Re-submission to Nature Methods is expected by July 2022.

Books or other non-periodical, one-time publications.

Nothing to report.

Other publications, conference papers, and presentations.

Nothing to report.

Website(s) or other Internet site(s)

Nothing to report.

Technologies or techniques

Nothing to report.

Inventions, patent applications, and/or licenses

A provisional patent application for the MIAAIM image analysis method was filed. This application formed the basis for a license agreement which was completed in May 2022. The license agreement will lead to the formation of a start-up company that aims to apply the developed methods in a commercial context generate tests to support clinical decision making for biopsied tissues.

Other Products

Nothing to report.

5. **PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS**

What individuals have worked on the project?

Name:	*Patrick Reeves, PhD*
Project Role:	*Principal Investigator*
Researcher Identifier:	*ORCID 0000-0001-8202-0723*
Nearest person month worked:	4 months
Contribution to Project:	*Coordinated and execute administrative tasks including IRB and HRPO approvals for samples, reagent and vendor purchases, collaborator, and team member efforts. Research efforts included experimental and antibody panel design, assay evaluation, development of analysis workflows, and generation of draft manuscript for data analysis methodology.*
Funding Support:	

Name:	*Chin-Lee Wu, MD PhD*
Project Role:	*Co-investigator*
Researcher Identifier:	*ORCID 0000-0002-0444-0068*
Nearest person month worked:	*0.6 months*
Contribution to Project:	*Worked on project idea development, study design including antibody selection, and prostate cancer sample selection*
Funding Support:	

Name:	*Mark C. Poznansky, MD PhD*
Project Role:	*Co-investigator*
Researcher Identifier:	*ORCID 0000-0003-1344-7103*
Nearest person month worked:	*0.6 months*
Contribution to Project:	*Advised on project coordination, study management, as well as clinical data and sample selection.*
Funding Support:	

Name:	*Ann Sluder, PhD*
Project Role:	*Project manager*

Researcher Identifier:	*ORCID 0000-0003-4415-0652*
Nearest person month worked:	*0.6 months*
Contribution to Project:	*Supported directly through project management, administrative and financial tracking. Advised on project coordination and vendor discussions.*
Funding Support:	

Name:	*Josh Hess*
Project Role:	*Technician*
Researcher Identifier:	
Nearest person month worked:	*6 months*
Contribution to Project:	*Development of analysis of computational data analysis tools and design of analysis workflow. Generation of draft manuscript related to data analysis methodology.*
Funding Support:	

Name:	*Shulin Wu, MD PhD*
Project Role:	*Research Fellow*
Researcher Identifier:	
Nearest person month worked:	*3 months*
Contribution to Project:	*Worked on selection and preparation prostate biopsy sample, image collection, data collection and annotation.*
Funding Support:	

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to report.

What other organizations were involved as partners?

Nothing to report.

6. **SPECIAL REPORTING REQUIREMENTS**

Not applicable.

7. **APPENDICES:**

Attached is a manuscript, “MIAAIM: Multi-omics image integration and tissue state mapping using topological data analysis and cobordism learning”, which was reviewed at Nature Methods and a pre-print is available on BioRxiv (DOI: 10.1101/2021.12.20.472858. The revised version will be submitted as update made available upon re-submission to the journal.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MIAAIM: Multi-omics image integration and tissue state mapping using topological data analysis and cobordism learning

Joshua M. Hess[1,2], Iulian Ilieş[3], Denis Schapiro[4,5,6,7], John J. Iskra[1,8], Walid M. Abdelmoula[9,10], Michael S. Regan[9], Georgios Theocharidis[11], Chin Lee Wu[12], Aristidis Veves[11], Nathalie Y.R. Agar[9,13], Ann E. Sluder[1], Mark C. Poznansky[1*+], Patrick M. Reeves[1*+], Ruxandra F. Sîrbulescu[1*+]

[1]Vaccine and Immunotherapy Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

[2]Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA

[3]Healthcare Systems Engineering Institute, Northeastern University, Boston, MA, USA.

[4]Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA.

[5]Broad Institute of Harvard Medical School and Massachusetts Institute of Technology, Boston, MA, USA.

[6]Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University Hospital and Heidelberg University, Heidelberg, Germany

[7]Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany

[8]Department of Mathematics, Emory & Henry College, Emory, VA

[9]Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[10]Current address: Invicro, LLC, Boston, MA, USA

[11]The Rongxiang Xu Center for Regenerative Therapeutics, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA.

[12]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

[13]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[*]RFS, PMR and MCP contributed equally to this manuscript

[+]Corresponding authors

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Corresponding authors:

Ruxandra F. Sîrbulescu, Vaccine and Immunotherapy Center, 149 13th St, Charlestown, MA 02129; Telephone: 617-953-2316; email: rsirbulescu@mgh.harvard.edu

Patrick M. Reeves, Vaccine and Immunotherapy Center, 149 13th St, Charlestown, MA 02129; Telephone: 617-726-3812; email: patrick_reeves@hms.harvard.edu

Mark C. Poznansky, Vaccine and Immunotherapy Center, 149 13th St, Charlestown, MA 02129; Telephone: 617-724-6375; email: mpoznansky@mgh.harvard.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

High-parameter tissue imaging enables detailed molecular analysis of single cells in their spatial environment. However, the comprehensive characterization and mapping of tissue states through multimodal imaging across different physiological and pathological conditions requires data integration across multiple imaging systems. Here, we introduce MIAAIM (Multi-omics Image Alignment and Analysis by Information Manifolds) a modular, reproducible computational framework for aligning data across bioimaging technologies, modeling continuities in tissue states, and translating multimodal measures across tissue types. We demonstrate MIAAIM's workflows across diverse imaging platforms, including histological stains, imaging mass cytometry, and mass spectrometry imaging, to link cellular phenotypic states with molecular microenvironments in clinical biopsies from multiple tissue types with high cellular complexity. MIAAIM provides a robust foundation for the development of computational methods to integrate multimodal, high-parameter tissue imaging data and enable downstream computational and statistical interrogation of tissue states.

INTRODUCTION

Increased availability of multiplexed imaging at cellular and subcellular resolution allows multiparameter molecular descriptions of organelle, cell, and tissue organization in a preserved spatial context[1-6]. A variety of high-parameter imaging technologies have been developed, each capturing individual aspects of multi-faceted tissue representations – ranging from untargeted methods such as mass spectrometry imaging[7] (MSI) and spatial transcriptomics[8, 9], which provide comprehensive molecular profiles, to targeted, multiplexed antibody-based labeling approaches[1-6]. Compared to advances in data acquisition[10, 11] and atlas mapping for single-cell multi-omics data in dissociated tissue[12, 13], analogous computational methods that preserve the spatial structure of tissues remain underdeveloped. Tissue cytoarchitecture and molecular composition are crucial for the comprehensive characterization of biological responses under different physiological conditions or disease states. Most importantly, multi-omics tissue portraits, including single-cell composition and spatial structure, could be used to infer equilibrium and transition states of tissue health and disease.

The integration of data across multiple imaging technologies can present significant challenges, including data set size and high dimensionality. In recent years, methods have been developed to align multimodal imaging data as it becomes increasingly relevant to analyze data acquired across multiple modalities[14, 15]. To date, no approach to transform data from separate imaging modalities into the same

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coordinate system that generalizes to the growing repertoire of high-parameter imaging systems has been developed. Dimensionality reduction is routinely used to model single-cell data, showing that generating functions of cellular phenotype and gene expression closely follow a nonlinear manifold structure. While manifold-based representations of single-cell systems enable cell state trajectory inference[16-18], tissue states emerging from cellular behavior are often discretized. Clustering is typically used to partition cells into distinct populations[19, 20] (i.e., cell types), and tissue states are defined based on overall abundances of cell types. We suggest that nonlinear mapping of a systems-level state space may better enable descriptions of continuous transitions from health to disease. For tissue atlases, such as the Human Tumor Atlas Network[21] and the HuBMAP Consortium[22], the transfer of information across tissue states may be used for downstream analysis, as is routinely done in dissociated tissue mapping. Comparable intact tissue mapping calls for a reproducible data integration framework that can be implemented at multiple institutions across computing architectures.

Here, we describe MIAAIM (Multi-omics Image Alignment and Analysis by Information Manifolds), a modular framework consisting of 4 new software packages (see **Methods**, **Key Resources Table**) with reproducible workflows to align, model, and transfer data across high-parameter imaging technologies. MIAAIM contains a manifold-based image registration approach, designed to compress and integrate data with diverse densities and spatial resolutions, that can be generalized to multiple high-parameter bioimaging systems. In addition, we introduce *cobordism learning* as a dimension reduction and domain transfer method that models phenotypic manifolds (i.e., tissue states) as the boundary of a higher-dimensional manifold, called a cobordism[23], that characterizes continuous transitions between tissue states. We demonstrate the applicability of MIAAIM using multimodal imaging data derived from a combination of mass spectrometry-based imaging technologies, including imaging mass cytometry (IMC) and mass spectrometry imaging (MSI), in addition to classical histology. Using this method, we investigate diverse cellular and molecular landscapes in tissue samples with high cellular complexity, including human tonsillar, diabetic foot ulcer (DFU) and prostate cancer biopsies.

RESULTS

MIAAIM workflow

MIAAIM is a sequential workflow aimed at providing comprehensive portraits of tissue states. It includes 4 processing stages: (i) image preprocessing and compression with the high-dimensional image preparation (HDIprep) workflow, (ii) image registration with the high-dimensional image registration

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(HDIreg) workflow, (iii) tissue state modeling with cobordism approximation and projection (PatchMAP), and (iv) cross-modality information transfer with i-PatchMAP (**Fig. 1**). Image integration in MIAAIM begins with two or more assembled images (level 2 data)[24, 25] or spatially resolved data sets (**Fig. 1,** assembled images). The size and standardized format of assembled images vary by technology. For example, cyclic fluorescence-based methods (e.g., CODEX[3], CyCIF[5, 6]) assemble BioFormats/OME-compatible[26] 20-60-plex full-tissue images after correcting uneven illumination (e.g., BaSiC[27]) and stitching tiles (e.g., ASHLAR[28]); other methods acquire 20 to 100-plex data within specified regions of interest (ROIs) (e.g. MIBI[1], IMC[2]) in the OME format. Additional methods quantify thousands of parameters at rasterized locations on full tissues or ROIs and are not stored in BioFormats/OME-compatible formats (e.g., the imzML format[29] often stores MSI data). MIAAIM accepts a variety of input data formats (**Fig. 1**), and its modular design will enable it to incorporate additional formats tailored to new imaging modalities.

Regardless of technology, assembled images contain high numbers of heterogeneously distributed parameters, which precludes comprehensive, manually-guided image alignment. In addition, high-dimensional imaging produces large feature spaces that challenge methods used in unsupervised settings. The HDIprep workflow in MIAAIM generates compressed images that preserve multiplex salient features to enable cross-technology statistical comparison while minimizing computational complexity (**Fig. 1,** HDIprep). For images acquired from histological staining, HDIprep provides parallelized smoothing and morphological operations that can be applied sequentially for preprocessing. Image registration with HDIreg produces transformations to combine modalities within the same spatial domain (**Fig. 1**, HDIreg). HDIreg uses Elastix[30], a parallelized image registration library to calculate transformations, and is optimized to transform large multichannel images with minimal memory use, while also supporting histological stains. HDIreg automates image resizing, padding, and trimming of borders prior to applying image transformations.

Aligned data are well-suited for established single-cell and spatial neighborhood analyses[31] and they can be segmented to capture multimodal single-cell measures (level 3 and 4 data)[24, 25], such as average protein expression or spatial features of cells, or analyzed at the pixel level. A common goal in pathology is to use composite tissue portraits to map healthy-to-diseased transitions. To address this, similarities between systems-level tissue states can be visualized with the PatchMAP workflow (**Fig. 1,** PatchMAP). PatchMAP models tissue states as phenotypic manifolds that are stitched to form a higher-dimensional cobordism. The result is a nested model capturing nonlinear intra-system states and cross-system

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

continuities. The cobordism paradigm can be applied as a tissue-based atlas-mapping tool to transfer information across modalities with i-PatchMAP (**Fig. 1,** i-PatchMAP).

MIAAIM's algorithms are nonparametric, using probability distributions supported by manifolds rather than training data models. MIAAIM is therefore technology-agnostic and generalizes to multiple imaging systems (**Supplementary Table 1**). Nonparametric image registration, however, is an iterative, parameter-tuning process rather than a "black-box" solution. This creates a substantial challenge for reproducible data integration across institutions and computing architectures. We therefore Docker[32] containerized MIAAIM's data integration workflows and developed a Nextflow[33] implementation to document human-in-the-loop processing and remove language-specific dependencies in accordance with the FAIR (finable, accessible, interoperable, and reusable) data stewardship principles[34]. In the following sections, we describe algorithms unique to MIAAIM and their application to multimodal imaging data sets across technologies and diverse tissue specimens.

High-Dimensional Image Compression with HDIprep

To compress high-parameter images, HDIprep performs dimensionality reduction on pixels using Uniform Manifold Approximation and Projection (UMAP)[35] (**Fig. 2a**). We performed a stringent comparison of dimensionality reduction algorithms using imaging data sets of human DFU, prostate cancer, and tonsil tissue biopsies acquired using MSI, IMC, and hematoxylin and eosin (H&E). Based on dimensionality reduction benchmarks, UMAP consistently outperformed competing algorithms in its robustness to noise and ability to efficiently preserve data complexity while capturing morphological structure (**Supplementary Figs. 1, 2, 3**).

HDIprep retains data complexity with the fewest degrees of freedom necessary by detecting *steady-state manifold embeddings*. To identify steady-state dimensionalities, information captured by UMAP pixel embeddings is computed (**Methods**, **Definition 1,** cross-entropy) across a range of embedding dimensionalities, and the first dimension where the observed cross-entropy approaches the asymptote of an exponential regression fit is selected. Steady state embedding calculations scale quadratically with the number of pixels, HDIprep therefore embeds spectral landmarks in the pixel manifold representative of its global structure (**Supplementary Fig. 4)**.

Pixel-level dimensionality reduction is computationally expensive for large images, i.e., at high resolution (e.g., ≤ 1 μm/pixel). To reduce compression time while preserving quality, we developed a subsampling approach to embed a spatially representative subset of pixels prior to spectral landmark

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection and project out-of-sample pixels into embeddings (**Supplementary Figs. 5, 6, 7)**. HDIprep combines all optimizations with a neural-network UMAP implementation[36] to scale to whole-tissue images. We demonstrate its efficacy on publicly available[37] 44-channel CyCIF images containing ~100 and ~256 million pixels (**Supplementary Fig. 8).** Thus, HDIprep presents an objective, pixel-level compression method applicable to multiple modalities (**Methods, Algorithm 1**).

High-Dimensional Image Registration (HDIreg)

MIAAIM connects the HDIprep and HDIreg workflows with a manifold alignment scheme parametrized by spatial transformations. To align pixel manifolds, we extended methodology to compute manifold α-entropy[38, 39] using entropic graphs to UMAP embeddings and adapted it to image registration using the entropic graph-based Rényi α-mutual information (α-MI)[40] (**Methods, HDIreg**). HDIreg selects for a transformation that maximizes image-to-image (manifold-to-manifold) α-MI (**Fig. 2b**). The entropic-graph based α-MI generalizes to arbitrary dimensionalities by considering distributions of k-nearest neighbor (KNN) graph lengths of compressed (i.e., embedded) pixels, rather than directly comparing the pixels themselves. Combining HDIprep compression with KNN α-MI extends intensity-based registration to images across technologies, without the need to match specific markers or analytes.

MIAAIM generates information on cellular phenotype, molecular ion distribution, and tissue state across scales

To demonstrate the capacity of MIAAIM to generate linked data between multimodal imaging with various scales and data densities, we applied the HDIprep and HDIreg workflows to MALDI-TOF MSI, H&E and IMC data from a DFU tissue biopsy containing a spectrum of tissue states, from the necrotic center of the ulcer to the healthy margin. Image acquisition covered 1.2 cm^2 for H&E and MSI data. Molecular imaging with MSI enabled untargeted mapping of lipids and small metabolites in the 400-1000 *m/z* range across the specimen at a resolution of 50 μm/pixel. Tissue morphology was captured with H&E at 0.2 μm/pixel, and 27-plex IMC data was acquired at 1 μm/pixel resolution from 7 ROIs on an adjacent section.

Cross-modality alignment was performed in a global-to-local fashion **(Fig. 2c)**. We utilized HDIprep compression for high-parameter data and HDIreg manifold alignment for registration of compressed images. Due to the destructive nature of IMC acquisition[2], we aligned full-tissue data from MSI, H&E (down-sampled to approximately 3.5 μm/pixel), and an IMC reference image first.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Deformations not captured at the full-tissue scale within each ROI were corrected using manual landmark guidance. Tissue deformations resulting from serial sectioning were accounted for with nonlinear transformations. Registrations were initialized by affine transformations for coarse alignment before nonlinear correction. Resolution differences were accounted for with a multiresolution smoothing scheme[30]. Final alignment proceeded by composing modality and ROI-specific transformations.

After locating single-cells using image segmentation with Ilastik[41], single-cell parameter quantification with MCMICRO[24], and antibody staining quality control, registered images yielded the following information for 7,114 cells: (i) average expression of 14 proteins (**Methods, Key Resources Table**) including markers for lymphocytes, macrophages, fibroblasts, keratinocytes, and endothelial cells, as well as extracellular matrix proteins, such as collagen and smooth muscle actin; (ii) morphological features, such as cell eccentricity, solidity, extent, and area, spatial positioning of each cell centroid; and (iii) the distribution of 9,753 *m/z* MSI peaks across the full tissue. Distances from each MSI pixel and IMC ROI to the center of the ulcer, identified by manual inspection of H&E, were also quantified. Through the integration of these modalities, MIAAIM provided cross-modal information that could not be gathered with a single imaging system, such as profiling of single-cell protein expression and microenvironmental molecular abundance.

Identification of molecular microenvironmental niches correlated with cell and disease states

We verified the existence of cross-modal associations from the integrated multimodal imaging data set by conducting a *microenvironmental correlation network analysis* (MCNA) on registered IMC and MSI data (**Fig. 3**). We performed community detection (i.e., clustering) on MSI analytes (*m/z* peaks) based on their correlations to single-cell protein measures and defined *microenvironmental correlation network modules* (MCNMs; different colors in **Fig. 3a**). Inspection of MCNMs with top correlations to protein levels identified with IMC revealed that sets of molecules, rather than individual peaks, were associated with cellular protein expression (**Fig. 3b**). MCNMs organized on an axis separating those with moderate positive correlations to cell markers indicative of inflammation and cell death (CD68, activated Caspase-3) and those with moderate positive correlations to markers of immune regulation (CD163, CD4, FoxP3) and vasculature (CD31). Some proteins, such as the myeloid marker CD14 and the cell proliferation marker Ki-67, were not strongly correlated with any of the detected *m/z* peaks across all cells.

To gain insights into the association of molecular distributions with tissue health, we plotted ion intensity distributions of MCNMs with respect to their proximity to the center of the ulcer (**Fig. 3c**). This

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis revealed a change in molecular profiles about 6 mm from the ulcer center point, as the tissue state progressed from healthy to injured. We validated our observations and the performance of HDIreg to align micron-scale structures by visualizing the distribution of top correlated ions within cellular microenvironments (**Figs. 3d, 3e**).

An advantage of the MIAAIM approach is the capacity to identify molecular variations in one modality (here, MSI) that correlate with cell states identified using a different modality (here, IMC). We investigated whether *m/z* peaks differentially associate with cell proliferation (Ki-67 marker in IMC). We identified cell phenotypes through unsupervised clustering on IMC segmented cell-level expression patterns (**Fig. 3f**) and conducted a differential correlation analysis between phenotypes within the well-separated CD3+ cluster, which likely identifies infiltrating T cells at the wound site, and the CD3- cell populations (**Fig. 3g**). Interestingly, we found that correlations to Ki-67 expression shifted with near significance (2σ) for multiple *m/z* peaks (Fisher transformed, one-sided z-statistics; Bonferroni corrected P-values) between CD3− populations and the CD3+ population (**Fig. 3h**).

We then utilized spatial context preserved with MIAAIM and observed that ion intensities of *m/z* peaks positively correlated to Ki-67 in CD3+ cells increased with distance from the wound, while molecules with Ki-67 negative correlations specific to CD3+ cells showed the opposite trend (**Fig. 3i**). This distribution of Ki-67 correlates suggests that proliferation of CD3+ T cells occurs predominantly near the healthy margin of the DFU and confirms that molecular correlates of T-cell proliferation can be identified through this unbiased analysis. Collectively, these results provide insights to molecular microenvironments associated with different functional and metabolic states of cell subtypes, and how these microenvironments are distributed in the spatial context in a gradient from injured to healthy tissue.

Mapping tissue states via cobordism approximation and projection (PatchMAP)

To model transitions between tissue states, such as healthy or injured, we generalized manifold learning to higher-order instances by developing a new algorithm, called PatchMAP, that integrates mutual nearest-neighbor calculations with UMAP **(Fig. 4a** and **Methods, Algorithm 2)**. We hypothesized that state spaces of system-level transitions could be represented as nonlinear manifolds. Accordingly, PatchMAP represents the disjoint union of phenotypic manifolds as the boundary of a nonlinear cobordism that maps the transition between phenotypic manifolds. Patches across boundary manifolds are connected by pairwise nearest-neighbor queries that represent cobordism geodesics and stitched using the t-norm to make their metrics compatible. Interpreting PatchMAP embeddings is analogous to existing

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dimensionality reduction algorithms – similar data within or across boundary manifolds are located close to each other, while dissimilar data are farther apart. PatchMAP incorporates both boundary manifold topological structure and continuities across boundary manifolds to produce cobordisms.

Currently, no machine learning method to form cobordisms exists – methods closest to doing this are single-cell data integration algorithms[13, 42, 43]. Therefore, to benchmark PatchMAP's manifold stitching, we compared it to the data integration methods BBKNN[43], Seurat v3[13], and Scanorama[42] in a stitching simulation using the "digits" machine learning method development data set (**Fig. 4b**). We split data by label into boundary manifolds, then applied each method to re-stitch the full data set. In this task, perfect stitching would result in the separation of projected boundary manifolds, which we quantified with the silhouette coefficient (SC). For controlled visualization, UMAP was used after data integration to provide analogous embeddings to PatchMAP's for all benchmark methods.

PatchMAP outperformed data integration methods at higher nearest-neighbor (NN) counts in the stitching simulation (**Fig. 4b**). All other methods incorrectly mixed boundary manifolds when there was no overlap, as expected given that lack of manifold connections violated their assumptions. By contrast, PatchMAP's stitching uses a fuzzy set intersection[35], which prunes incorrectly connected data across manifolds while reinforcing correct connections. We also validated that PatchMAP preserves boundary manifold organization while embedding higher-order structures between similar boundary manifolds **(Supplementary Fig. 9)**. At low NN values and when boundary manifolds are similar, PatchMAP resembles UMAP projections (**Supplementary Fig. 9**). At higher NN values, manifold annotations are strongly weighted, which results in less mixing and better manifold separation.

Information transfer across imaging technologies and tissues (i-PatchMAP)

We posited that the transfer of information across biological states should account for continuous transitions and be robust to manifold connection strength (including the lack thereof). The i-PatchMAP workflow therefore uses PatchMAP as a paired domain transfer and quality control visualization method to propagate information between different imaging methods (**Fig. 4a,** information transfer). To transfer information across data sets, i-PatchMAP first computes one-step Markov chain transition probabilities between boundary manifolds of "*reference*" and "*query*" data along cobordisms (**Fig. 4a,** transition probabilities), and then linearly interpolates measures from reference to query data (**Fig. 4a,** propagate information). Quality control of i-PatchMAP can be performed by visualizing connections between boundary manifolds in PatchMAP embeddings (**Fig. 4a,** visualize manifold connections).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To benchmark i-PatchMAP, we compared it to other nonparametric domain transfer tools, Seurat v3 and a modification of UMAP incorporating a transition probability-based interpolation similar to i-PatchMAP (UMAP+), first on data multimodal single-cell data from DFU ROIs (see **Fig. 3**) and then separately, using a publicly available cord blood mononuclear cell (CBMC) CITE-seq data set[11]. UMAP+ utilized a directed NN graph from query to reference data for data interpolation, rather than PatchMAP's metric-compatible stitching, therefore acting as a control for PatchMAP. We tiled ROIs from to construct 23 evaluation instances and performed leave-one-out cross-validation using single-cell MSI profiles to predict IMC protein expression. We assessed accuracy using Spearman's correlation between the predicted spatial autocorrelation (Moran's I) and the true autocorrelation for each parameter to account for resolution differences between imaging modalities (**Fig. 4c**). i-PatchMAP outperformed tested methods on its ability to transfer IMC measures to query data based on MSI profiles (**Fig. 4d**). However, all methods demonstrated consistently poor performance for parameters with no original spatial autocorrelation within tiles (TGF-β, FoxP3, CD163). For the CITE-seq data set, we created 15 evaluation instances and used single-cell RNA profiles to predict antibody derived tag (ADT) abundance. We quantified performance using Pearson's correlation between true and predicted ADT values (**Fig. 4e**), finding that i-PatchMAP performed as well or slightly better than other tested methods for all parameters.

i-PatchMAP transfers multiplexed protein distributions across tissues based on molecular microenvironmental profiles

To assess whether i-PatchMAP can be used to transfer molecular signature information across imaging modalities and further, across different tissue samples, we used single-cell IMC/MSI protein measures (see **Fig. 3**) to extrapolate IMC information from discrete ROIs to the full DFU sample, as well as to distinct prostate tumor and tonsil specimens, based on MSI profiles. A PatchMAP embedding of single cells in DFU ROIs and individual pixels across tissues based on MSI parameters revealed that single-cell molecular microenvironments in the DFU ROIs provided a good representation of the overall DFU molecular profile (**Fig. 4f**). Accordingly, we used i-PatchMAP to transfer DFU single-cell protein measures from the ROIs to the full DFU tissue based on molecular similarities (**Fig. 4g**). i-PatchMAP predicted that the wound area of the DFU tissue would show high expression levels for CD68, a marker of pro-inflammatory macrophages, and activated Caspase-3, a marker of apoptotic cell death. By contrast, the healthy margin of the DFU biopsy was predicted to contain higher levels of CD4, indicating infiltrating T cells, and the cell proliferation marker Ki-67. Interestingly, the PatchMAP visualization revealed that

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

molecular microenvironments corresponding to specific single-cell measures in the DFU (e.g., CD4) were strongly connected with MSI pixels in the tonsil tissue sample (**Fig. 4h**). In the tonsil tissue, which is lymphocyte rich, i-PatchMAP predictions for CD4 matched well with follicular structure, and areas that lacked cell content were accurately predicted to not contain CD4. By contrast, there were no strong connections between the molecular profiles of the DFU biopsy and the prostate cancer sample, which showed poor immune cell infiltration by IMC examination (**Supplementary Fig. 10**). Thus, in the current datasets, strong inter-sample cellular and molecular connections are supported by the presence or absence of common immune cell populations.

DISCUSSION

We developed MIAAIM, a modular, reproducible computational framework to align, analyze, and transfer information across high-parameter bioimaging modalities. With MIAAIM we have demonstrated that highly detailed and mathematically grounded structures of high-dimensional bioimaging modalities can be linked and propagated across tissues, tissue states, and modalities. As part of MIAAIM, we have developed a manifold learning algorithm based on cobordism theory to translate multiplexed information across tissues. Our combined IMC-MSI approach provided insights into how molecular microenvironmental niches of individual cells may correlate with their states and localization within tissue sections. We showed MIAAIM's generalizability across diverse healthy and pathological human tissue types and imaging technologies, including those that generate large, whole-tissue datasets.

In its current version, MIAAIM does not incorporate deep learning, which would require large training sets across multiple imaging technologies for generalized application. To our knowledge, no learning models for multiplexed image registration exist, likely due to the large degree of variability in tissue organization across specimens. Similar limitations likely restrict generalized application of domain transfer neural networks for antibody-free labeling of growing repositories of publicly available images. Improvements in deep learning robustness may make its application routine in these settings – a key will be limiting model complexity through appropriate parametrization and regularization. MIAAIM currently provides a first framework to generate standardized training sets for the development of such methods. Its modular implementation will enable versatile workflow incorporation of new method developments such as those from the deep learning community.

With larger sample sizes, our multimodal imaging approach is well positioned to identify novel biomarkers to support diagnosis and prognosis in clinical cases in a disease-agnostic manner. PatchMAP

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

applied to single-cell data may provide a coordinate system on a manifold *between* single-cell network states (i.e., boundary manifolds) that could be used to infer a "single-cell network phenotypic tree" (i.e., cobordism). Computational views of biological systems have proven to be useful[44]; cobordism theory in this context may indicate a deeper relevance of bridging pure mathematics with modern biology. Indeed, connections between cobordisms and vector spaces are the core of a topological field theory[45] – translating these concepts to biology may provide a formal, algebraic language to map transitions between cellular system states.

In conclusion, MIAAIM provides a conceptual and practical foundation for novel applications in high-dimensional bioimage integration, interrogating composite tissue states, and transferring data across growing numbers of high-parameter imaging modalities. MIAAIM is technology-agnostic and works with any data in BioFormats/OME-compatible, imzML, NIfTI, and HDF5 formats. We envision MIAAIM to have broad applicability, from single-cell proteomic and transcriptomic imaging to untargeted molecular imaging. The harmonization of these data is poised to provide more comprehensive tissue portraits across health and disease states, and transitions between them.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nature medicine* **20**, 436 (2014).
2. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11**, 417-422 (2014).
3. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968-981. e915 (2018).
4. Gut, G., Herrmann, M.D. & Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science* **361** (2018).
5. Lin, J.-R., Fallahi-Sichani, M. & Sorger, P.K. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature communications* **6**, 1-7 (2015).
6. Lin, J.-R. et al. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* **7** (2018).
7. McDonnell, L.A. & Heeren, R.M. Imaging mass spectrometry. *Mass spectrometry reviews* **26**, 606-643 (2007).
8. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods* **16**, 987-990 (2019).
9. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
10. Peterson, V.M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nature biotechnology* **35**, 936-939 (2017).
11. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865-868 (2017).
12. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell*.
13. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
14. Abdelmoula, W.M. et al. Automatic 3D nonlinear registration of mass spectrometry imaging and magnetic resonance imaging data. *Analytical chemistry* **91**, 6206-6216 (2019).
15. Abdelmoula, W.M. et al. Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding. *Analytical chemistry* **86**, 9204-9211 (2014).
16. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845 (2016).
17. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature biotechnology* **37**, 451-460 (2019).
18. Moon, K.R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* **37**, 1482-1492 (2019).
19. Levine, J.H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184-197 (2015).
20. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L. & Nolan, G.P. Automated mapping of phenotype space with single-cell data. *Nature methods* **13**, 493-496 (2016).
21. Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236-249 (2020).
22. Consortium, H. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23. Thom, R. Quelques propriétés globales des variétés différentiables. *Commentarii Mathematici Helvetici* **28**, 17-86 (1954).
24. Schapiro, D. et al. MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging. *bioRxiv* (2021).
25. Schapiro, D. et al. MITI Minimum Information guidelines for highly multiplexed tissue images. *arXiv preprint arXiv:2108.09499* (2021).
26. Linkert, M. et al. Metadata matters: access to image data in the real world. *Journal of Cell Biology* **189**, 777-782 (2010).
27. Peng, T. et al. A BaSiC tool for background and shading correction of optical microscopy images. *Nature communications* **8**, 1-7 (2017).
28. Muhlich, J.L., Chen, Y.-A., Russell, D.P.W. & Sorger, P.K. Stitching and registering highly multiplexed whole slide images of tissues and tumors using ASHLAR software. *bioRxiv* (2021).
29. Schramm, T. et al. imzML—a common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of proteomics* **75**, 5106-5110 (2012).
30. Klein, S., Staring, M., Murphy, K., Viergever, M.A. & Pluim, J.P. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging* **29**, 196-205 (2009).
31. Schapiro, D. et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods* **14**, 873 (2017).
32. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal* **2014**, 2 (2014).
33. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nature biotechnology* **35**, 316-319 (2017).
34. Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3**, 1-9 (2016).
35. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
36. Sainburg, T., McInnes, L. & Gentner, T.Q. Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning. *arXiv preprint arXiv:2009.12981* (2020).
37. Rashid, R. et al. Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer. *Scientific data* **6**, 1-10 (2019).
38. Costa, J.A. & Hero, A.O. in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3 iii-988 (IEEE, 2004).
39. Costa, J.A. & Hero, A.O. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* **52**, 2210-2221 (2004).
40. Staring, M., Van Der Heide, U.A., Klein, S., Viergever, M.A. & Pluim, J.P. Registration of cervical MRI using multifeature mutual information. *IEEE transactions on medical imaging* **28**, 1412-1421 (2009).
41. Berg, S. et al. ilastik: Interactive machine learning for (bio) image analysis. *Nature Methods*, 1-7 (2019).
42. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology* **37**, 685-691 (2019).
43. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44. Regev, A. & Shapiro, E. Cellular abstractions: Cells as computation. *Nature* **419**, 343-343 (2002).
45. Atiyah, M.F. Topological quantum field theory. *Publications Mathématiques de l'IHÉS* **68**, 175-186 (1988).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

MIAAIM implementation. MIAAIM workflows are implemented in Python and connected via the Nextflow pipeline language to enable automated results caching and dynamic processing restarts after alteration of workflow parameters, and to streamline parallelized processing of multiple images. MIAAIM is also available as a Python package. Each data integration workflow is containerized to enable reproducible environments and eliminate any language-specific dependencies. MIAAIM's output interfaces with a number of existing image analysis software tools (see **Supplementary Note 1**, **Combining MIAAIM with existing bioimaging software**). MIAAIM therefore supplements existing tools rather than replaces them.

High-dimensional image compression and pre-processing (HDIprep). HDIprep is implemented by specifying sequential processing steps. Options include image compression for high-parameter data, and filtering and morphological operations for single-channel images. Processed images were exported as 32-bit NIfTI-1 images using the NiBabel library in Python. NIfTI-1 was chosen as the default file format for many of MIAAIM's operations due to its compatibility with Elastix, ImageJ for visualization, and its memory mapping capability in Python.

To compress high-parameter images, HDIprep identifies a steady-state embedding dimensionality for pixel-level data. Compression is initialized with optional, spatially-guided subsampling to reduce data set size. We then implement UMAP to construct a graph representing the data manifold and its underlying topological structure (*FuzzySimplicialSet, **Algorithm 1***). UMAP aims to optimize an embedding of a high-dimensional fuzzy simplicial set (i.e., a weighted, undirected graph) so that the fuzzy set cross-entropy between the embedded simplicial set and the high-dimensional counterpart is minimized, where the fuzzy set cross-entropy is defined as[35]:

Definition 1. Given a reference set A and membership functions $u: A \to [0,1]$, $v: A \to [0,1]$, the *fuzzy set cross-entropy* C of (A, u) and (A, v) is defined as:

$$C\big((A,u),(A,v)\big) \triangleq \sum_{a \epsilon A} \left(u(a) \log\left(\frac{u(a)}{v(a)}\right) + \big(1-u(a)\big) \log\left(\frac{1-u(a)}{1-v(a)}\right) \right)$$

The fuzzy set cross-entropy is a global measure of agreement between simplicial sets, aggregated across members of the reference set A (here, graph edges). Calculating its exact value scales quadratically

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the number of data points, restricting its use for large data sets. UMAP's current implementation does not, therefore. compute the exact cross entropy during its optimization of low-dimensional embeddings. Instead, it relies on probabilistic edge sampling and negative sampling to reduce runtimes for large data sets[35]. Congruently, to identify steady-state embedding dimensionalities, we compute patches on the data manifold that are representative of its global structure, and we use these patches in the calculation of the exact cross-entropy after projecting them with UMAP over a range of dimensionalities. The result is a global estimate of the dimensionality required to accurately capture manifold complexity.

To identify globally representative patches on the data manifold, we subject the fuzzy simplicial set to a variant of spectral clustering. We iteratively project the spectral centroids into Euclidean spaces of increasing dimensionality using UMAP and calculate the fuzzy set cross-entropy in each case, then min-max normalize the obtained values. To identify the steady-state embedding dimensionality, we fit a least-squares exponential regression to normalized cross-entropy as a function of dimensionality, then simulate samples along the regression line to find the first dimensionality falling within the 95% confidence interval of the exponential asymptote. The subsampled data is embedded into the steady-state dimensionality, and out-of-sample pixels are projected into this embedding using the native nearest-neighbor based method in UMAP (*transform* () function). Finally, all pixels are mapped back to their original spatial coordinates to construct a compressed image with the number of channels equal to the steady-state embedding dimensionality. These steps are summarized in pseudo-code below:

Algorithm 1: Image Compression.

Input: Multichannel image (X), SVD dimensionality (b), k-means clusters (k), embedding dimensions (n)

Output: Compressed image (I)

function `Compress`

{*Spatially Subsample Pixels*}

$p \longleftarrow \subseteq \{a_{x,y} \mid a_{x,y} \in X\}$

{*Compute Data Manifold*}

$g \longleftarrow$ *FuzzySimplicialSet* (p) ▷ UMAP

{*Compute Spectral Centroids*}

$s \longleftarrow$ *RandomizedSVD* (g, b)

$c \longleftarrow$ *MiniBatchKmeans* (s, k)

{*Compute Manifold Embedding Error*}

$g_c \longleftarrow$ *FuzzySimplicialSet* (c)

for $i = 1 \ldots n$ **do**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$e_c^i \leftarrow EmbedSimplicialSet\ (g_c, i)$
{*Compute Fuzzy Set Cross Entropy*}
$E_i \leftarrow FuzzySetCrossEntropy\ (g_c, e_c^i)$
$E \leftarrow \frac{E - E_{min}}{E_{max} - E_{min}}$ ▷ Min-max Normalization
{*Compute Steady-state Dimensionality*}
$\hat{y} \leftarrow (y_0 - c)e^{-kE} + c$ ▷ Exponential Regression
for $i = 1 \ldots n$ **do**
    if $\hat{y}_i$ **in** $c \pm 1.96\sigma$ **do** ▷ Model Gaussian Residual Process
        $j \leftarrow i$
        break
{*Compute Pixel Manifold Embedding*}
for $k_{x,y}$ **in** $\{a_{x,y} \mid a_{x,y} \in X \wedge a_{x,y} \notin p\}$ **do**
    $M \leftarrow ProjectData\ (k_{x,y}, e_c^j)$ ▷ Project data (UMAP *transform* ())
{*Reconstruct Image*}
for $k_{x,y} \in M$ **do**
    $I(x, y) \leftarrow k_{x,y}$
return I

Image data subsampling. Subsampling is performed at pixel level and is optional for image compression. Implemented options include uniformly spaced grids within the (*x, y*) plane, random coordinate selection, and random selection initialized with uniformly spaced grids ("pseudo-random"). HDIprep also supports specification of masks for sampling regions, which may be useful for extremely large data sets.

By default, images with fewer than 50,000 pixels are not subsampled, images with 50,000-100,000 pixels are subsampled using 55% pseudo-random sampling initialized with 2x2 pixel uniformly spaced grids, images with 100,000-150,000 pixels are subsampled using 15% pseudo-random sampling initialized with 3x3 pixel grids, and images with more than 150,000 pixels are subsampled with 3x3 pixel grids. These default values are based on empirical studies (**Supplementary Figs. 4, 5, and 6**).

No subsampling was used for presented MSI data. Subsampling rates used for presented IMC data were determined on a case-by-case basis from empirical studies and match those used in the spectral landmark sampling experiments. Subsampling with 10x10 pixel uniformly spaced grids was used for CyCIF data compression.

Fuzzy simplicial set generation. To construct a pixel-level data manifold, we represent each pixel as a d-dimensional vector, where d is the number of channels in the given high-parameter image (i.e., discarding

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spatial information). We then implement the UMAP algorithm and extract the resulting fuzzy simplicial set representing the manifold structure of these d-dimensional points. For all presented results, we used the default UMAP parameters to generate this manifold: 15 nearest neighbors and the Euclidean metric.

Manifold landmark selection with spectral clustering. Spectral landmarks are identified using a variant of spectral clustering. We use randomized singular value decomposition (SVD) followed by mini-batch k-means to scale spectral clustering to large data sets, following the procedure introduced in the potential of heat diffusion for affinity-based transition embedding (PHATE) algorithm[18]. Given a symmetric adjacency matrix A representing pairwise similarities between nodes (here, pixels) originating from a d-dimensional space $\mathbb{R}^d$, we first compute the eigenvectors corresponding to the k largest eigenvalues of A. We then perform mini-batch k-means on the nodes of A using these k eigenvectors as features. Spectral landmarks are defined as the d-dimensional centroids of the resulting clusters.

By default, the input data is reduced to 100 components using randomized SVD and then split into 3,000 clusters using mini-batch k-means. These default parameter values are based on empirical studies (**Supplementary Fig. 7**). Due to steady-state embeddings of MSI and IMC data only being available after experimental tests, no landmark selection was used for processing or determining the optimal embedding dimensionality of these data sets. Instead, full or subsampled datasets were used. All other steady-state embeddings for image data was compressed using the above default parameters.

Steady-state UMAP embedding dimensionalities. By default, HDIprep embeds spectral landmarks into Euclidean spaces with 1-10 dimensions to identify steady-state embedding dimensionalities. Exponential regressions on the spectral landmark fuzzy set cross entropy are performed using built-in functions from the Scipy Python library. These default parameters were used for all presented data.

Histology image preprocessing. HDIprep processing options for hematoxylin and eosin (H&E) stained tissues and other low-channel histological stains include image filters (e.g., median), thresholding (e.g., manually set or automated), and successive morphological operations (e.g., thresholding, opening and closing). Presented H&E and toluidine-blue stained images were processed using median filters to remove salt-and-pepper noise, followed by Otsu thresholding to create a binary mask representing the foreground. Sequential morphological operations were then applied to the mask, including morphological opening to

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remove small connected foreground components, morphological closing to fill small holes in foreground, and filling to close large holes in foreground.

Image compression with UMAP parametrized by a neuronal network. We implemented parametric UMAP[36] using the default parameters and neural architecture with a TensorFlow backend. The default architecture was comprised of a 3-layer 100-neuron fully connected neuronal network. Training was performed using gradient descent with a batch size of 1,000 edges and the Adam optimizer with a learning rate of 0.001.

High-dimensional image registration (HDIreg). HDIreg is a containerized workflow implementing the open-source Elastix software[30] in conjunction with custom-written Python modules to automate the image resizing, padding, and trimming often applied before registration. HDIreg incorporates several different registration parameters, cost functions, and deformation models, and additionally allows manual definition of point correspondences for difficult problems, as well as composition of transformations for fine-tuning (see **Supplemental Note 2**, **Notes on the HDIreg workflow's expected performance**).

High-parameter images are registered using a manifold alignment scheme parametrized by spatial transformations, which aims to maximize image similarity. Formally, we view registration as the following optimization problem[40]:

Given a fixed d-dimensional image $I_F\colon \mathbb{R}^d \to \mathbb{R}^2$ with domain Ω_F and a moving q-dimensional image $I_M\colon \mathbb{R}^q \to \mathbb{R}^2$ with domain Ω_m, we aim to optimize

$$\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} S(\boldsymbol{T}_{\boldsymbol{\mu}}; I_F, I_M, \Omega_F) \quad (1)$$

where $\boldsymbol{T}_{\boldsymbol{\mu}}\colon \Omega_F \to \Omega_M$ is a smooth transformation defined by the vector of parameters $\boldsymbol{\mu} \subset \mathbb{R}^M$, and S is a similarity measure maximized when $I_M \circ \boldsymbol{T}_{\boldsymbol{\mu}}$ and I_F are aligned.

Differential geometry and manifold learning: MIAAIM's manifold alignment scheme uses the entropic graph-based Rényi α-mutual information (α-MI)[40] as the similarity measure S in **Equation 1**, which extends to manifold representations of images (i.e., compressed images) embedded in Euclidean space with potentially differing dimensionalities. This measure is justified in the HDIreg manifold alignment scheme through the notion of *intrinsic manifold information* (i.e., entropy). In what follows, we introduce

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

613 basic differential geometric concepts that enable us to expand existing foundations of intrinsic manifold
614 entropy estimation to the UMAP algorithm. We assume familiarity with the definition of a topological
615 space (see Armstrong[46] for an introduction).
616
617 **Definition 2:** Let X and Y be topological spaces. A function $f: X \to Y$ is *continuous* if for each point $x \in$
618 X and each open neighborhood N of $f(x) \subseteq Y$ the set $f^{-1}(N)$ is an open neighborhood of $x \in X$. A
619 function $f: X \to Y$ is a *homeomorphism* if it is one-to-one, onto, continuous, and has a continuous inverse.
620 When a homeomorphism exists between spaces X and Y, they are called *homeomorphic* spaces.
621
622 **Definition 3.** A *manifold $\mathcal{M}$ of dimension n* (i.e., an n-manifold) is a second-countable Hausdorff space,
623 each point of which has an open neighborhood homeomorphic to n-dimensional Euclidean space, $\mathbb{R}^n$. For
624 any open set $U \subset \mathcal{M}$ we can define a chart (φ, U) where $\varphi: U \to \widetilde{U} \subset \mathbb{R}^n$ is a homeomorphism. We can
625 say that (φ, U) acts as a local coordinate system for $\mathcal{M}$, and we can define a transition between two charts
626 (φ, U) and (ω, V) as $\varphi \circ \omega^{-1}: \omega(U \cap V) \to \varphi(U \cap V)$ when $U \cap V$ is non-empty.
627
628 **Definition 4.** A *smooth manifold* is a manifold where there exists a smooth transition map between each
629 chart of $\mathcal{M}$. A *Riemannian metric* g is a mapping that associates to each point $y \in \mathcal{M}$ an inner product
630 $g_y(\cdot, \cdot)$ between vectors tangent to $\mathcal{M}$ at y. We denote tangent spaces of y as $T_y\mathcal{M}$. A *Riemannian*
631 *manifold*, written $(\mathcal{M}, g)$, is a smooth manifold $\mathcal{M}$ together with a Riemannian metric g. Given a
632 Riemannian manifold, the *Riemannian volume element* provides a means to integrate a function with
633 respect to volume in local coordinates. Given $(\mathcal{M}, g)$, we can express the volume element ω in terms of
634 the metric g and the local coordinates of a point $x = x_1, \ldots, x_n$ as $\omega = g(x)dx_1 \wedge \ldots \wedge dx_n$ where $g(x) >$
635 0 and $\wedge$ indicates the wedge product. The volume of $\mathcal{M}$ under this volume form is given by $Vol(\mathcal{M}) =$
636 $\int_{\mathcal{M}} \omega$.
637
638 **Definition 5.** An *immersion* of a smooth n-manifold $\mathcal{M}$ into $\mathcal{N}$ is a differentiable mapping $\psi: \mathcal{M} \to \mathcal{N}$
639 such that $d\psi_p: T_p\mathcal{M} \to T_{\psi(p)}\mathcal{N}$ is injective for all points $p \in \mathcal{M}$. ψ is therefore an immersion if its
640 derivative is everywhere injective.
641

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition 6. An *embedding* between smooth manifolds $\mathcal{M}$ and $\mathcal{N}$ is a smooth function $f: \mathcal{M} \to \mathcal{N}$ such that f is an immersion and an embedding of topological spaces (i.e., is an injective homeomorphism).

Let $(\mathcal{M}, g)$ be a compact, n-dimensional Riemannian manifold immersed in an ambient $\mathbb{R}^d$, where $n \ll d$, and let $\mathcal{X}_j = \{X_1, \ldots, X_j\}$ be a set of independent and identically distributed random vectors with values drawn from a distribution supported by $\mathcal{M}$. Define $\mathcal{U}_j = \{U_1, \ldots, U_j\}$ to be open neighborhoods of the elements of $\mathcal{X}_j$. An aim of manifold learning is to approximate an embedding f such that a measure of distortion D is minimized between $\mathcal{U}_j$ and $f(\mathcal{U}_j) = \{f(U_1), \ldots, f(U_j)\}$. The manifold learning problem can therefore be written as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} D(\mathcal{U}_j, f(\mathcal{U}_j)) \tag{2}$$

where $\mathcal{F}$ represents the family of possible measurable functions taking $x \mapsto f(x)$. In machine learning settings, open neighborhoods $U_i \in \mathcal{U}_j$ about vectors $X_i \in \mathcal{X}_j$ are often defined to be KNN neighborhoods with geodesic distances (or probabilistic encodings thereof[47]) within each open neighborhood approximated with a positive definite kernel[48], which allows the computation of inner products in a Riemannian framework (as compared with a pseudo-Riemannian framework which need not be positive definite). Measures of distortion vary by algorithm (see **Supplementary Note 3, HDIprep dimension reduction validation** for examples). Of interest in our exposition are the measures induced by the embedded geodesics via the volume elements of these coordinate patches. These provide the components needed to quantify the *intrinsic Rényi α-entropy* of embedded data manifolds.

Entropic graph estimators. Given Lebesgue density f and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with values in a compact subset of $\mathbb{R}^d$, the *extrinsic* Rényi α-entropy of f is given by:

$$H_\alpha^{\mathbb{R}^d}(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(x)\, \mathrm{d}x \tag{3}$$

where $\alpha \in (0,1)$.

Definition 7 (adapted from Costa and Hero[38]). Let $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be identically distributed random vectors with values in a compact subset of $\mathbb{R}^d$, the nearest neighbor of $\boldsymbol{X}_i \in \mathcal{X}_n$ under the Euclidean metric is given by:

$$\arg\max_{\boldsymbol{X} \in \mathcal{X}_n \setminus \{\boldsymbol{X}_i\}} \|\boldsymbol{X} - \boldsymbol{X}_i\|_2 \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A k-nearest neighbor (KNN) graph puts and edge between each $\boldsymbol{X}_i \in \mathcal{X}_n$ and its k-nearest neighbors. Let $\Gamma_{k,i} = \Gamma_{k,i}(\mathcal{X}_n)$ be the set of k-nearest neighbors of $\boldsymbol{X}_i \in \mathcal{X}_n$. Then the total edge length of the KNN graph for $\mathcal{X}_n$ is given by:

$$L_{\gamma,k}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{X} \in \Gamma_{k,i}} \|\boldsymbol{X} - \boldsymbol{X}_i\|_2^{\gamma} \tag{5}$$

where $\gamma > 0$ is a power-weighting constant.

In practice, the *extrinsic* Rényi α-entropy of f can be suitably approximated using a class of graphs known as continuous quasi-additive graphs, including k-nearest neighbor (KNN) Euclidean graphs[49], as their edge lengths asymptotically converge to the Rényi α-entropy of feature distributions as the number of feature vectors increases[50]. This property leads to the convergence of KNN Euclidean edge lengths to the *extrinsic* Rényi α-entropy of a set of random vectors with values in a compact subset of $\mathbb{R}^d$ with $d \geq 2$ [38]. This is a direct corollary of the **Beardwood-Halton-Hammersley Theorem** outlined below.

Beardwood-Halton-Hammersley (BHH) Theorem[38, 49]. *Let $(\mathcal{M}, g)$ be a compact Riemannian m-manifold immersed in an ambient $\mathbb{R}^d$. Suppose $\mathcal{X}_n = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ are identically distributed random vectors with values in a compact subset of $\mathbb{R}^d$ and Lebesgue density f. Assume $d \geq 2, 1 \leq \gamma < d$ and define $\alpha = \frac{(d-\gamma)}{d}$. Then with probability 1,*

$$\lim_{n\to\infty} \frac{L_{\gamma,k}(\mathcal{X}_n)}{n^{\frac{d'-\gamma}{d'}}} = \beta_{d,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{x}) d\boldsymbol{x} \tag{6}$$

The value that determines the right side of the limit in **Equation 6** is the extrinsic Rényi α-entropy given by Equation 4. When identically distributed random vectors are restricted to a compact smooth m-manifold, $\mathcal{M}$, in an ambient $\mathbb{R}^d$, the **BHH Theorem** generalizes to enable an estimation of *intrinsic* Rényi α-entropy $H_{\alpha}^{\mathcal{M}}(f)$ of the multivariate density f on $\mathcal{M}$ defined by

$$H_{\alpha}^{\mathcal{M}}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \tag{7}$$

by incorporating the measure μ_g naturally induced by the Riemannian metric via the Riemannian volume element[38]. This is formalized by the following given by Costa and Hero[38]:

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Theorem 1 (Costa and Hero[38]): *Let* $(\mathcal{M}, g)$ *be a compact Riemannian* m*-manifold immersed in an ambient* $\mathbb{R}^d$*. Suppose* $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ *are identically distributed random vectors of* $\mathcal{M}$ *with bounded density* f *relative to the differential volume element* μ_g *induced by the metric* g*. Assume* $d \geq 2, 1 \leq \gamma < d$ *and define* $\alpha = \frac{(d-\gamma)}{d}$*. Then with probability 1,*

$$\lim_{n\to\infty} \frac{L_{\gamma,k}(\mathcal{Y}_n)}{n^{\frac{d'-\gamma}{d'}}} = \begin{cases} \infty, & d' < m \\ \beta_{m,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y}) \mu_g(d\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases} \quad (8)$$

where $\beta_{m,\gamma,k}$ *is a constant that is independent of* f *and* $(\mathcal{M}, g)$*. Similarly, the expectation* $E\left[L_{\gamma,k}(\mathcal{Y}_n)\right]/n^{\alpha}$ *converges to the same limit.*

The quantity that determines the limit when $d' = m$ is the intrinsic Renyi alpha entropy of f given by **Equation 7. Theorem 1** has been used in conjunction with manifold learning algorithms Isomap and a variant C-Isomap to estimate the intrinsic dimensionality of embedded manifolds[39]. In contrast to these results that use all pairwise geodesic approximations for each point in the data set to estimate the α-entropy, we aim to provide a similar formulation utilizing local information contained in data manifolds, following the results of our dimension reduction benchmark which shows that local information preserving algorithms are well-suited for the task of high-dimensional image data compression (**Supplementary Figs 1, 2, 3**). Indeed, the *information density* of volumes of continuous regions of model families (i.e., collections of output embedding spaces or input points) have been recognized in defining an information geometry of statistical manifold learning[47], motivating the title of our software.

Entropic graph estimators on local information of embedded manifolds: In what follows, we utilize two concepts to show that the intrinsic information of multivariate probability distributions supported by embedded manifolds in Euclidean space with the UMAP algorithm can be approximated using the **BHH Theorem**: (i.) the compactness of constructed manifolds and (ii.) the conservation of Riemannian volume elements. We address (i.) with a simple proof, and we provide a motivational example of conservation of volume elements using UMAP to address (ii.).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition 8. A topological space X is *compact* if every open cover $\mathcal{A}$ of X contains a finite sub collection that also covers X. By finite open cover, we mean that the elements of $A_i \subset \mathcal{A}$ with $i \in \{1, \dots, n\}$ are open sets in X, and that the union of the elements of $\mathcal{A}$ equals X: $\bigcup_{i=1}^{n} A_i = X$.

Proposition 1. *Let $n > d$ and suppose that $\mathcal{M}$ is a compact manifold of dimension r with $r \leq d$ that is immersed in ambient $\mathbb{R}^n$. Then the image $f(\mathcal{M})$ of $\mathcal{M}$ under a projection $f: \mathcal{M} \to \mathbb{R}^d$ is compact.*

Proof. Let $(\mathcal{M}, g)$ be a compact Riemannian manifold (e.g., a manifold constructed with UMAP) with metric g in an ambient $\mathbb{R}^n$ and f a projection from $\mathcal{M}$ to $\mathbb{R}^d$. Since f is a projection, it is continuous, which takes compact sets to compact sets. ■

Proposition 1 shows that a d-dimensional Euclidean projection of a compact Riemannian manifold take values in a compact subset of $\mathbb{R}^d$, a necessary condition in the **BHH Theorem**. The UMAP algorithm considers fuzzy simplicial sets, i.e., manifolds, constructed from finite extended pseudo-metric spaces (see finite fuzzy realization functor, Definition 7, given by McInnes and Healy[35]). By finite, we mean that these extended pseudo-metric spaces are constructed from a finite collection of points. If one considers this finiteness condition, then the compactness of UMAP manifolds follows naturally from **Definition 8** – given an open cover on a manifold, one can find a finite subcover. Thus, UMAP projections are compact, following **Proposition 1**.

To extend the **BHH Theorem** to the calculation of intrinsic α-entropy of UMAP embeddings as in **Equation 7**, we must show that volume elements induced via embedding are well approximated. Note that these results apply to any dimension reduction algorithm that can provably preserve distances within open neighborhoods upon embedding a compact manifold in Euclidean space. In what follows, we do not provide a proof that UMAP preserves distances within open neighborhoods about points, although, this would be an ideal scenario. Rather, we assume that this ideal scenario exists, and we describe how to find the optimal dimensionality for projecting data to satisfy this assumption.

In contrast to global data preserving algorithms, such as Isomap that calculate all pairwise geodesic distances or approximations thereof with landmark based approaches, UMAP approximates geodesic distances in open neighborhoods local to each point (see **Lemma 2** below). Given Lebesgue density μ and identically distributed random vectors $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ with values constrained to lie on a compact, uniformly distributed Riemannian manifold $\mathcal{M}$, geodesics between samples $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ drawn from μ are encoded probabilistically with UMAP and represent a scaled exponential distribution[51]:

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$P_{j|i} = \exp\left(\frac{-\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\| - \rho_i}{\sigma_i}\right) \tag{9}$$

$$P_{ij} = P_{j|i} + P_{i|j} - P_{j|i}P_{i|j} \tag{10}$$

where ρ_i is the distance from vector $\boldsymbol{Y}_i$ to its nearest neighbor and σ_i is an adaptively chosen normalization factor. Using the terminology in **Equation 2,** the objective of embedding in UMAP is given by minimizing the fuzzy simplicial set cross-entropy (**Definition 1**), which represents distortion D. Formally, given probability distribution P_{ij} encoding geodesics between samples $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ and probability distribution Q_{ij} encoding distances between samples $f(\boldsymbol{Y}_i)$ and $f(\boldsymbol{Y}_j)$, we can represent the cross-entropy loss employed by UMAP as (adapted from Narayan et. al[51]):

$$CE(P\|Q) = -\sum_{ij} P_{ij}\log Q_{ij} + (1 - P_{ij})\log(1 - Q_{ij}) \tag{11}$$

where Q_{ij} is the probability distribution formed from low dimensional positions of embedded vectors $f(\boldsymbol{Y}_i)$ and $f(\boldsymbol{Y}_j)$ by $Q_{ij}(a,b) = \frac{1}{\left(1+a\|f(\boldsymbol{Y}_i)-f(\boldsymbol{Y}_j)\|_2^{2b}\right)}$ with a, b user-defined parameters to control embedding spread.

Minimizing **Equation 11** is not, in general, a convex optimization problem. Optimization over the family $\mathcal{F}$ from **Equation 2** is restricted to a subset rather than the full family and thus represents, in the best case, a local optimum. We include a larger family of measurable functions in order to more accurately approach optimal embedding of geodesic distances within open neighborhoods of vectors through the identification steady-state embedding dimensionalities, as outlined in the HDIprep workflow in a "pseudo-global" optimization procedure.

Using the notation in **Equation 2**, the volume elements induced by the embedding $\hat{f}$ are similarly those that minimize distortion between the volume elements of $\mathcal{M}$ given for local coordinates $x_1, \ldots, x_n$ by $\sigma = g(x)dx$ where $dx = dx_1 \wedge \ldots \wedge dx_n$ and those of $\hat{f}(\mathcal{M})$ given for local coordinates $\hat{f}x_1, \ldots, \hat{f}x_n$ by $\tau = \hat{f}g(x)d(\hat{f}x)$, assuming that the dimensionality n of local coordinates is preserved. Under a globally optimal solution to **Equation 7 (**i.e., when $P_{ij} = Q_{ij}$), volume elements are preserved: $\int_{\mathcal{M}} \sigma = \int_{\mathcal{M}} \tau$. The existence of volume preserving diffeomorphisms for compact manifolds in consideration are proven by Moser[52].

In order to identify manifold embeddings that minimize distortion of volume elements, we viewed increases in the dimensionality of real-valued data in a natural way by modeling increases in dimensionality as exponential increases in potential positions of points (i.e., increasing copies of the real

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

line, $\mathbb{R}^n$). The relationship between radii of open neighborhoods, volume, and density of manifolds in a learning setting is formalized by Narayan et. al[51] in an application where density is preserved with a given dimensionality for embeddings by altering open neighborhood radii; however, we can readily extend this in an adapted scenario of volume preservation under fixed radii to infer a dimensionality that satisfies the assumption of geodesic distance preservation. Consider the following example:

Let $\boldsymbol{Y}_i \in \mathcal{M}$ be a vector of manifold $\mathcal{M}$ immersed in an ambient $\mathbb{R}^d$, and assume that the k-nearest neighbors of $\boldsymbol{Y}_i$ are uniformly distributed in a ball B_{r_d} of radius r_d with proportional volume $V_d \propto r_d^d$. Assume that a map f takes the open neighborhood B_{r_d} of $\boldsymbol{Y}_i$ to an m-dimensional ball $B_{r_m} \subseteq \mathcal{N}$ of manifold $\mathcal{N}$ with radius r_m while preserving its structure, including the uniform distribution and the induced Riemannian volume element V_m. Following Narayan et al.[51], we can use the proportion $V_d \propto r_m^m$ to infer a power law relationship $r_m \propto r_d^{d-m}$ between the local radii r_m in embedding spaces and the original radius of B_{r_d}.

To extend this example, suppose that r_m and r_d are fixed (we can assume that radii are fixed in the original space, and that those in the embedding space are controlled by the a, b parameters influencing Q_{ij} in **Equation 11**). Assume that the ambient metrics of the embedding space and the original space are the same, and that they give rise to geodesic distances within B_{r_m} and B_{r_d} using the native UMAP method (**Lemma 2** below). Since ambient metrics and radii are preserved and that $r_m \propto r_d^{d-m}$, this implies that the geodesic distances δ_m and δ_d between points in B_{r_m} and B_{r_d} also exhibit a power law relationship, $\delta_m \propto \delta_d^{d-m}$. If we further assume that $\delta_m = \delta_d^d$ (i.e., that geodesic distances between points in open neighborhoods are preserved, the ideal scenario) we can solve for the relationship between geodesics and dimensionality m. Specifically, suppose that $\delta_m = \delta_d^d$. Substituting δ_d^d for δ_m, we can see that

$$\delta_d^d = \delta_d^{d-m}$$

$$\rightarrow \delta_d^d = \frac{\delta_d^d}{\delta_d^m}$$

$$\rightarrow \delta_d^m = \frac{\delta_d^d}{\delta_d^d} = 1$$

which implies that the geodesics within open neighborhoods of points of the original space have an exponential relationship with dimensionality m.

Using the power-law relationship between volumes of open neighborhoods in UMAP manifolds and their embedded counterparts, we can attempt to identify the dimensionality m that such that geodesics

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

within open neighborhoods is preserved with exponential regression. Note that geodesic distance preservation is stronger than volume preservation; however, volume preservation is implied. As a result, steady-state manifold embeddings provide a Euclidean dimensionality to approximate manifold geodesics of vectors sampled from $\mathcal{M}$, and thus the volume elements of $\mathcal{M}$. KNN graph functionals calculated in the steady-state embedding space provide the necessary machinery to calculate the intrinsic α-entropy of embedded data manifolds in MIAAIM by applying the **BHH Theorem** using the induced measure across all coordinate patches as in **Theorem 1**.

Note, though, that the distances outside of open neighborhoods of points are not guaranteed to be accurately modelled in the embedding space with UMAP if one makes the assumptions introduced in our example. Therefore, applying **Theorem 1** in conjunction with UMAP by replacing KNN graph lengths with those obtained with length functionals of geodesic minimal spanning trees (GMST), another type of entropic graph, should not be expected to reproduce the intrinsic entropy originally reported by Costa and Hero[39]. Our primary contribution here is combining the KNN intrinsic entropy estimator with a local information preserving dimension reduction algorithm. We wish to extend these results to the setting where two such manifolds are compared, as this serves the basis for our image registration application. An entropic graph-based estimator of α-MI in an image registration setting is described by the following[40]:

Let $z(x_i) = [z_i(x_i), \dots, z_d(x_i)]$ be a d-dimensional vector encoding the features of point x_i. Let $Z_f(x) = \{z^f(x_1), \dots, z^f(x_N)\}$ be the feature set of a fixed image, $Z_m\left(\boldsymbol{T_\mu}(x)\right) = \{z^m(\boldsymbol{T_\mu}(x_1)), \dots, z^m(\boldsymbol{T_\mu}(x_N))\}$ be the feature set of a transformed moving image at points in $T_\mu(x)$, and $z_i^{fm} = [z^f(x_i), z^m\left(\boldsymbol{T_\mu}(x_i)\right)]$ be the concatenation of the feature vectors of the fixed and transformed moving image at x_i. Then,

$$\widehat{\alpha\text{-}MI}\left(\boldsymbol{\mu}; Z_f, Z_m, Z_{fm}\right) = \frac{1}{\alpha - 1}\log\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\frac{\Gamma_i^{fm}}{\sqrt{\Gamma_i^f\Gamma_i^m(\boldsymbol{\mu})}}\right)^{2\gamma} \tag{12}$$

is a graph-based estimator for α-MI, where $\gamma = d(1-\alpha)$, $0 < \alpha < 1$, and three graphs

$$\Gamma_i^f = \sum_{i=1}^{k}\left\|z^f(x_i) - z^f\left(x_{ip}\right)\right\| \tag{13}$$

$$\Gamma_i^m(\boldsymbol{\mu}) = \sum_{i=1}^{k}\left\|z^m(\boldsymbol{T_\mu}(x_i)) - z^m(\boldsymbol{T_\mu}(x_{ip}))\right\| \tag{14}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\Gamma_i^{fm}(\boldsymbol{\mu}) = \sum_{i=1}^{k} \left\| z^{fm}(x_i, \boldsymbol{T}_{\boldsymbol{\mu}}(x_i)) - z^{fm}(x_{ip}, \boldsymbol{T}_{\boldsymbol{\mu}}(x_{ip})) \right\| \quad (15)$$

represent the Euclidean graph functionals (lengths) of a feature vector z to its p^{th} nearest neighbor over k considered nearest neighbors.

The Rényi α-MI provides a quantitative measure of association between the intrinsic structure of multiple manifold embeddings constructed with the UMAP algorithm. The Rényi α-MI measure extends to feature spaces of arbitrary dimensionality, which MIAAIM utilizes in combination with its image compression method to quantify similarity between steady-state embeddings of image pixels in potentially differing dimensionalities.

Multimodal image registration studies. Since data acquisition removed tissue context at regions of interest on the IMC full-tissue reference image, we first aligned full tissue sections then used the coordinates of IMC regions to extract data from all modalities for fine-tuning. Custom Python scripts were used to propagate alignment across imaging modalities. Manual landmark correspondences were used where unsupervised alignment proved suboptimal. We accounted for alignment errors around IMC regions following full-tissue registration by padding regions with extra pixels prior to cropping. All registrations involving MSI or IMC data were conducted using KNN α-MI, as in **Equation 12**, with $\alpha = 0.99$ and 15 nearest neighbors. All registrations aligning low-channel slides (IMC reference toluidine blue image and H&E) were conducted using histogram-based MI after grayscale conversion for rapid processing.

For full-tissue images, a two-step registration process was implemented by first aligning images using an affine model for the vector of parameters $\boldsymbol{\mu}$ (**Equation 1**) and subsequently aligning images with a nonlinear model parametrized by B-splines. Hierarchical Gaussian smoothing pyramids were used to account for resolution differences between image modalities, and stochastic gradient descent with random coordinate sampling was used for optimization. We additionally optimized final control point grid spacings for B-spline models and the number of hierarchical levels to include in pyramidal smoothing for each MSI data set to H&E alignment individually (**Supplementary Figs. 1, 2, 3**). A final control point spacing of 300 pixels for nonlinear B-spline registrations of MSI data to corresponding H&E data balanced correct alignment with unrealistic warping, which we identified visually and by inspecting the determinants of the spatial Jacobian matrices for values that deviated substantially from 1. H&E and IMC reference tissue registrations utilized a final grid spacing of 5 pixels. Similar optimizations for numbers of pyramidal levels were carried out for these data. All data that underwent image registration were

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exported and stored as 32-bit NIfTI-1 images. IMC data was not transformed and was kept in 16-bit OME-TIF(F) format.

Cobordism Approximation and Projection (PatchMAP). PatchMAP is an algorithm that constructs a smooth manifold by representing lower dimensional manifolds as its boundary, and it projects the higher-order manifold in a lower dimensional space for visualization. Higher-order manifolds produced by PatchMAP can be understood as cobordisms, which are described by the following set of definitions:

Definition 9. Given a family of sets $\{S_i : i \in I\}$ and an index set I, their *disjoint union,* denoted $S = \coprod_{i \in I} S_i$ is a set together with injective functions $\phi_i : S_i \to S$ for each S_i. The disjoint union corresponds to the coproduct of sets.

Definition 10. Two closed n-manifolds $\mathcal{M}$ and $\mathcal{N}$ are *cobordant* if their disjoint union, denoted $\mathcal{M} \coprod \mathcal{N}$, is the boundary of an n+1-manifold $\mathcal{W}$. We call the manifold $\mathcal{W}$ a *cobordism*. By boundary of a d-manifold, we mean the set of points on $\mathcal{M}$ which are homeomorphic to the upper half plane, $\mathbb{R}^d_+$. We denote the boundary of $\mathcal{W}$ as $\partial\mathcal{W}$.

PatchMAP addresses cobordism learning in a semi-supervised manner, where data is assumed to follow the structure of a nonlinear cobordism, and our task is to glue lower dimensional manifolds to the boundary of a higher-dimensional manifold to produce a cobordism. Here, we would like for coordinate transformations across the cobordism to have their own geometries independent of the metrics of boundary manifolds. In practice, this property allows us to explore data within boundary manifolds without being reliant on the specific geometry of the cobordism. Ultimately, cobordism geodesics are the base component of downstream applications such as the i-PatchMAP workflow. Furthermore, we would like for the cobordism to highlight boundary manifolds whose data points are highly similar. A natural way to satisfy both of these conditions is to use the fuzzy set-theoretic foundation of the UMAP algorithm.

The primary goal of PatchMAP then is to identify a smooth manifold whose boundary is the disjoint union of smooth manifolds of lower dimensionality, and which has a metric independent of each boundary manifolds' metric that we choose to represent. We address this with a two-step algorithm that uncouples the computation of boundary manifolds from the cobordism. First, boundary manifolds are computed by applying the UMAP algorithm to each set of data with a user-provided metric. Practically,

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the result of this step are symmetric, weighted graphs that represent geodesics within each boundary manifold. Our task is to construct from a finite set of n boundary manifolds $\mathcal{F} = \{(\mathcal{M}_1, g_1) \ldots (\mathcal{M}_n, g_n)\}$ a manifold $(\mathcal{M}_\mathcal{F}, g)$, with metric g such that the disjoint union $\coprod_{i \in I} \mathcal{M}_i = \partial \mathcal{M}_\mathcal{F}$. We wish to approximate the geodesics of $\mathcal{M}_\mathcal{F}$, and for each element of $p \in \mathcal{M}_\mathcal{F}$, the tangent space $T_p\, \mathcal{M}_\mathcal{F}$ with inner product g_p.

Lemma 2 (McInnes and Healy[35]). *Let $(\mathcal{M}, g)$ be a Riemannian manifold in an ambient $\mathbb{R}^n$, and let $p \in \mathcal{M}$ be a point. If g is locally constant about p in an open neighborhood $U \subseteq \mathcal{M}$ such that g is a constant diagonal matrix in ambient coordinates, then in a ball $B \subseteq U$ centered at p with volume $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$ with respect to g, the geodesic distance from p to any point in $q \in B$ is $\frac{1}{r} d_{\mathbb{R}^n}(p, q)$, where r is the radius of the ball in the ambient space $\mathbb{R}^n$ and $d_{\mathbb{R}^n}$ is the metric on the ambient space.*

If we assume that data points across boundary manifolds can be compared with a suitable metric, we can compute the geodesic distances between points on the projections of each boundary manifold under a user-provided ambient metric using **Lemma 2** (for a proof of **Lemma 2**, see Appendix A of McInnes and Healy[35]) to their disjoint union. Given two boundary manifolds $\mathcal{M}_i$ and $\mathcal{M}_j$, we compute the pairwise geodesic distances between points $p_i, p_j \in \mathcal{M}_i \coprod \mathcal{M}_j$ where $p_i \in \mathcal{M}_i \to \mathcal{M}_i \coprod \mathcal{M}_j$ and $p_j \in \mathcal{M}_j \to \mathcal{M}_i \coprod \mathcal{M}_j$ to obtain the components needed to construct the metric on $\mathcal{M}_{ij}$ where the boundary $\partial \mathcal{M}_{ij} = \mathcal{M}_i \coprod \mathcal{M}_j$. By extension, concatenation of all pairwise distance calculations between boundary manifolds $\mathcal{M}_i$ and $\mathcal{M}_j$ where $i, j \in \{1, 2, \ldots, n\}; i \neq j$ to their disjoint union $\coprod_{i \in I} \mathcal{M}_i$ provides the components to construct the geodesics on the full cobordism $\mathcal{M}_\mathcal{F}$ across all pairwise combinations of boundary manifolds, as desired. The result of using **Lemma 2** to approximate the projections of manifold geodesics, however, is that we have directed, incompatible views of the geodesics across the cobordism, to and from, boundary manifolds. That is, $\mathcal{M}_{ij} \neq \mathcal{M}_{ji}$. We can interpret these directed geodesics as being defined on *oriented* cobordisms. We aim to encode the directed geodesics in the oriented cobordisms and the boundary manifold geodesics with a single data representation.

Our goal then is to construct an *unoriented* cobordism, where the directed geodesics of oriented cobordisms are resolved into a single, symmetric matrix representation. To do this, we can translate each of the extended pseudo-metric spaces described above into a fuzzy simplicial set using the fuzzy singular set functor (see Definition 9 by McInnes and Healy[35]), which captures both a topological representation

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the oriented cobordisms and the underlying metric information. The incompatibilities in oriented cobordism geodesics can be resolved with a norm of our choosing. A natural choice for the fuzzy set representation that we have chosen is the t-norm (also known as a *fuzzy intersection*). If we interpret the fuzzy simplicial set representations of oriented cobordisms in a probabilistic manner, then their intersection corresponds to the joint distribution of the oriented cobordism metric spaces, which highlights the directed cobordism geodesics that occur in both directions, to and from, with a high probability.

After symmetrizing the cobordisms between each pairwise boundary manifold, the final step is to integrate the boundary manifold geodesics with the symmetric cobordism geodesics obtained with the fuzzy set intersection. Additionally, we should ensure that cobordisms with shared domains (i.e., boundary manifolds) are identified and integrated with each other. We can do this by taking a *fuzzy set union* (probabilistic t-conorm) over the family of extended pseudo-metric spaces, as in the original UMAP implementation. The result is a cobordism that contains its own geometric structure that is captured in cobordism geodesics, in addition to individual boundary manifolds that contain their own geometries.

Optimizing a low dimensional representation of the cobordism can be achieved with a number of methods – we choose to optimize the embedding using the fuzzy set cross entropy (**Definition 1**), as in the original UMAP implementation for consistency. Note that since our algorithm produces a symmetric matrix, PatchMAP could be repeatedly applied to construct "nested" cobordisms of hierarchical dimensionality.

PatchMAP implementation. To construct cobordisms, PatchMAP first computes boundary manifolds by constructing fuzzy simplicial sets from each provided set of data, i.e., tissue state, by applying the UMAP algorithm (*FuzzySimplicialSet,* **Algorithm 2**). Then, pairwise, directed nearest neighbor (NN) queries between boundary manifolds are computed in the ambient space of the cobordism (*DirectedGeodesics,* **Algorithm 2**). Directed NN queries between boundary manifolds are weighted according to the native implementation in UMAP, the method of which we refer the reader to **Equations 5 and 6**. Resulting pairwise directed NN graphs between UMAP boundary manifolds are weighted, and they reflect Riemannian metrics that are not compatible. That is, they cannot be simply added or multiplied to integrate their weights. We therefore stitch the cobordism metric and make directed NN queries compatible by applying a fuzzy simplicial intersection, which results in a weighted, symmetric graph (*FuzzyIntersection,* **Algorithm 2**). The final cobordism produced by PatchMAP is obtained by taking a fuzzy union over the family of all fuzzy simplicial sets that represent cobordisms between boundary manifolds (*FuzzyUnion,*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Algorithm 2). To represent connections between boundary manifolds in PatchMAP cobordism projections, we implemented the hammer edge bundling algorithm in the Datashader Python library. Pseudo-code outlining the PatchMAP algorithm is shown in the following:

Algorithm 2: PatchMAP

Input: Data sets $(D_1, D_2 \ldots D_n)$, boundary manifold ambient metric (g_f), cobordism ambient metric (g)
Output: Cobordism $(\mathcal{W})$
function `Stitch`
    {*Compute boundary manifolds*}
    for $i \ldots n$ **do**
        $\mathcal{M}_i \leftarrow FuzzySimplicialSet\ (D_i, g_f)$ ▷ UMAP
    {*Compute geodesic projections in cobordism*}
    for $i \ldots n$ **do**
        $\mathcal{M}_{i\to j} \leftarrow DirectedGeodesics\ (D_j, g) : i \neq j$ ▷ Project data (UMAP *transform* ())
        $\mathcal{M}_{j\to i} \leftarrow DirectedGeodesics\ (D_i, g) : i \neq j$
        $\mathcal{M}_{ij} \leftarrow FuzzyIntersection\ (\mathcal{M}_{i\to j}, \mathcal{M}_{j\to i})$ ▷ t-norm
    {*Integrate Cobordism*}
    $\mathcal{W} \leftarrow FuzzyUnion\ \{\mathcal{M}_i, \mathcal{M}_{ij}\}$ ▷ t-conorm
    return $\mathcal{W}$

Domain/Information Transfer (i-PatchMAP). Let $\mathcal{M}_{rq}$ be a cobordism with boundary manifolds $\mathcal{M}_r$ and $\mathcal{M}_q$ obtained with PatchMAP that come a reference and query data set, respectively. Specifically, $\mathcal{M}_{rq}$ is a matrix where rows represent points in the reference boundary manifold, columns represent the nearest neighbors of reference manifold points in the query boundary manifold under a user defined metric, and the i, j^{th} entry represents the geodesics between points $p_i, p_j \in \mathcal{M}_r \coprod \mathcal{M}_q$ such that $p_i \in \mathcal{M}_r \to \mathcal{M}_r \coprod \mathcal{M}_q$ and $p_j \in \mathcal{M}_q \to \mathcal{M}_r \coprod \mathcal{M}_q$. We compute a new feature matrix of predictions P_q for the query data set by multiplying the feature matrix to be transferred, F, with the transpose of the weight matrix W_{rq} obtained through ℓ_1 normalization of $\mathcal{M}_{rq}$:

$$P_q = F{W_{rq}}^T \tag{16}$$

In this context, the matrix W_{rq} can be interpreted as a single-step transition matrix of a Markov chain between states $p_i$and p_j derived from geodesic distances on the cobordism $\mathcal{M}_{rq}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Biological Methods. All patient tissue samples were obtained with approval from the Institutional Review Boards (IRB) of Massachusetts General Hospital (protocol #2005P000774) and Beth Israel Deaconess Medical Center (protocol #2018P000581).

Generation of imaging mass cytometry data. Frozen tissues were sectioned serially at a thickness of 10 μm using a Microm HM550 cryostat (Thermo Scientific) and thaw-mounted onto SuperFrost^{TM} Plus Gold charged microscopy slides (Fisher Scientific). After temperature equilibration to room temperature, tissue sections were fixed in 4% paraformaldehyde (Ted Pella) for 10 min, then rinsed 3 times with cytometry-grade phosphate-buffered saline (PBS) (Fluidigm). Unspecific binding sites were blocked using 5% bovine serum albumin (BSA) (Sigma Aldrich) in PBS including 0.3% Triton X-100 (Thermo Scientific) for 1 hour at room temperature. Metal conjugated primary antibodies (Fluidigm) at appropriately titrated concentrations were mixed in 0.5% BSA in DPBS and applied overnight at 4 °C in a humid chamber. Sections were then washed twice with PBS containing 0.1% Triton X-100 and counterstained with iridium (Ir) intercalator (Fluidigm) at 1:400 in PBS for 30 min at room temperature. Slides were rinsed in cytometry-grade water (Fluidigm) for 5 min and allowed to air dry. Data acquisition was performed using a Hyperion Imaging System (Fluidigm) and CyTOF Software (Fluidigm), in 33 channels, at a frequency of 200 pixels/second and with a spatial resolution of 1 μm. Images were visualized with MCD Viewer software (Fluidigm) before exporting the data as text files for further analysis. After imaging, slides were rapidly stained with 0.1% toluidine blue solution (Electron Microscopy Sciences) to reveal gross morphology. Slides were digitized at a resolution of approximately 2.75 μm/pixel using a digital camera.

Generation of mass spectrometry imaging data. Paired 10μm thick sections from the same tissue blocks as used for imaging mass cytometry were thaw mounted onto Indium-Tin-Oxide (ITO) coated glass slides (Bruker Daltonics). Tissue sections were coated with 2.5-dihydroxybenzoic acid (40 mg/mL in 50:50 acetonitrile:water including 0.1% TFA) with an automated matrix applicator (TM-sprayer, HTX imaging). Mass spectrometry imaging of sections was performed using a rapifleX MALDI Tissuetyper (Bruker Daltonics, Billerica, MA). Data acquisition was performed using FlexControl software (Bruker Daltonics, Version 4.0) with the following parameters: positive ion polarity, mass scan range (*m/z*) of 300–1000, 1.25 GHz digitizer, 50 μm spatial resolution, 100 shots per pixel, and 10 kHz laser frequency. Regions of interest for data acquisition were defined using FlexImaging software (Bruker Daltonics, version 5.0), and individual images were visualized using both FlexImaging and SCiLS Lab (Bruker Daltonics). After data

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

acquisition, sections were washed with PBS and subjected to standard hematoxylin and eosin histological staining followed by dehydration in graded alcohols and xylene. The stained tissue was digitized at a resolution of 0.5 μm/pixel using an Aperio ScanScope XT brightfield scanner (Leica Biosystems).

Mass spectrometry imaging data preprocessing. Data were processed in SCiLS LAB 2018 using total ion count normalization on the mean spectra and peak centroiding with an interval width of $\pm$25mDa. For all analyses, a peak range of *m/z* 400-1,000 was used after peak centroiding, which resulted in 9,753 *m/z* peaks. No peak-picking was performed for presented data unless explicitly stated. Data were exported from SCiLS Lab as imzML files for further analysis and processing.

Single-cell segmentation. To quantify parameters of single-cells in IMC and registered MSI data within the DFU dataset, we performed cell segmentation on IMC ROIs using the pixel classification module in Ilastik (version 1.3.2)[41], which utilizes a random forest classifier for semantic segmentation. For each ROI, two 250 μm by 250 μm areas were cropped from IMC data and exported in the HDF5 format to use for supervised training. To ensure each cropped area was a representative training sample, a global threshold for each was created using Otsu thresholding on the Iridium (nuclear) stain with Scikit-image Python library. Cropped regions were required to contain greater than 30% of pixels over their respective threshold.

Training regions were annotated for "background", "membrane", "nuclei", and "noise". Random forest classification incorporated Gaussian smoothing features, edges features, including Laplacian of Gaussian features, Gaussian gradient magnitude features, and difference of Gaussian features, and texture features, including structure tensor eigenvalues and Hessian of Gaussian eigenvalues. The trained classifier was used to predict each pixels' probability of assignment to the four classes in the full images, and predictions were exported as 16-bit TIFF stacks. To remove artifacts in cell staining, noise prediction channels were Gaussian blurred with a sigma of 2 and Otsu thresholding with a correction factor of 1.3 was applied, which created a binary mask separating foreground (high pixel probability to be noise) from background (low pixel probability to be noise). The noise mask was used to assign zero values in the other three probability channels from Ilastik (nuclei, membrane, background) to all pixels that were considered foreground in the noise channel. Noise-removed, three-channel probability images of nuclei, membrane, and background were used for single-cell segmentation in CellProfiler (version 3.1.8)[53].

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single-cell parameter quantification. Single-cell parameter quantification for IMC and MSI data were performed using an in-house modification of the quantification (MCQuant) module in the multiple-choice microscopy software (MCMICRO)[24] to accept NIfFTI-1 files after cell segmentation. IMC single-cell measures were transformed using 99th percentile quantile normalization prior to downstream analysis.

Imaging mass cytometry cluster analysis. Cluster analysis was performed in Python using the Leiden community detection algorithm with the leidenalg Python package. UMAP's simplicial set (weighted, undirected graph) created with 15 nearest neighbors and Euclidean metric was used as input to community detection.

Microenvironmental correlation network analysis. To calculate associations across MSI and IMC modalities, we used Spearman's correlation coefficient in the Python Scipy library. *M/z* peaks from MSI data with no correlations to IMC data with Bonferroni corrected P-values above 0.001 were removed from the analysis. Correlation modules were formed with hierarchical Louvain community detection using the Scikit-network package. The resolution parameter used for community detection was chosen based on the elbow point of a graph plotting resolution vs. modularity of community detection results. UMAP's simplicial set, created with 5 nearest neighbors and the Euclidean metric, was used as input for community detection after inverse cosine transformation of Spearman's correlation coefficients to form metric distances[54]. Visualization of MSI correlation module trends to IMC parameters were computed using exponential-weighted moving averages in the Pandas library in Python after standard scaling IMC and MSI single-cell data. MSI moving averages were additionally min-max scaled to a range of 0-1 for plotting purposes. Differential correlations of variables u from MSI data and v from IMC data between conditions a and b were quantified and ranked using the formula given by Hsu et al.[55]:

$$D_{a\to b}^{u,v} = \left(\rho_b^{u,v} - \rho_a^{u,v}\right) \times \max\left(\left|\rho_a^{u,v}\right|, \left|\rho_b^{u,v}\right|\right) \quad (17)$$

where change in correlation coefficients for each pair u, v between conditions are weighted according to the maximal absolute correlation coefficient among both conditions. Significance of differential correlations $\rho_b^{u,v} - \rho_a^{u,v}$ were calculated using one-sided, Bonferroni corrected z-statistics after Fisher transformation.

Dimensionality reduction algorithm benchmarking. Methods used for benchmarking dimensionality reduction algorithms are outlined in **Supplementary Note 3, HDIprep dimension reduction validation.**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Spatial subsampling benchmarking. Default subsampling parameters in MIAAIM are based on experiments across IMC data from DFU, tonsil, and prostate cancer tissues recording Procrustes transformation sum of squares errors between subsampled UMAP embeddings with subsequent projection of out-of-sample pixels and full UMAP embeddings using all pixels. Spatial subsampling benchmarking was performed across a range of subsampling percentages.

Spectral landmark benchmarking. Subsampling percentages and dimensionalities to validate landmark-based steady-state UMAP dimensionalities were determined on a case-by-case basis from empirical studies and match those used for presented, registered data. Parameters were chosen due to the computational burden of computing values for cross-entropy on large data. To compare landmark steady-state dimensionality selections to subsampled data, we compared the shape of exponential regression fits from both sets of data using sum of squared errors. Sum of squared errors were calculated across a range of resulting landmarks.

Cobordism boundary manifold stitching simulation. Simulations were performed using the MNIST digits dataset in the Python Scikit-learn library using the default parameters for BKNN, Seurat v3, Scanorama, and PatchMAP across a range of nearest neighbor values. Data points were split into according to their digit label and stitched together using each method. Integrated data from each tested method excluding PatchMAP was then visualized with UMAP. Quality of boundary manifold stitching for each algorithm was quantified using the silhouette coefficient in the UMAP embedding space, implemented in Python with the Scikit-learn library. The silhouette coefficient is a measure of dispersion for a partition of a dataset. A high value indicates that data from the same label/type are tightly grouped together, whereas a lower value indicates that data from different types are grouped together. The silhouette coefficient (SC) is the average silhouette score s computed across each data point in the dataset, given by the following:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & if\ a(i) > b(i) \end{cases} \tag{18}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $a(i)$ is the average distance of data point i to all points with its label and $b(i)$ is the average distance of point i to all other data that do not have the same label.

CBMC CITE-seq data transfer. CBMC CITE-seq data were preprocessed according to the vignette provided by the Satija lab at https://satijalab.org/seurat/articles/multimodal_vignette.html. RNA profiles were log transformed and ADT abundances were normalized using centered log ratio transformation. RNA variable features were then identified, and the dimensionality of the RNA profiles of cells were reduced using principal components analysis. The first 30 principal components of single-cell RNA profiles were used to predict single-cell ADT abundances. The CBMC dataset was split randomly into 15 evaluation instances with 75% training and 25% testing data. Training data was used to predict test data measures. Prediction quality was quantified using Pearson's correlation coefficient between true and predicted ADT abundances. Correlations were calculated using the Python library, Scipy. Seurat was implemented using the TransferData function after finding transfer anchors (FindTransferAnchors function) using the default parameters. PatchMAP and UMAP+ were applied with 80 nearest neighbors and the Euclidean metric in the PCA space.

Spatially resolved image data transfer. To benchmark MSI to IMC information transfer, we performed leave-one-out cross validation with segmented single-cells from 23 image tiles (number of cells ranging from ~100 to ~500 each) from the DFU dataset. IMC ROIs were split into four evenly sized quadrants to create 24 tiles. One tile was removed due to lack of cell content. Data was transformed prior to information transfer using principal components analysis with 15 components using the Scikit-learn library. Seurat was implemented using the TransferData function after finding transfer anchors (FindTransferAnchors function) using the default parameters and 15 principal components. PatchMAP and UMAP+ were implemented using 80 nearest neighbors and the Euclidean metric in the PCA space. Information transfer quality was computed for each predicted IMC parameter by calculating Pearson's correlation in Python with the Scipy library between the Moran's autocorrelation index of ground truth data and predicted data. Moran's autocorrelation index (I) is given by the following[13]:

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{19}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where N is the number of spatial dimensions in the data (2 for our purposes), x is the abundance of a protein of interest, $\bar{x}$ is the mean abundance of protein x, w_{ij} is a spatial weight matrix, and W is the sum of all w_{ij}.

We acknowledge the Satija lab for utilizing spatial autocorrelation to quantify information transfer quality in spatial data[13]. We calculated Moran's I using the R package ape. Spatial weight matrices were calculated using the R package stats. To account for resolution differences between IMC and MSI imaging modalities, spatial weights for predicted IMC parameters were calculated using $w = \frac{1}{d^{50}}$ and weights for true IMC parameters were calculated using $w = \frac{1}{d}$ where d is the Euclidean distance between segmented cell centroids in the tiled image. Since there was no spatial deconvolution for spatial parameter predictions, IMC predictions were at a lower resolution than the true IMC data. Thus, a value of 50 was chosen to reduce the spatial weighting of predicted parameters and provide a fair comparison between true IMC data and predicted IMC data. Full tissue IMC parameter predictions for the DFU and tonsil biopsy tissue section were performed using the same parameters as the tiled ROIs from the benchmark study above.

Key Resources Table

Reagent or Resource	Source	Identifier
Biological Samples		
Diabetic Foot Ulcer IMC/MSI	Fresh operative tissue sample at BIDMC	N/A
Prostate Cancer MSI	Fresh operative tissue sample at MGH	N/A
Tonsil MSI	Fresh operative tissue sample at MGH	N/A
Antibodies		
anti-SMA 141 Pr	Fluidigm	Cat: 3141017D
anti-CD19 142 Nd	Fluidigm	Cat: 3142014D
anti-CD117 143 Nd	Fluidigm	Cat: 3143001B
anti-CD14 144 Nd	Fluidigm	Cat: 3144025D
anti-CD16 146 Nd	Fluidigm	Cat: 3146020D
anti-CD163 147 Sm	Fluidigm	Cat: 3147021D
anti-CD11b 149 Sm	Fluidigm	Cat: 3149028D
anti-CD86 150 Nd	Fluidigm	Cat: 3150020B
anti-CD31/PECAM-1 151 Eu	Fluidigm	Cat: 3151025D
anti-CD45 152 Sm	Fluidigm	Cat: 3152016D

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

anti-CD44 153 Eu	Fluidigm	Cat: 3153029D
anti-CD11c 154 Sm	Fluidigm	Cat: 3154025D
anti-FoxP3 155 Gd	Fluidigm	Cat: 3155016D
anti-CD4 156 Gd	Fluidigm	Cat: 3156033D
anti-CD73 158 Gd	Fluidigm	Cat: 3158031D
anti-CD68 159 Tb	Fluidigm	Cat: 3159035D
anti-CD20 161 Dy	Fluidigm	Cat: 3161029D
anti-Pan-Keratin 162 Dy	Fluidigm	Cat: 3162033D
anti-TGF-β 163 Dy	Fluidigm	Cat: 3163010B
anti-Arginase-1 164 Dy	Fluidigm	Cat: 3164027D
anti-β-Catenin 165 Ho	Fluidigm	Cat: 3165032D
anti-IL-10 166 Er	Fluidigm	Cat: 3166008B
anti-Ki-67 168 Er	Fluidigm	Cat: 3168022D
anti-Collagen type 1 169 Tm	Fluidigm	Cat: 3169023D
anti-CD3 170 Er	Fluidigm	Cat: 3170019D
anti-Caspase-3 172 Yb	Fluidigm	Cat: 3172027D
anti-Pan-Actin 175 Lu	Fluidigm	Cat: 3175032D
anti-Histone-3 176 Yb	Fluidigm	Cat: 3176023D
Critical Software/Algorithms		
MIAAIM	This study	https://github.com/JoshuaHess12/miaaim and https://github.com/JoshuaHess12/miaaim-python
HDIprep application	This study	https://github.com/JoshuaHess12/hdi-prep
HDIreg application	This study	https://github.com/JoshuaHess12/hdi-reg
(i)-PatchMAP application	This study	https://github.com/JoshuaHess12/patchmap
HDIutils application	This study	https://github.com/JoshuaHess12/hdi-utils
Elastix	Klein and Staring, et al.[30]	https://github.com/SuperElastix/elastix
Docker	Merkel[32]	https://www.docker.com
Nextflow	Tommaso et al.[33]	https://www.nextflow.io
MCMICRO	Schapiro et al.[24]	https://mcmicro.org
SCiLS Lab (2018)	Bruker	https://scils.de
Ilastik	Berg et al.[41]	https://www.ilastik.org
CellProfiler	McQuin et al.[53]	https://cellprofiler.org
Leiden Algorithm	Traag et al.[56]	https://leidenalg.readthedocs.io/en/stable/intro.html
Scipy	Virtanen et al.[57]	https://www.scipy.org/index.html
UMAP	McInnes and Healy[35]	https://umap-learn.readthedocs.io/en/latest/
Scikit-learn	Pedregosa et al.[58]	https://scikit-learn.org/stable/
Scikit-image	van der Walt et al.[59]	https://scikit-image.org
Scikit-network	Bonald et al.[60]	https://scikit-network.readthedocs.io/en/latest/
Seurat v3	Stuart et al.[13]	https://satijalab.org/seurat/
BBKNN	Polański et al.[43]	https://github.com/Teichlab/bbknn

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Scanorama	Hie et al.[42]	https://github.com/brianhie/scanorama

13. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
18. Moon, K.R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* **37**, 1482-1492 (2019).
24. Schapiro, D. et al. MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging. *bioRxiv* (2021).
30. Klein, S., Staring, M., Murphy, K., Viergever, M.A. & Pluim, J.P. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging* **29**, 196-205 (2009).
32. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal* **2014**, 2 (2014).
33. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nature biotechnology* **35**, 316-319 (2017).
35. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
36. Sainburg, T., McInnes, L. & Gentner, T.Q. Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning. *arXiv preprint arXiv:2009.12981* (2020).
38. Costa, J.A. & Hero, A.O. in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3 iii-988 (IEEE, 2004).
39. Costa, J.A. & Hero, A.O. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* **52**, 2210-2221 (2004).
40. Staring, M., Van Der Heide, U.A., Klein, S., Viergever, M.A. & Pluim, J.P. Registration of cervical MRI using multifeature mutual information. *IEEE transactions on medical imaging* **28**, 1412-1421 (2009).
41. Berg, S. et al. ilastik: Interactive machine learning for (bio) image analysis. *Nature Methods*, 1-7 (2019).
42. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology* **37**, 685-691 (2019).
43. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2020).
46. Armstrong, M.A. Basic topology. (Springer Science & Business Media, 2013).
47. Sun, K. & Marchand-Maillet, S. in International Conference on Machine Learning 1-9 (PMLR, 2014).
48. Jayasumana, S., Hartley, R., Salzmann, M., Li, H. & Harandi, M. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE transactions on pattern analysis and machine intelligence* **37**, 2464-2477 (2015).
49. Yukich, J.E. Probability theory of classical Euclidean optimization problems. (Springer, 2006).
50. Hero, A.O., Ma, B., Michel, O.J. & Gorman, J. Applications of entropic spanning graphs. *IEEE signal processing magazine* **19**, 85-95 (2002).
51. Narayan, A., Berger, B. & Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 1-10 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. Moser, J. On the volume elements on a manifold. *Transactions of the American Mathematical Society* **120**, 286-294 (1965).
53. McQuin, C. et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS biology* **16**, e2005970 (2018).
54. Van Dongen, S. & Enright, A.J. Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv preprint arXiv:1208.3145* (2012).
55. Hsu, C.-L., Juan, H.-F. & Huang, H.-C. Functional analysis and characterization of differential coexpression networks. *Scientific reports* **5**, 1-14 (2015).
56. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1-12 (2019).
57. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).
58. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
59. Van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
60. Bonald, T., de Lara, N., Lutz, Q. & Charpentier, B. Scikit-network: Graph Analysis in Python. *Journal of Machine Learning Research* **21**, 1-6 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data availability statement

All newly generated data from this study are available at https://www.synapse.org/MIAAIM_manuscript.

Code availability statement

All software and code that produced the findings of the study, including all main and supplemental figures, are free and publicly available via GitHub at https://github.com/JoshuaHess12/miaaim-development. A user manual and training guide to get started using the MIAAIM software and associated workflows is available at https://joshuahess12.github.io/miaaim-python. Python modules that allow for individual applications of MIAAIM workflows can be found at https://github.com/JoshuaHess12/hdi-prep, https://github.com/JoshuaHess12/hdi-reg, and https://github.com/JoshuaHess12/patchmap. A Python package implementing MIAAIM is available at https://github.com/JoshuaHess12/miaaim-python.

Author contribution statements

J.M.H., M.C.P., P.M.R. and R.F.S. designed the study. J.M.H. conceived and designed software and algorithms. R.F.S., P.M.R., and M.C.P. supervised the work. R.F.S. and P.M.R. conceived and performed MSI, IMC, and H&E multimodal imaging. J.M.H. performed processing and analysis of MSI, IMC, H&E, and CyCIF images. J.M.H., R.F.S., I.I., M.C.P. and P.M.R. designed benchmark studies and interpreted analyses. J.M.H. and D.S developed and implemented cell segmentation and quantification pipeline. J.M.H, J.J.I, and I.I. critically reviewed mathematics of the paper. M.S.R., G.T., C.L.W., A.V., N.Y.R.A. provided MSI, IMC, and H&E images. W.M.A. and A.E.S. provided critical input. J.M.H., I.I., M.C.P, P.M.R., and R.F.S. wrote the manuscript. J.M.H, I.I., M.C.P., P.M.R, and R.F.S. finalized the manuscript with input from all authors.

ACKNOWLEDGEMENTS

J.M.H, M.C.P., P.M.R. and R.F.S. were supported by the Vaccine and Immunotherapy Center Innovation and Education Funds at Massachusetts General Hospital. J.M.H was previously supported and PMR is also supported by a Congressionally Directed Medical Research Program grant W81XWH-20-1-0301. J.M.H. is supported by a National Science Foundation Graduate Research Fellowship under grant no. 1746886. D.S. was supported by the University of Zurich BioEntrepreneur-Fellowship (BIOEF-17-001), a Swiss National Science Foundation Early Postdoc Mobility fellowship (P2ZHP3_181475, BMBF (01ZZ2004) and was a Damon Runyon Fellow supported by the Damon Runyon Cancer Research

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Foundation (DRQ-03-20). N.Y.R.A was partially supported by NIH grants U54-CA210180 and P41-EB028741. We thank Enterprise Research Infrastructure & Services at Partners HealthCare for their in-depth support and provision of computing resources.

OUTSIDE INTERESTS

D.S. is a scientific consultant to Roche Glycart AG; none of these relationships are directly related to the topic of this study. R.F.S., J.M.H., P.M.R, and M.C.P. are authors on patent PCT/US21/48928 applied for by the General Hospital Corporation covering the described methods.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Fig. 1 | Summary of MIAAIM workflow and key outputs

A schematic representation of the MIAAIM workflow to integrate level 2 multi-modal images derived from various imaging methodologies (cyan boxes) and tissue specimen types (red boxes) to enable tissue state mapping across modalities and tissues. Key inputs and outputs (yellow boxes) of each MIAAIM module (shaded boxes) are connected through MIAAIM's Nextflow implementation (solid arrows) or exploratory analysis modules in Python (dashed arrows). Algorithms unique to MIAAIM (orange boxes) are detailed in corresponding figures in the remainder of this study (black bolded text). Dataflows associated with high-parameter images (i.e., greater than 3 channels) follow the left-side arrows through the HDIprep (green) and HDIreg (navy blue) modules for pixel-level image compression and pixel manifold-based image registration, while low-parameter images follow the right-side arrows (HDIprep and HDIreg modules, white boxes). Aligned multi-modal data may be exported in various file formats (HDIreg, left-side yellow boxes), and MIAAIM's Nextflow implementation of the Docker containerized HDIprep and HDIreg modules enables level 2 data integration reproducibility (HDIreg, right-side yellow boxes). MIAAIM's nonparametric data integration allows for human-in-the-loop quality control of aligned multi-modal data (HDIreg, right-side white box). External analysis techniques that interface with MIAAIM's key outputs are included in the schematic (bottom right white boxes). Aligned multi-modal data or multi-modal single-cell data (i.e., from cell segmentation) across tissue types and tissue states can be modeled using the PatchMAP module (pink), and data can be transferred across tissue states and modalities using the i-PatchMAP module (light blue).

Fig. 2 | HDIprep compression and spatially constrained manifold alignment with HDIreg

a. HDIprep compression algorithm. **(i)** High-dimensional modality. **(ii)** Spatial subsampling. **(iii)** Data manifold represented as a fuzzy simplicial set (UMAP). Edge bundled connectivity of the manifold is shown on two axes of the resulting steady state embedding (*fractal-like structure may not reflect biologically relevant features). **(iv)** High-connectivity landmarks identified with spectral clustering. **(v)** Spectral landmarks are embedded into a range of dimensionalities and exponential regression identifies steady-state dimensionalities. Pixel locations are used to reconstruct compressed image. **b.** HDIreg manifold alignment. **(i)** Spatial transformation is optimized to align moving image to fixed image. KNN graph lengths between resampled points (yellow) are used to compute α-MI. Edge-length distribution panels show Shannon MI between distributions of intra-graph edge lengths at resampled locations before

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and after alignment (α-MI converges to Shannon MI as $\alpha \longrightarrow 1$). MI values show increase in information shared between images after alignment. KNN graph connections show correspondences across modalities. **(ii)** Optimized transformation aligns images. Shown are results of transformed H&E image (green) to IMC (red). **c.** Example alignment. **(i)** Full-tissue MSI-to-H&E registration produces T_0. **(ii)** H&E is transformed to IMC full-tissue reference, producing T_1. **(iii)** ROI coordinates extract underlying MSI and IMC data in IMC reference space. (**iv**) H&E ROI is transformed to correct in IMC domain, producing T_2. Final alignment applies modality-specific transformations. Shown are results for an IMC ROI.

Fig. 3 | Microenvironmental correlation network analysis (MCNA) links protein expression with molecular distributions in the DFU niche

a. MCNA UMAP of *m/z* peaks grouped into modules. **b.** Exponential-weighted moving averages of normalized ion intensities for top five positive and negative correlates to proteins. Colors indicate module assignment. Heatmaps (right) indicate Spearman's rho. **c.** Exponential-weighted moving averages of normalized average ion intensity per modules ordered as distance from center of wound in DFU increases. **d.** Raw IMC nuclear (Ir) and CD3 staining in ROI (left) (scale bars = 80 μm). Masks showing CD3 expression (middle-left). Aligned MSI showing one of top CD3 correlates (middle-right). Overlay of CD3 expression and a top molecular correlate (right). **e.** Same as **d** at different ROI. **f.** Unsupervised phenotyping. Shaded box indicates CD3+ population. Heatmap indicates normalized protein expression. **g.** MCNA UMAP colored to reflect ions' correlations to Ki-67 within CD3+ and CD3- populations. Colors indicate Spearman's rho and size of points indicates negative log transformed, Benjamini-Hochberg corrected P-values for correlations. **h.** Tornado plot showing top five CD3+ differential negative and positive correlates to Ki-67 compared to the CD3- cell populations. X-axis indicates CD3+ specific Ki-67 values. Color of each bar indicates change in correlation from CD3- to CD3+ populations. **i.** Boxplots showing ion intensity and of top differentially correlated ions (top, positive; bottom; negative) to CD3+ specific Ki-67 expression across ROIs on the DFU. Tissue maps of top differentially associated CD3+ Ki-67 correlates (top, positive; bottom; negative) with boxes (white) indicating ROIs on the tissue that contain CD3+ cells.

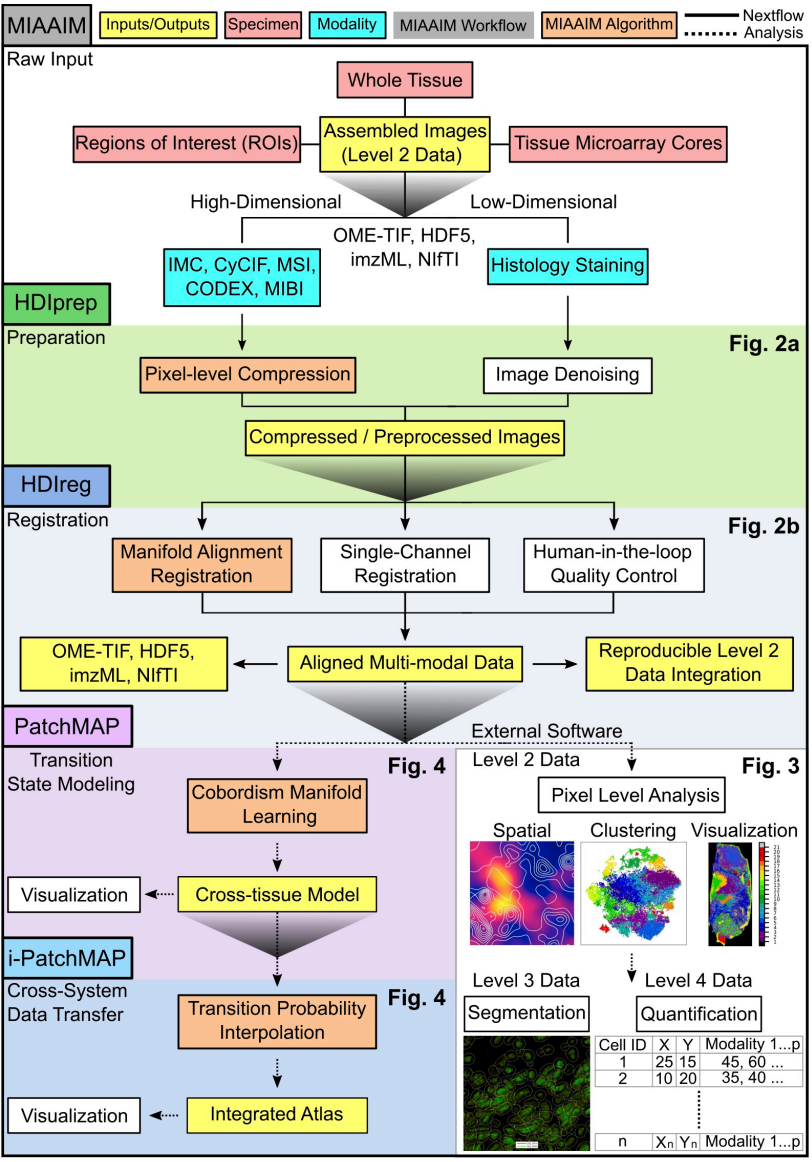
Fig. 4 | Cobordism projection with PatchMAP and domain transfer with i-PatchMAP

a. Schematic representing the PatchMAP algorithm's cobordism geodesic stitching between boundary manifolds (reference and query data) to form cobordism (represented as grey surface), information transfer

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

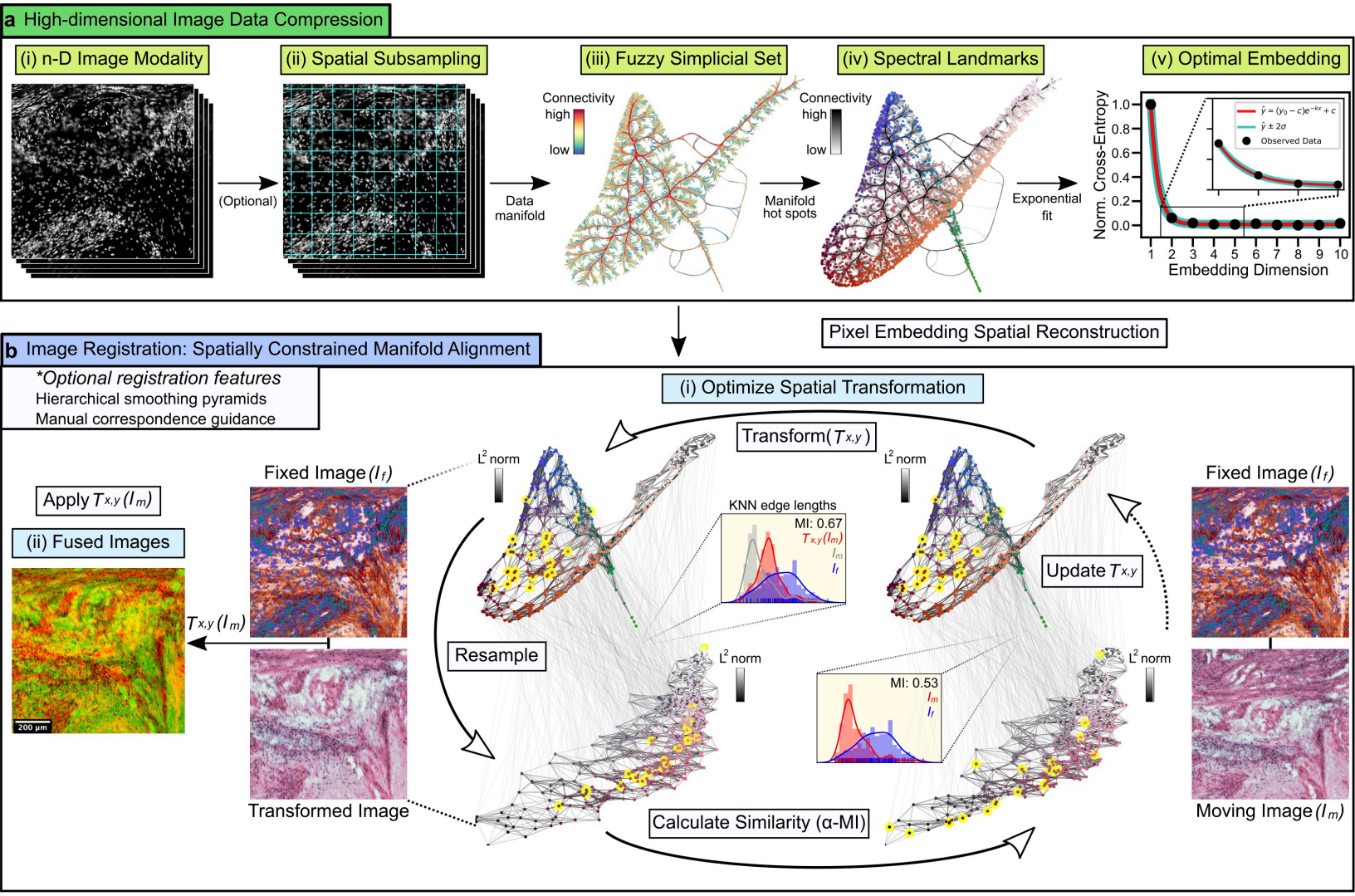
across cobordism geodesics with the i-PatchMAP workflow (top right), and cobordism projection visualization (bottom). In PatchMAP, cobordisms sharing common boundary manifolds are glued along their shared domain, and geodesic distances in the cobordism are made compatible with boundary manifold geodesics. **b.** Boundary manifold stitching simulation. PatchMAP projection (manually drawn dashed lines indicate stitching) and UMAP projections of integrated data are shown at NN values that maximized the SC for each method. **c.** MSI-to-IMC data transfer with i-PatchMAP. Line plots show Spearman's rho between predicted and true spatial autocorrelation values. **d.** MSI-to-IMC data transfer benchmark. **e.** CBMC multimodal CITE-seq data transfer benchmark. **f.** PatchMAP of DFU segmented single-cells (blue), DFU full tissue (red), tonsil tissue (green), and prostate tissue (orange) pixels based on MSI profile. Individual plots show IMC expression for DFU single-cells (right). **g.** MSI-to-IMC data transfer from DFU single-cells to the DFU full tissue using i-PatchMAP. **h.** MSI-to-IMC data transfer from DFU single-cells to the tonsil tissue.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



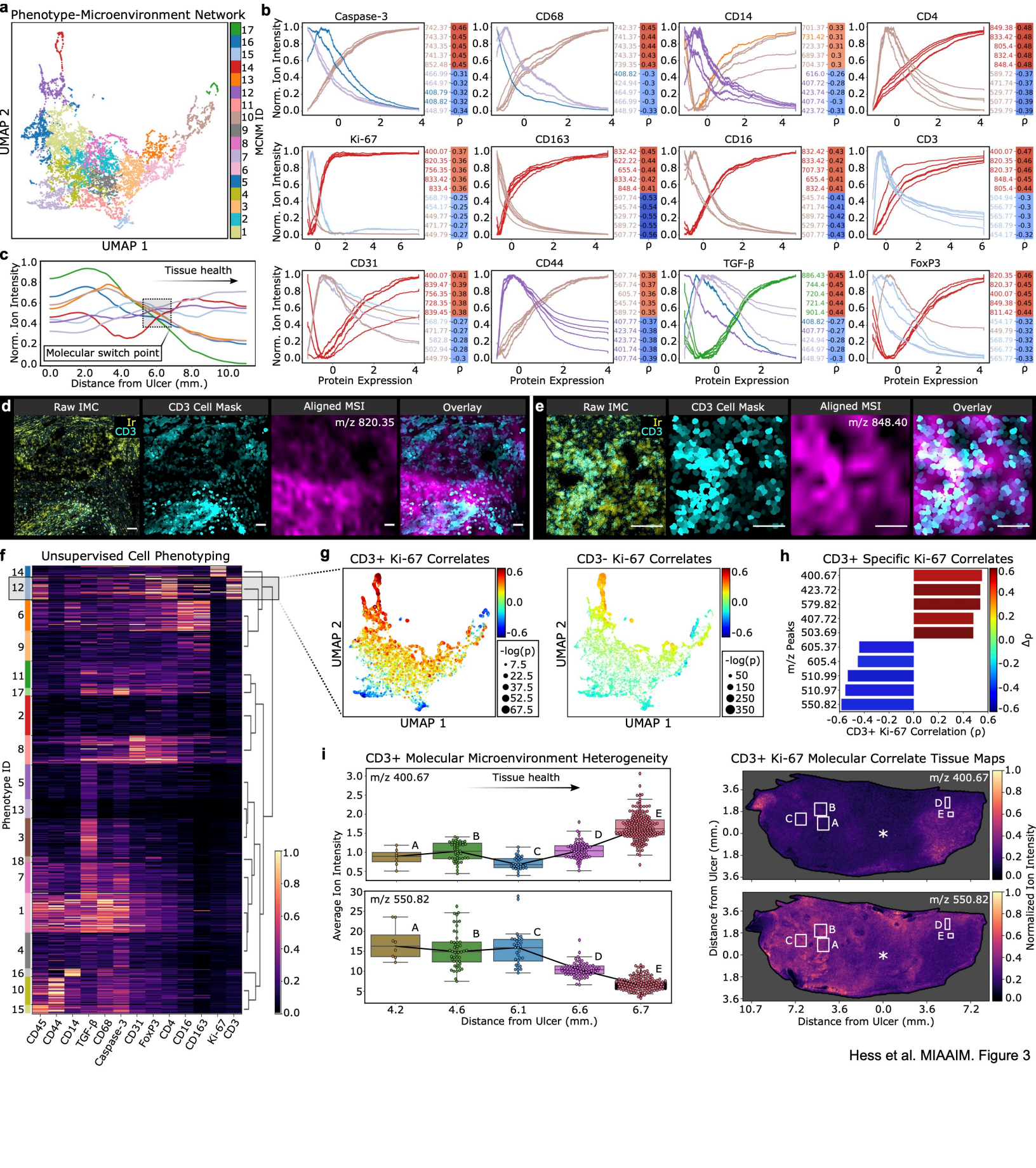
Hess et al. MIAAIM. Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



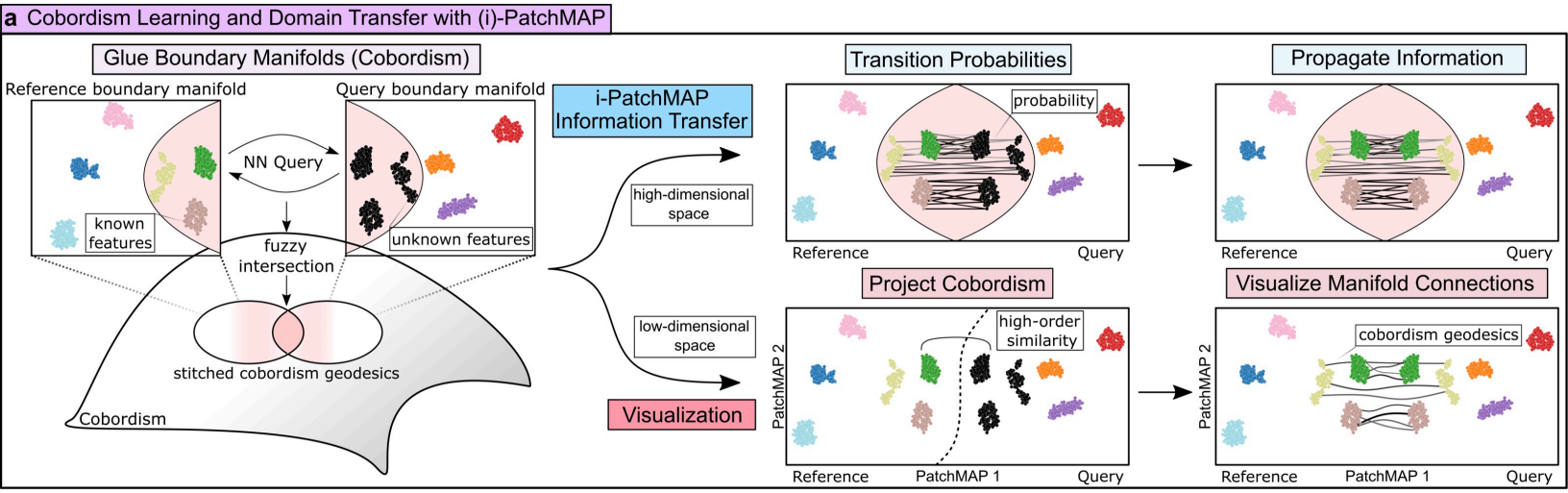
Hess et al. MIAAIM. Figure 2

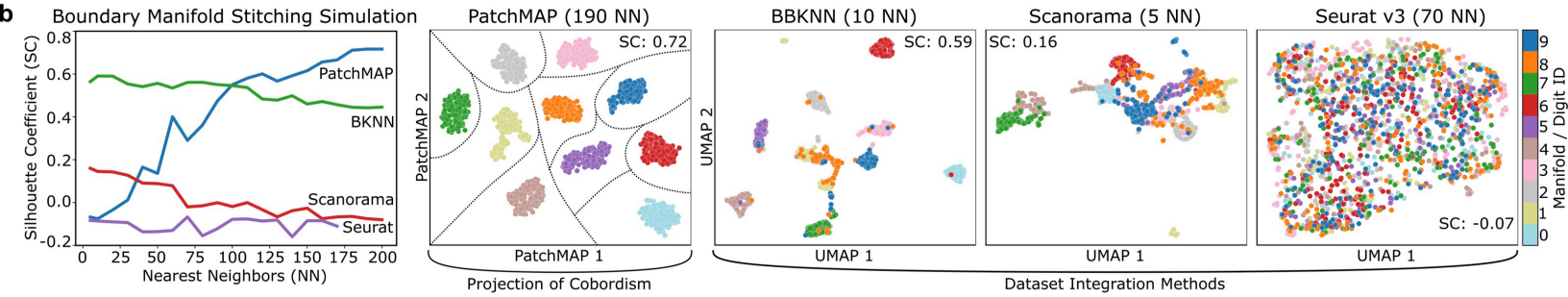
bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

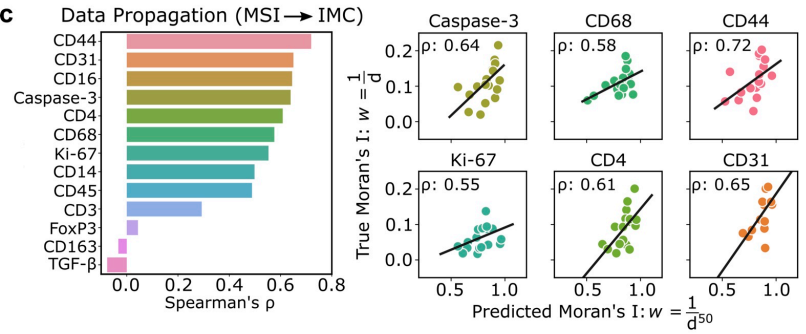


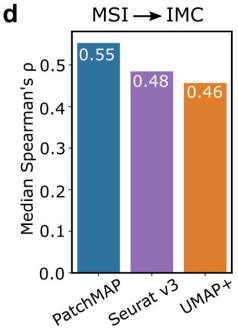
Hess et al. MIAAIM. Figure 3

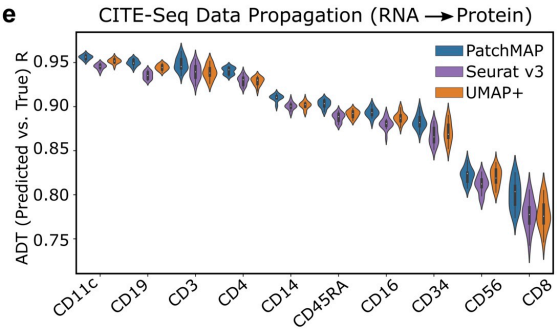
bioRxiv preprint doi: https://doi.org/10.1101/2021.12.20.472858; this version posted December 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

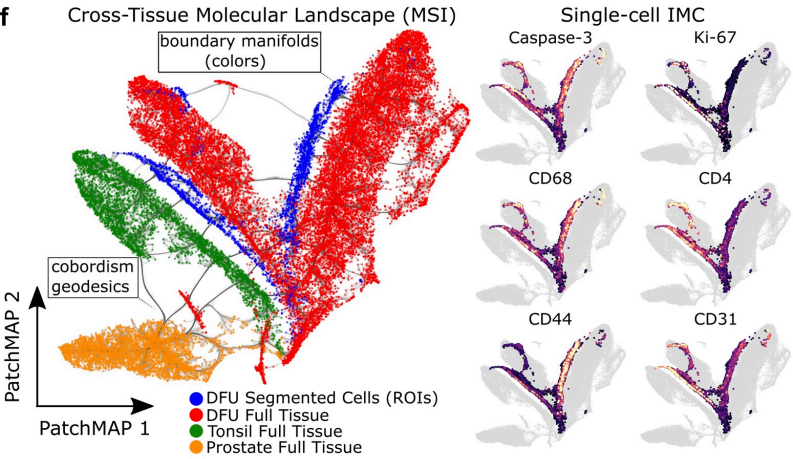


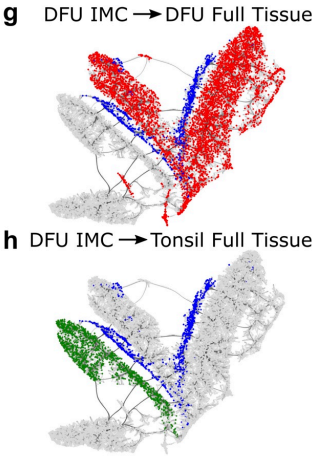


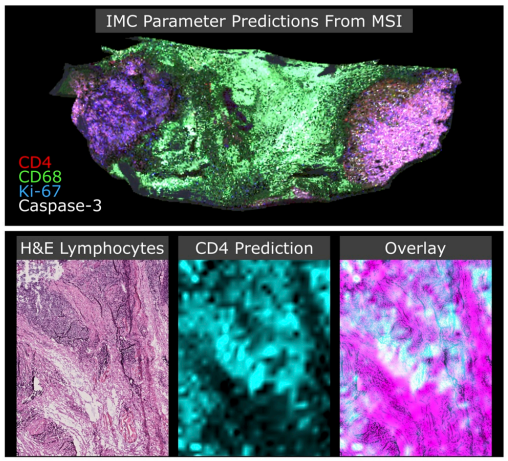












Hess et al. MIAAIM. Figure 4