Imputing Single-Cell Protein Abundance in Multiplex Tissue Imaging

Raphael Kirchgaessner[1,2], Cameron Watson[1,2], Allison Creason[1,2], Kaya Keutler[3,] and Jeremy Goecks[1,4]

[1]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239
[2]The Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239
[3]Department of Chemical Physiology and Biochemistry, Oregon Health & Science University, Portland, OR 97210, USA
[4]Department of Machine Learning, Moffitt Cancer Center, Tampa, FL 33612
*Corresponding author: jeremy.goecks@moffitt.org

Abstract: Multiplex tissue imaging are a collection of increasingly popular single-cell spatial proteomics and transcriptomics assays for characterizing biological tissues both compositionally and spatially. However, several technical issues limit the utility of multiplex tissue imaging, including the limited number of RNAs and proteins that can be assayed, tissue loss, and protein probe failure. In this work, we demonstrate how machine learning methods can address these limitations by imputing protein abundance at the single-cell level using multiplex tissue imaging datasets from a breast cancer cohort. We first compared machine learning methods' strengths and weaknesses for imputing single-cell protein abundance. Machine learning methods used in this work include regularized linear regression, gradient-boosted regression trees, and deep learning autoencoders. We also incorporated cellular spatial information to improve imputation performance. Using machine learning, single-cell protein expression can be imputed with mean absolute error ranging between 0.05-0.3 on a [0,1] scale. Our results demonstrate (1) the feasibility of imputing single-cell abundance levels for many proteins using machine learning to overcome the technical constraints of multiplex tissue imaging and (2) how including cellular spatial information can substantially enhance imputation results.

Introduction

Multiplex tissue imaging (MTI) are a set of single-cell spatial proteomics and transcriptomics assays for highly detailed profiling of biological tissues. With MTI, single-cell abundance levels and spatial distribution of 10-150 of proteins and/or 500-2000 RNAs can be quantified simultaneously (Francisco-Cruz et al., 2020; Sheng et al., 2023). MTI enables characterization of individual cells as well as tissue organization, and MTI has been used in studies of healthy tissue (Neumann et al., 2022), COVID (Werlein et al., 2023), cancer (Lewis et al., 2021), and other diseases (Kitko et al., 2022; McCaffrey et al., 2022; Sepe et al., 2022). There are many MTI platforms, including cyclic immunofluorescence (CycIF) (Lin et al., 2018), CO-Detection by indEXing (CODEX) (Black et al., 2021), CosMx (Z. R. Lewis et al., 2022), Xenium (Janesick et al., 2022) and multiplex immunohistochemistry (Tsujikawa et al., 2017). MTI has been used to generate large datasets in NIH consortia such as the NIH Human BioMolecular Atlas Program (HuBMAP Consortium, 2019) and the NCI Cancer Moonshot Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020). MTI is also an increasingly common assay in cancer (Tan et al., 2020), where it has proven important for quantifying tumor spatial organization and microenvironment heterogeneity (Blise et al., 2022) and connecting these features to cancer subtypes, prognosis, and therapy response (Friebel et al., 2020; Steele et al., 2020).

However, several key factors limit the usefulness of multiplex tissue imaging. Only 10-150 proteins and/or several thousand RNAs can be assayed in a single experiment, and hence the information obtained from a single experiment is bounded. Further, MTI assays can suffer from several technical issues that reduces the information obtained, including tissue loss or folding, probe failure, illumination artifacts, or errors in downstream image processing. These limitations have a profound impact on MTI data quality and substantially reduce the overall utility of MTI. To mitigate these limitations and improve utility of MTI, machine learning and deep learning approaches can be used to computationally increase the numbers of proteins/RNAs available from MTI and mitigate assay failures. Computationally increasing—or imputing—additional data by filling in missing data with predicted values is already common in other molecular assays, such as single cell RNA sequencing (scRNA) (Chen et al., 2022; Gong et al., 2018; He et al., 2020; Kharchenko et al., 2014; Talwar et al., 2018; Tran et al., 2022; van Dijk et al., 2018; Xu et al., 2021), bulk genomics (Qiu et al., 2020), and bulk transcriptomics (Patruno et al., 2021). While imputation has been applied to MTI images (Sims & Chang, 2023; Ternes et al., 2021), to the best of our knowledge imputation on MTI single-cell datasets has not been explored. Imputing single-cell data is especially valuable because single-cell datasets require fewer computational resources to process than images and can be readily integrated with other molecular datasets.

In this study, we applied machine learning (ML) and deep learning (DL) methods to impute protein abundance in tissue-based cyclic immunofluorescence (t-cycIF) (Lin et al., 2018) datasets obtained from breast cancer tissues. We evaluated the performance of ML and DL methods to predict protein abundance levels in t-cycIF single-cell datasets that included 20 proteins. Three distinct ML/DL approaches—regularized linear regression, gradient-boosted trees, and auto encoders—were used and evaluated for in-patient and across-patient imputation. Spatial information was introduced as well to improve imputation results. Overall, our results indicate that accurate imputation is possible for many proteins and that spatial information significantly improves imputation results.

Results

Study Cohort and Analysis Overview:

The multiplexed tissue imaging single-cell datasets used in this study were generated using a 20-plex t-CycIF (Lin et al., 2018) protein panel applied to a cohort of hormone receptor-positive (HR+), HER-2 negative metastatic breast cancer biopsies. The tissue biopsies and datasets are part of the NCI Cancer Moonshot Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020) and have detailed associated clinical metadata. Our

Table 1: Overview of proteins assayed using t-cycIF and their use as functional or lineage proteins. Lineage proteins are used to identify cell types whereas functional proteins are used to characterize cell function.

Lineage Protein	Functional Protein
CD45	Ki67
αSMA	pERK
eCadherin	PR
CK19	EGFR
CK14	p21
CK17	pRB
CK7	AR
Vimentin	HER2

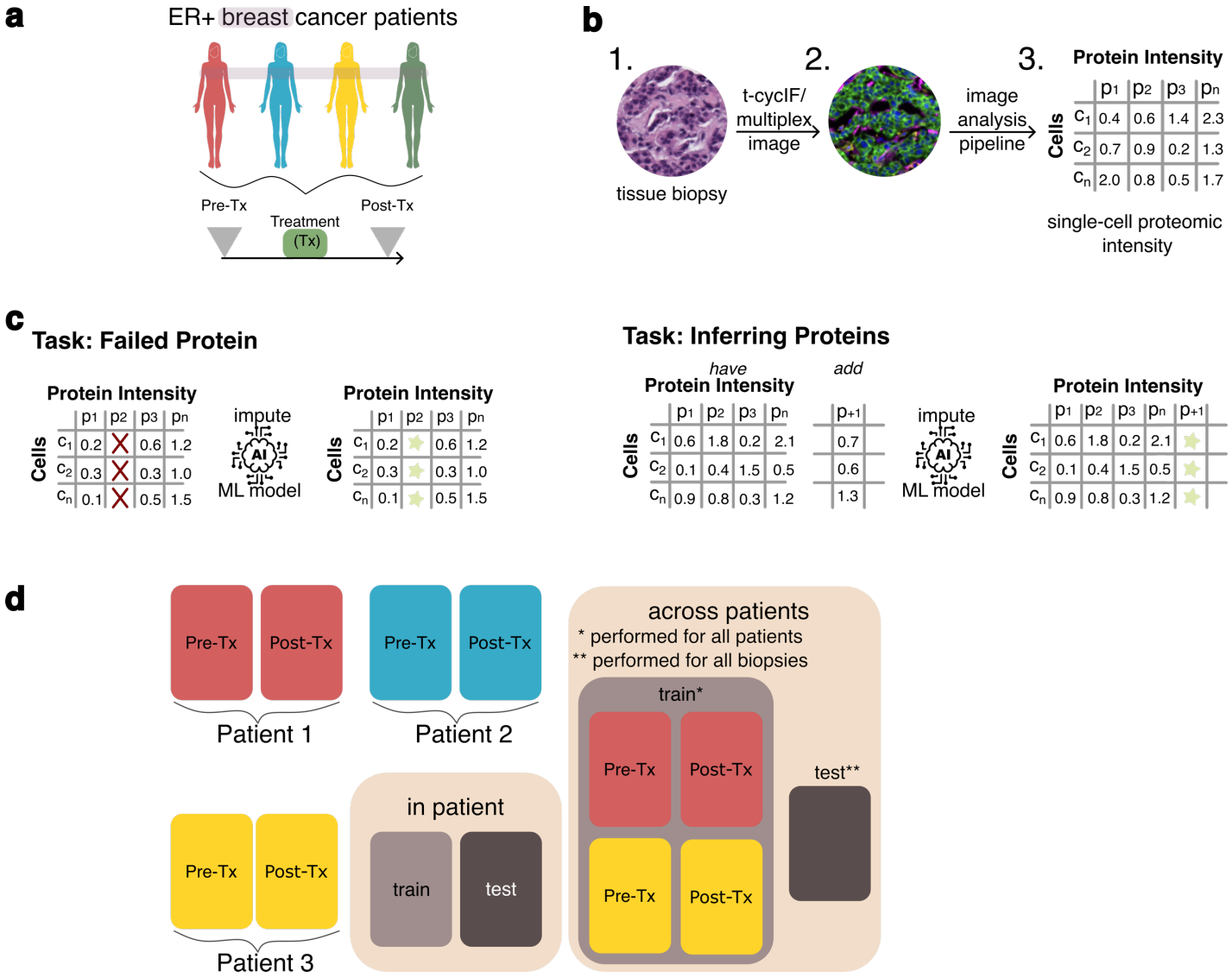

Figure 1: Overview of dataset, study motivations, and analysis approaches. a: Biopsies were obtained from four HR+ breast cancer patients before and after the same standard-of-care therapy for a total of eight biopsies. **b:** each biopsy was assayed using the multiplexed tissue imaging assay t-cycIF to quantify abundance levels of 20 proteins and then processed using an image analysis pipeline to create single-cell feature tables (total number of cells identified: 475359); **c:** the key tasks addressed by this work are imputing failed proteins and inferring additional proteins not present in an MTI experiment; **d:** approaches for training and testing ML models for imputing proteins within and across patients.

dataset includes a total of 8 biopsies derived from 4 patients (**Fig. 1a**) that received a CDK4/6 inhibitor in combination with endocrine therapy, which is a common combination therapy in metastatic HR+ breast cancer. Each patient contributed a pair of biopsies, a pre-treatment biopsy and a biopsy taken at the time of tumor progression. Image stacks collected from t-CycIF were processed using the MCMICRO image analysis pipeline (Schapiro et al., 2021) to generate single-cell feature tables (**Fig. 1b**). Each row in the table is a single cell identified in the image, and the table columns are the protein abundance levels calculated via mean pixel intensity per cell. In total 475359 single cells were identified across all biopsies, with an average of 59400 cells per biopsy. To perform in-patient evaluation, either the pretreatment or the posttreatment biopsy was used for training a machine learning model while the remaining biopsy was used for testing model performance. Biopsy timing was not used in this study. In total 16 proteins were shared between all biopsies, including 8 proteins for identifying cell types (lineage proteins) and 8 proteins for characterizing cellular functional states (functional proteins) (**Table 1**).

The imputation task in this study was to predict protein abundance levels for a withheld protein or set of proteins. For each machine learning experiment, one or more proteins was withheld and used as the target variable(s) for the predictive model, and the remaining protein abundances were used as input features for the model. This task simulates the key application for imputation in multiplexed tissue imaging: computationally increasing proteins not originally included in an MTI assay or inferring protein levels where the assay failed (**Fig. 1c**). Three machine learning methods were used for imputation: elastic-net (EN) regularized linear regression (Su et al., 2012), light gradient-boosting machine (LGBM) (Ke et al., 2017), and neural network autoencoders (AE) (Zhai et al., 2018). Each model has advantages and disadvantages. As a linear model, EN is constrained to learning linear relationships between model inputs and outputs whereas LGMB and AE can learn non-linear relationships. EN and LGBM are easy to set up and minimal adjustments are required to get started. However, EN and LGBM can impute only one protein per model, resulting in the need to train a model for each protein of interest. This in turn requires a time-consuming process of training many models. In contrast, a single AE model can be trained once and then used to impute any single protein or even multiple proteins at once. However, due to the compression and reconstruction techniques applied in AEs, they sometimes approximate data with some loss of precision, which may lead to reduced performance compared to EN and LGBM approaches.

Imputation model training and evaluation was performed both within patients and across patients **(Fig. 1d).** For the in-patient (IP) strategy, a model was trained on a single patient biopsy, and then model performance was measured using the other biopsy of the same patient. For the across-patient (AP) strategy, a model was trained on all biopsies except those from a single patient, and then model performance was evaluated using the biopsies of the single patient held out of training. The IP approach was anticipated to produce optimal performance because inter-patient variability is not present to reduce accuracy, whereas the AP approach is likely to be less accurate due to patient variability. Importantly, the AP approach approximates real-world usage. We evaluated the model performance using the mean absolute error (MAE). Preprocessing was performed to remove all columns from the datasets except protein intensities, followed by a min-max scaling approach, which maps values between 0 and 1. Thus, model error is in the range [0,1] where lower error represents better performance.

Protein abundance imputation with elastic net and light gradient-boosting machines

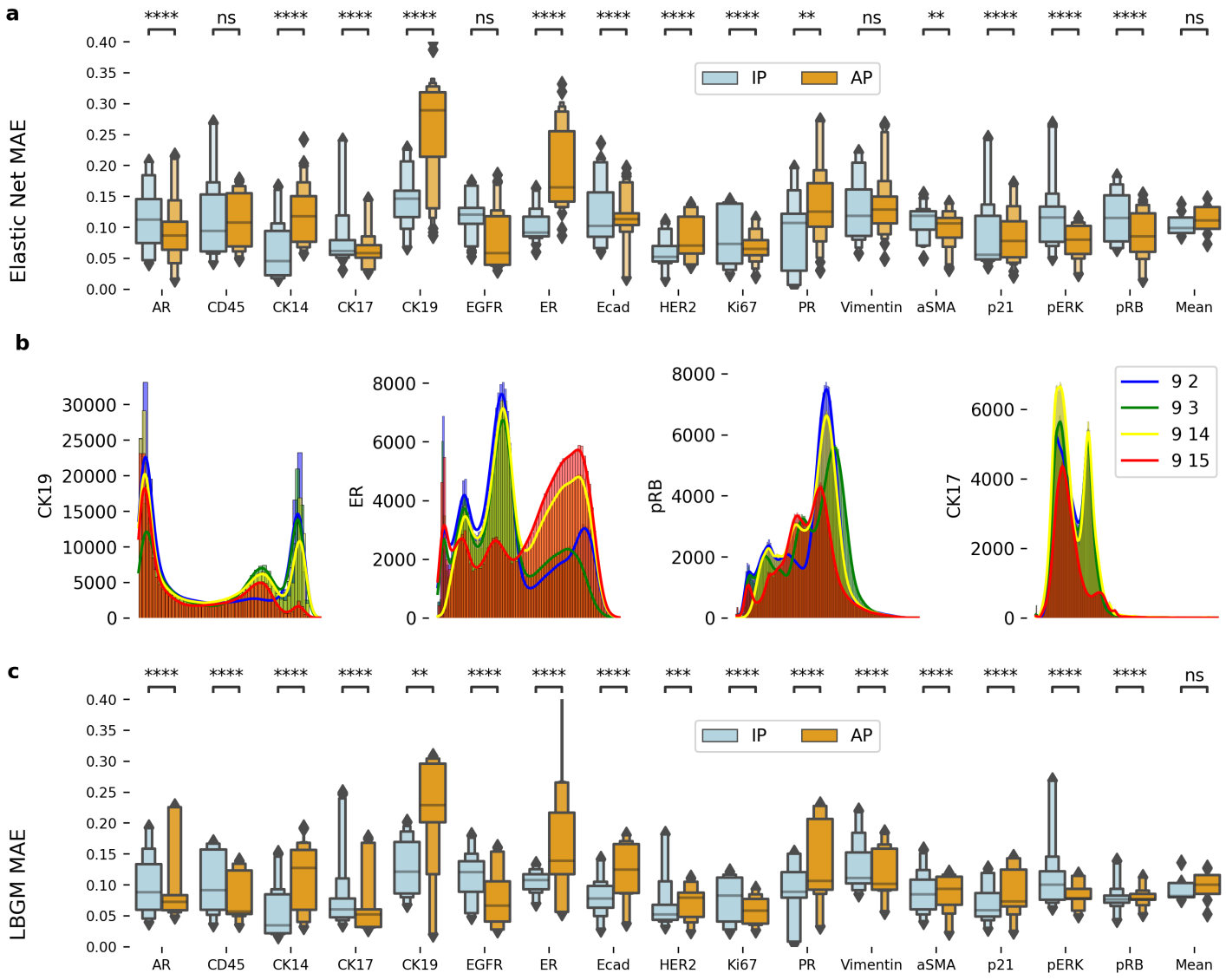

Figure 2: Imputation results for elastic net and Light GBM machine learning models within and across patients. **a:** imputation results using elastic net for all proteins demonstrates low mean absolute error for many proteins (scale is [0,1]). **b**: patient protein expression for four proteins (CK19, ER, pRB, CK17) is highly variable; **c:** imputation results using LGBM show lower MAE than elastic net.
p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

We used the elastic net (EN) regularized linear regression approach as a baseline model because it is simple, robust, and widely used in modeling high-dimensional omics datasets (Vinga, 2021). EN imputation accuracy ranged from 0.05 and 0.20 MAE in both IP and AP experiments for twelve of sixteen proteins (**Fig. 2a**), demonstrating that it is possible to accurately impute many protein abundance levels. Proteins CK17 and Ki67 were most accurately imputed with MAE of 0.05. Proteins where imputation MAE was worse than 0.2 included CK19, ER, CK14, and PR. CK19 (1.28) and ER (0.32) protein expression exhibited some of the highest variances (**Table S3**). These markers are likely to exhibit patient-specific patterns that may be difficult for ML models to learn. In addition, for these proteins the AP MAE was notably higher than the IP setting. Reduced imputation performance for these proteins can be explained by noting that HR+ breast cancers are enriched in these proteins, and patient-specific patterns in these proteins are common. Wide-spread abundance levels of these proteins and patient-specific patterns prevent accurate imputation (**Fig. 2b**). Unexpectedly, imputation accuracy for proteins EGFR, Ki67, and Vimentin was higher in AP experiments compared to the IP experiments.

Using Light Gradient Boosting Machine (LGBM) yielded improved imputation accuracy compared to EN. Imputation accuracy differences between IP and AP analyses using the LGBM were less pronounced relative to EN, especially for proteins CK19 and ER (**Fig. 2c)**. Like the EN, LGBM performance for the same twelve of sixteen proteins was between 0.05 and 0.20 MAE in IP and AP settings. LGBM imputation accuracy for CK19 and ER are like the EN and greater than 0.2 MAE. Overall, LGBM displayed more accurate imputation results than EN both within and across patients (**Table 2**). Of particular significance were imputation accuracy for Ki67, pERK, EGFR, and AR in the AP context, which show improved performance compared to the IP setting. The difference between IP and AP performance of

Table 2: Mean and standard deviation of performance for EN, LGBM and AEs in IP and AP settings.

Model/Network	EN	LGBM	AE Single	AE Multi
IP Mean (Std)	0.10 (0.05)	0.09 (0.05)	0.12 (0.07)	0.12 (0.08)
AP Mean (Std)	0.11 (0.07)	0.10 (0.06)	0.13 (0.09)	0.13 (0.09)

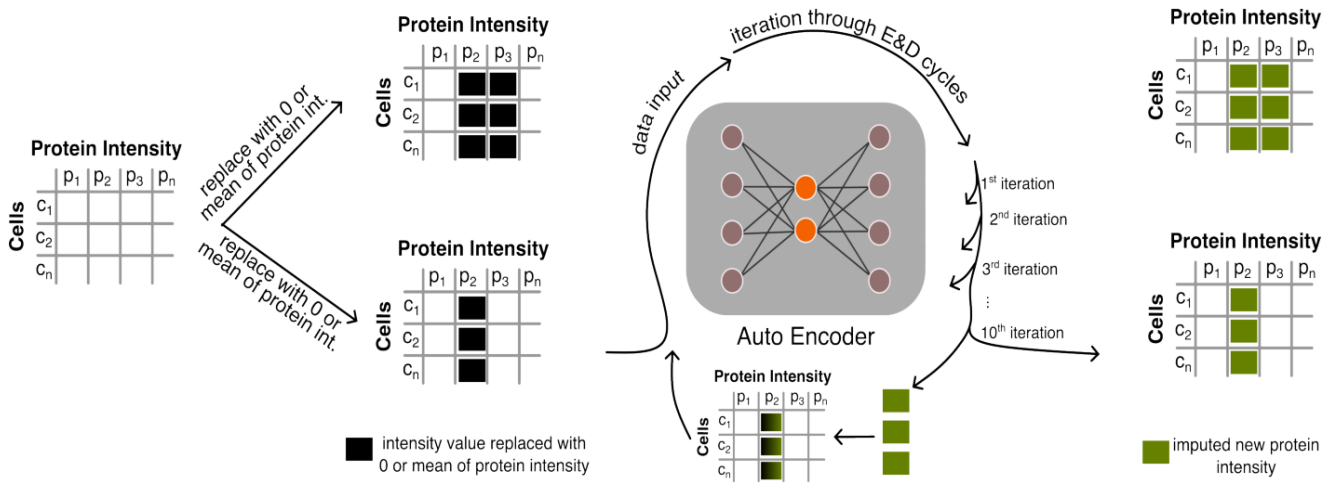

Figure 3: **Using a deep learning autoencoder for imputation via iterative analyses**. The autoencoder is trained and then uses an iterative approach to impute single or multiple proteins. To start, proteins to be imputed are replaced with either zero or the mean of the intensity values in the training set. Then, the autoencoder is used iteratively to predict protein intensities using output values as new input values for each iteration.

the CK19 protein is also much smaller when using LGBM compared to the EN. Importantly, LGBM imputation accuracy in the AP experiments is comparable to that in IP experiments, and mean accuracy as well as the standard deviation across all proteins was improved relative to the EN model.

<u>Protein abundance imputation using autoencoders and all model comparisons:</u>

Auto Encoders (AE) are deep learning neural networks for accurate reconstruction of high-dimensional data that include two distinct components: (1) an encoder network that maps a high-dimensional input to a lower-dimensional representation in a latent space and (2) a decoder network that reconstructs the original high-dimensional input from the low-dimensional latent space representation. The goal of an AE is to perform information-preserving dimensionality reduction of its input to the latent space so that it can then accurately reconstruct the input from the latent representation. AEs have been successfully used for imputation in various biological domains, including single-cell RNA (Grønbech et al., 2020; Hou et al., 2020; Lotfollahi et al., 2020; Talwar et al., 2018) and genomics (Qiu et al., 2020) and more (McCoy et al., 2018).

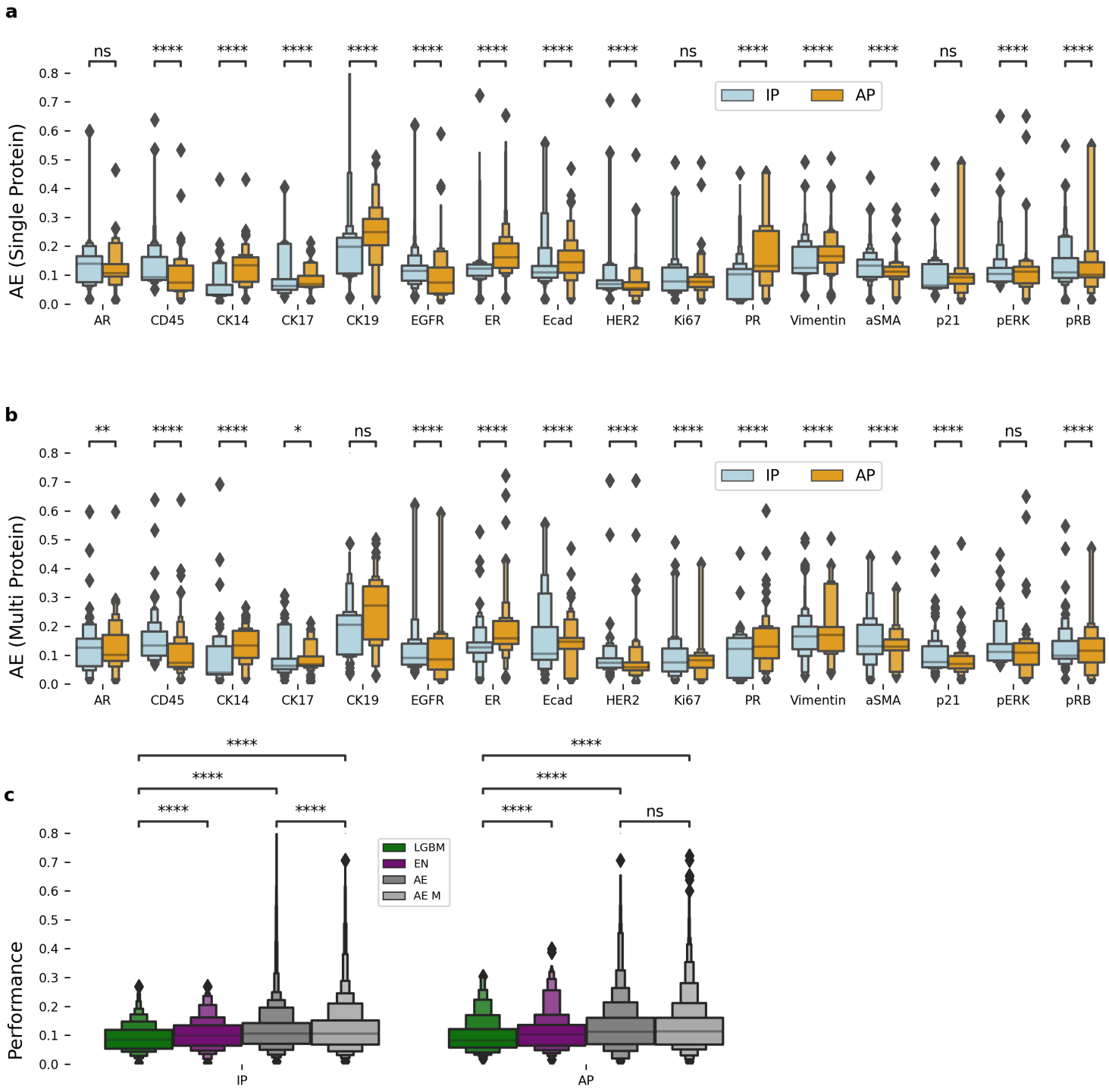

Figure 4: Autoencoder imputation results and performance comparison between machine learning models. **a**: AE single protein imputation performance **b**: AE multi-protein imputation performance. **c**: performance comparison between all evaluated ML models shows similar performance overall and that LGBM performs best, followed by EN and finally AE. There is no significant difference between single and multi-protein imputation performance for AE in an AP setting.
p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

Unlike LGBM and EN models, AEs can impute multiple features simultaneously because their output is a full reconstruction of the entire input. Leveraging this capability, we performed both single-protein and multi-protein imputation experiments. Multi-protein imputation was performed based on the order in which the proteins were assayed during t-CycIF's multiple imaging rounds. After training the AEs, we

used a standard iterative approach for imputing cellular protein abundances using three steps: (i) for each target protein to be predicted, start by replacing the target(s) abundance with 0 or the mean protein abundance from the training dataset; (ii) repeatedly use the AE to predict protein abundances and then use these predictions as new inputs to the AE; and lastly (iii) use the last AE prediction for each protein as the final AE prediction (**Fig. 3)**.

AEs accurately imputed proteins in both single and multi-protein experiments (**Fig. 4a, Fig. 4b**). Imputation accuracy within patients is better than across patients, but the difference in accuracy for most proteins is less than 0.05 MAE. Imputation accuracy of CK19 levels is between 0.15-0.35 MAE, while imputation of the best performing proteins, CK17 and p21, is between 0.05 and 0.10 MAE. Like the EN

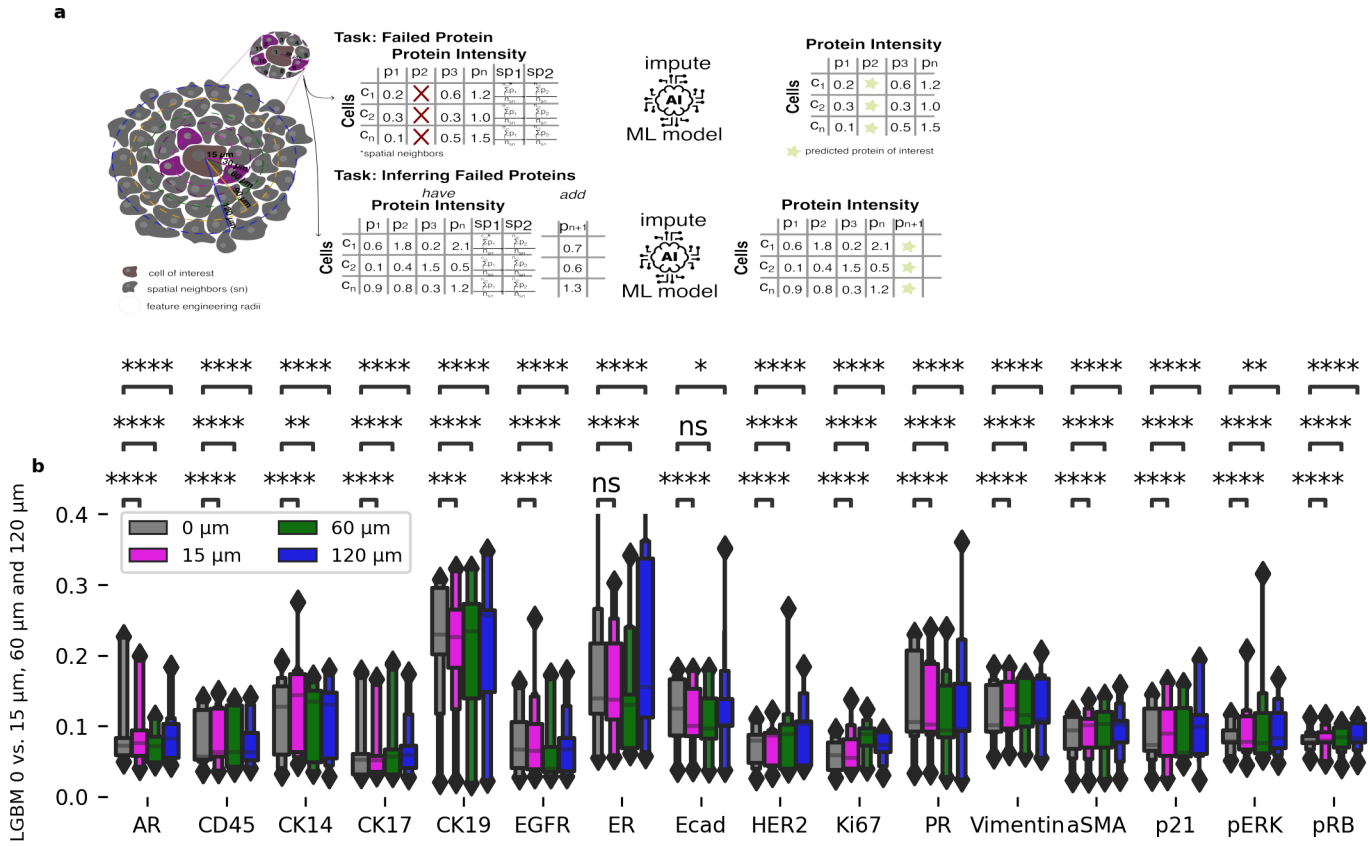

Figure 5: Using spatial information improves imputation performance for LGBM. a: Schematic for creating a feature table based on spatial neighbors found in selected radii. Exemplary 15 µm radius is shown. Red marks the cell of interest (or origin) and protein abundance levels of cells in its neighborhood are averaged to get neighborhood abundance levels. **b:** LGBM imputation results across patients with MAE scores for 0 µm,15 µm, 60 µm and 120 µm reveal significant improvement for several proteins. p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-04

and LGBM models, imputation performance is worst for the proteins with the most variable abundance levels in our breast cancer cohort, including CK19, ER, and PR. Performance differences between the imputation of single proteins and that of multi-proteins are minimal (**Fig. 4c**).

We next compared performance for all three machine learning models used for imputation. Overall, LGBM performed best, followed by the EN and the AEs. These performance differences are consistent in both IP and AP settings (**Fig. 4c**). However, performance differences between the models are relatively modest, with the LGBM achieving a mean accuracy of 0.10 MAE in the AP experiments, followed by the EN with a mean accuracy of 0.11 MAE, and the AEs with a mean accuracy of 0.13 MAE.

Using cellular spatial information to improve imputation:

A key advantage of multiplexed tissue imaging datasets is that the spatial coordinates of each cell are known, making it possible to quantify spatial information around individual cells. We hypothesized that the spatial information available in t-CycIF could be used to improve imputation performance. To test this hypothesis, we quantified the spatial cellular context surrounding a target cell by calculating the mean protein abundance of neighboring cells. Average abundance levels of all proteins in neighboring cells were then added to our prior set of input features to create a feature set that includes both single-cell protein abundances plus average neighbor abundances (**Fig. 5a**). When no neighboring cells were

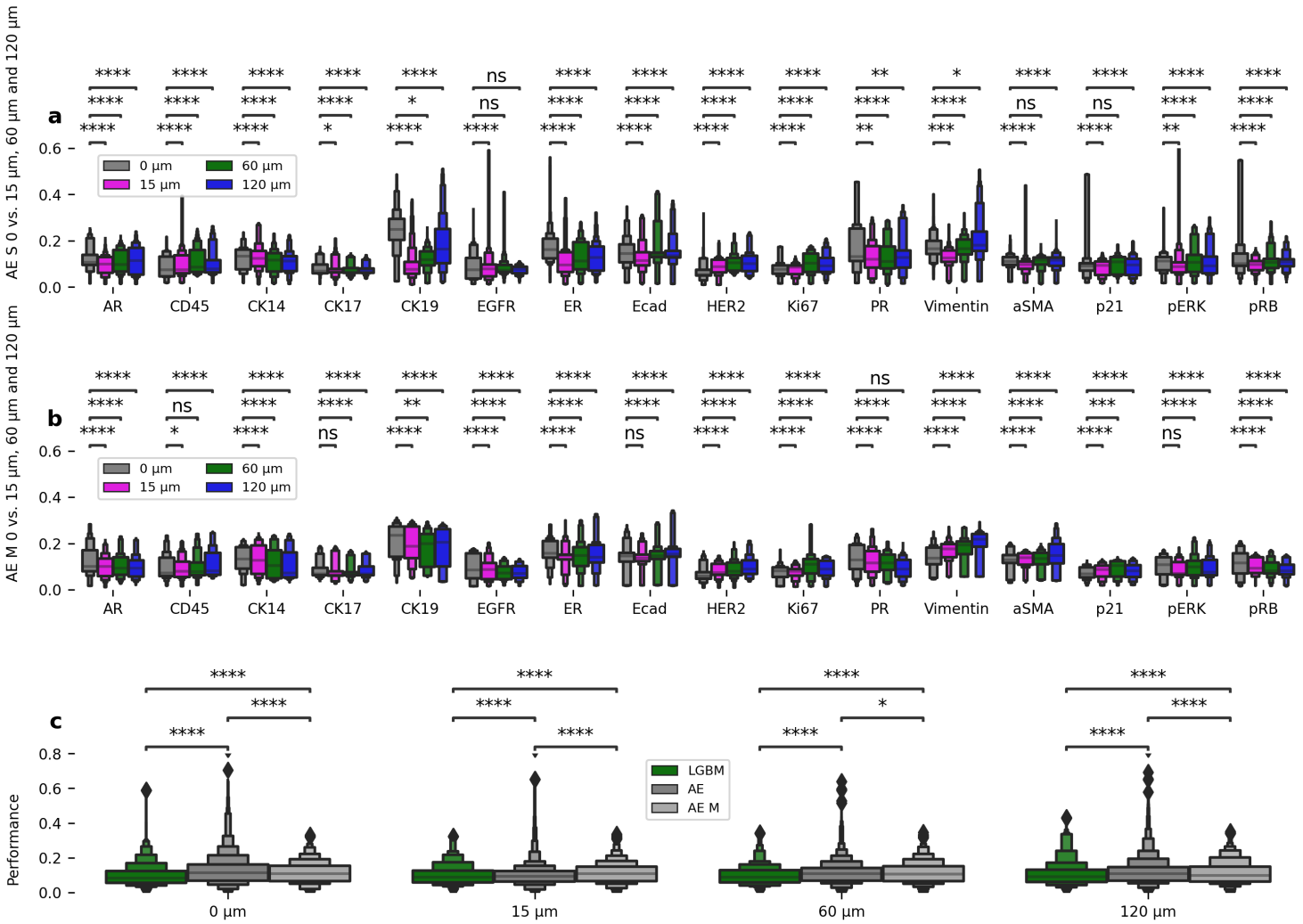

Figure 6 Autoencoder imputation performance improves significantly using spatial information across multiple radii. **a:** Single protein imputation MAE in an AP approach for 0 µm, 15 µm, 60 µm and 120 µm leads to improved imputation accuracy. **b:** Multi-protein imputation MAE scores in an AP setting for 0 µm, 15 µm, 60 µm and 120 µm leads to improved imputation accuracy. **c:** Comparison of LGBM and AE imputation performance for 0,15, 60 and 120µm shows similar performance of all models. p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-04

detected, a value of 0 was assigned for neighbors' protein abundances. Radii of 15, 30, 60, 90 and 120 micrometers were used to identify neighboring cells and assess the impact of using different sizes of radii on imputation performance. Only features for one radius setting were used for training a model, and hence a single set of spatial features was included as input for a predictive model.

A radius of 15 micrometers captures most of the immediate neighbors of a cell, whereas larger radii capture the extended neighborhood of a cell.

In these imputation experiments, only LGBM and AEs were used due to LGBM's superior performance compared to ENs and AE's multi-protein imputation capability. Using spatial information improved overall imputation accuracy. Importantly, imputation accuracy for the proteins that had proven difficult to impute due to their very high levels of variance (**Supplements**) - CK19, ER, and PR - was improved significantly with spatial information (**Fig. 5b**). LGBM performance also improved for other proteins such as CK17, CD45, Ecad, ASMA and p21. Performance of the AE achieved improvements in single-protein imputation for most markers, with CK19 showing the greatest improvement (**Fig. 6a**). Multi-protein imputation also benefited from spatial information integration (**Fig. 6b**), however performance gains were not as pronounced compared to the single protein imputation model. Aligned with prior research (Fischer et al., 2022), imputation accuracy generally improves up to a certain neighborhood radius and then plateaus or declines (**Table 3, S1, and S2**). However, the LGBM does not show the same improvement up until a certain radius, but instead remains largely steady, with a peak performance observed using 60μm (**Table 3, S1, and S2**). Performance in both IP and AP imputation is improved by incorporating spatial information (**Table S1 and S2**).

Table 3: Mean and standard deviation of performance for LGBM and AEs for different radii. Mean and (standard deviation) for each model are listed. 0μm is considered as baseline without any use of spatial information.

Network	**0 μm [Baseline]**	**15 μm**	**60 μm**	**120 μm**
LGBM	0.10 (0.06)	0.10 (0.06)	0.097 (0.05)	0.11 (0.07)
AE Single Protein	0.13 (0.09)	0.10 (0.06)	0.113 (0.06)	0.12 (0.07)
AE Multi Protein	0.12 (0.09)	0.11 (0.07)	0.117 (0.07)	0.12 (0.08)

Discussion

In this study, we used machine learning models to accurately impute single-cell protein abundance levels using datasets obtained from the t-CycIF multiplexed tissue imaging assay. Our datasets included eight biopsies from a cohort of four metastatic breast cancer patients, making it possible to train and evaluate ML models for both in-patient (IP) and across patient (AP) imputation. Within a range of [0,1], imputation performance ranged from 0.05-0.15 mean absolute error for most proteins. Proteins with particularly high variance in our cohort, such as CK19 and ER, were more difficult to impute and imputation MAE for these proteins were 0.15-0.35. The LGBM model, a gradient-boosted regression tree approach, was the best performing model amongst all tested ML algorithms. Making spatial features available to ML models, in the form of neighboring cells' protein abundance levels, improved their performance with a reduction of 0.02 in MAE on average. Improved performance was especially notable for proteins with high variance that were otherwise difficult to impute. This result complements recent research indicates that cell communication may vary and requires careful evaluation using multiple cellular neighborhoods (Fischer et al., 2022). Our results are concordant with this observation as they show a similar pattern of improved protein performance using a diverse set of radii.

While the LGBM shows the overall best performance, there are tradeoffs to consider when choosing a machine learning model for imputation for multiplexed tissue imaging datasets. Traditional ML models such as LGBM and EN can only impute one protein per model, which requires training and storing a model for each protein to be imputed. Using a single model for each protein is time and cost inefficient. In contrast, an AE is capable of imputing multiple proteins at once and can also impute all proteins included in their training data, requiring only a single training session and model. While AEs perform marginally worse than LGBM and EN for protein imputation, their capability for multi-protein imputation offers an advantage in reduced training time and cost. Multi-protein imputation, as opposed to sequential imputation, also models inter-protein relationships, and potentially yields more biologically pertinent relationships to explore.

Limitations of this work include the small number of proteins used for imputation, the cohort composition of metastatic breast cancers, and the small sample size. This analysis used the sixteen proteins that were shared amongst all biopsies, and it is uncertain if other proteins can be imputed as accurately as these sixteen. This analysis also focused on breast cancer biopsies and diseased tissue, and imputation results may be different in healthy tissue or in other diseases. Furthermore, this study uses data derived from four patients. Data from these patients may not be representative of a larger patient cohort, and hence it is difficult to understand how well our results generalize. A study like ours would benefit from using a larger and more diverse cohort, potentially with more proteins and different multiplexed tissue imaging assays.

In summary, this study demonstrates that machine learning can be used to accurately impute single-cell protein abundance levels within and across patients using multiplexed tissue imaging datasets. Three diverse ML models were all able to impute single-cell protein abundance levels well. Finally, incorporating spatial information significantly improves imputation performance, suggesting that local cellular neighborhoods are important determinants of cellular protein abundance levels.

Data availability

The datasets used for this study are available in DataVerse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/RBIJSQ&version=1.0 and Table 4 lists the HTAN Biopsy and Biospecimen IDs used in this study. More information on these biopsies and biospecimens is available at https://humantumoratlas.org/explore

Table 4: Human Tumor Atlas Network (HTAN) biopsy and biospecimen IDs.

HTAN Biopsy ID	HTAN Biospecimen ID
9 2 1	HTA9_2_11
9 2 2	HTA9_2_21
9 3 1	HTA9_3_11
9 3 2	HTA9_3_21
9 14 1	HTA9_14_6
9 14 2	HTA9_14_14
9 15 1	HTA9_15_7
9 15 2	HTA9_15_15

Code availability

The source code of this work is freely available in the GitHub repository.
https://github.com/goeckslab/MTIProteinImputation

References:

Black, S., Phillips, D., Hickey, J. W., Kennedy-Darling, J., Venkataraaman, V. G., Samusik, N., Goltsev, Y., Schürch, C. M., & Nolan, G. P. (2021). CODEX multiplexed tissue imaging with DNA-conjugated antibodies. In *Nature Protocols* (Vol. 16, Issue 8, pp. 3802–3835). Nature Research. https://doi.org/10.1038/s41596-021-00556-8

Blise, K. E., Sivagnanam, S., Banik, G. L., Coussens, L. M., & Goecks, J. (2022). Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. *Npj Precision Oncology*, *6*(1). https://doi.org/10.1038/s41698-022-00253-z

Chen, Y., Wang, Y., Chen, Y., Cheng, Y., Wei, Y., Li, Y., Wang, J., Wei, Y., Chan, T. F., & Li, Y. (2022). Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-34550-9

Fischer, D. S., Schaar, A. C., & Theis, F. J. (2022). Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01467-z

Francisco-Cruz, A., Parra, E. R., Tetzlaff, M. T., & Wistuba, I. I. (2020). Multiplex Immunofluorescence Assays. In *Methods in Molecular Biology* (Vol. 2055, pp. 467–495). Humana Press Inc. https://doi.org/10.1007/978-1-4939-9773-2_22

Friebel, E., Kapolou, K., Unger, S., Núñez, N. G., Utz, S., Rushing, E. J., Regli, L., Weller, M., Greter, M., Tugues, S., Neidert, M. C., & Becher, B. (2020). Single-Cell Mapping of Human Brain Cancer Reveals Tumor-Specific Instruction of Tissue-Invading Leukocytes. *Cell*, *181*(7), 1626-1642.e20. https://doi.org/10.1016/j.cell.2020.04.055

Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., & Garry, D. J. (2018). DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2226-y

Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., & Winther, O. (2020). scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, *36*(16), 4415–4422. https://doi.org/10.1093/BIOINFORMATICS/BTAA293

He, Y., Yuan, H., Wu, C., & Xie, Z. (2020). DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02083-3

Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02132-x

HuBMAP Consortium. (2019). The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. In *Nature* (Vol. 574, Issue 7777, pp. 187–192). Nature Publishing Group. https://doi.org/10.1038/s41586-019-1629-x

Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Rouault, M., Beliakoff, G., Faria De Oliveira, M., Kohlway, A., Abousoud, J., Morrison, C. A., Drennon, Y., Mohabbat, S. H., Williams, S. R., Teams, D., & Taylor, S. E. B. (2022). *High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue*. https://doi.org/10.1101/2022.10.06.510405

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. https://github.com/Microsoft/LightGBM.

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742. https://doi.org/10.1038/nmeth.2967

Kitko, C. L., Arora, M., DeFilipp, Z., Abu Zaid, M., Di Stasi, A., Radojcic, V., Betts, C. B., Coussens, L. M., Meyers, M. L., Qamoos, H., Ordentlich, P., Kumar, V., Quaranto, C., Schmitt, A., Gu, Y., Blazar, B. R., Wang, T. P., Salhotra, A., Pusic, I., … Lee, S. J. (2022). Axatilimab for Chronic Graft-Versus-Host Disease After Failure of at Least Two Prior Systemic Therapies: Results of a Phase I/II Study. *J Clin Oncol*, *41*, 1864–1875. https://doi.org/10.1200/JCO.22

Lewis, S. M., Asselin-Labat, M. L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., & Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods 2021 18:9*, *18*(9), 997–1012. https://doi.org/10.1038/s41592-021-01203-6

Lewis, Z. R., Phan-Everson, T., Geiss, G., Korukonda, M., Bhatt, R., Brown, C., Dunaway, D., Phan, J., Rosenbloom, A., Filanoski, B., Meredith, R., Chantranuvatana, K., Liang, Y., Brown, E., Birditt, B., Ong, G., Yi, H. S., Piazza, E., Devgan, V., … Beechem, J. (2022). Subcellular characterization of over 100 proteins in FFPE tumor biopsies with CosMx Spatial Molecular Imager. *Cancer Research*, *82*(12_Supplement), 3878–3878. https://doi.org/10.1158/1538-7445.AM2022-3878

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P. M., Santagata, S., & Sorger, P. K. (2018). *Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes*. https://doi.org/10.7554/eLife.31657.001

Lotfollahi, M., Dony, L., Agarwala, H., & Theis, F. (2020). *Out-of-distribution prediction with disentangled representations for single-cell RNA sequencing data*. https://doi.org/10.1101/2021.09.01.458535

McCaffrey, E. F., Donato, M., Keren, L., Chen, Z., Delmastro, A., Fitzpatrick, M. B., Gupta, S., Greenwald, N. F., Baranski, A., Graf, W., Kumar, R., Bosse, M., Fullaway, C. C., Ramdial, P. K., Forgó, E., Jojic, V., Van Valen, D., Mehra, S., Khader, S. A., … Angelo, M. (2022). The immunoregulatory landscape of human tuberculosis granulomas. *Nature Immunology*, *23*(2), 318–329. https://doi.org/10.1038/s41590-021-01121-x

McCoy, J. T., Kroon, S., & Auret, L. (2018). Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit. *IFAC-PapersOnLine*, *51*(21), 141–146. https://doi.org/10.1016/j.ifacol.2018.09.406

Molino, P., Dudin, Y., & Miryala, S. S. (2019). *Ludwig: a type-based declarative deep learning toolbox*. http://arxiv.org/abs/1909.07930

Neumann, E. K., Patterson, N. H., Rivera, E. S., Allen, J. L., Brewer, M., deCaestecker, M. P., Caprioli, R. M., Fogo, A. B., & Spraggins, J. M. (2022). Highly multiplexed immunofluorescence of the human kidney using co-detection by indexing. *Kidney International*, *101*(1), 137–143. https://doi.org/10.1016/J.KINT.2021.08.033

Patruno, L., Maspero, D., Craighero, F., Angaroni, F., Antoniotti, M., & Graudenzi, A. (2021). A review of computational strategies for denoising and imputation of single-cell transcriptomic data. In *Briefings in Bioinformatics* (Vol. 22, Issue 4). Oxford University Press. https://doi.org/10.1093/bib/bbaa222

Qiu, Y. L., Zheng, H., & Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *GigaScience*, *9*(8). https://doi.org/10.1093/gigascience/giaa082

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., Ashenberg, O., Cerami, E., Coffey, R. J., Demir, E., Ding, L., Esplin, E. D., Ford, J. M., Goecks, J., Ghosh, S., Gray, J. W., Guinney, J., Hanlon, S. E., Hughes, S. K., … Zhuang, X. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. In *Cell* (Vol. 181, Issue 2, pp. 236–249). Cell Press. https://doi.org/10.1016/j.cell.2020.03.053

Schapiro, D., Sokolov, A., Yapp, C., Chen, Y.-A., Muhlich, J. L., Hess, J., Creason, A. L., Nirmal, A. J., Baker, G. J., Nariya, M. K., Lin, J.-R., Maliga, Z., Jacobson, C. A., Hodgman, M. W., Ruokonen, J., Farhi, S. L., Abbondanza, D., McKinley, E. T., Persson, D., … Sorger, P. K. (2021). MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nature Methods*. https://doi.org/10.1038/s41592-021-01308-y

Sepe, J. J., Gardner, R. T., Blake, M. R., Brooks, D. M., Staffenson, M. A., Betts, C. B., Sivagnanam, S., Larson, W., Kumar, S., Bayles, R. G., Jin, H., Cohen, M. S., Coussens, L. M., & Habecker, B. A. (2022). *Therapeutics That Promote Sympathetic Reinnervation Modulate the Inflammatory Response After Myocardial Infarction VISUAL ABSTRACT*. https://doi.org/10.1016/j.jacbts.2022.04.009

Sheng, W., Zhang, C., Mohiuddin, T. M., Al-Rawe, M., Zeppernick, F., Falcone, F. H., Meinhold-Heerlein, I., & Hussain, A. F. (2023). Multiplex Immunofluorescence: A Powerful Tool in Cancer

Immunotherapy. *International Journal of Molecular Sciences*, *24*(4), 3086. https://doi.org/10.3390/ijms24043086

Sims, Z., & Chang, Y. H. (2023). *A Masked Image Modeling Approach to Cyclic Immunofluorescence (CyCIF) Panel Reduction and Marker Imputation*. https://doi.org/10.1101/2023.05.10.540265

Steele, N. G., Carpenter, E. S., Kemp, S. B., Sirihorachai, V. R., The, S., Delrosario, L., Lazarus, J., Amir, E. ad D., Gunchick, V., Espinoza, C., Bell, S., Harris, L., Lima, F., Irizarry-Negron, V., Paglia, D., Macchia, J., Chu, A. K. Y., Schofield, H., Wamsteker, E. J., … Pasca di Magliano, M. (2020). Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. *Nature Cancer*, *1*(11), 1097–1112. https://doi.org/10.1038/s43018-020-00121-4

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198

Talwar, D., Mongia, A., Sengupta, D., & Majumdar, A. (2018). AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-34688-x

Tan, W. C. C., Nerurkar, S. N., Cai, H. Y., Ng, H. H. M., Wu, D., Wee, Y. T. F., Lim, J. C. T., Yeong, J., & Lim, T. K. H. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. In *Cancer Communications* (Vol. 40, Issue 4, pp. 135–153). John Wiley and Sons Inc. https://doi.org/10.1002/cac2.12023

Ternes, L., Dane, M., Labrie, M., Mills, G., Gray, J., Heiser, L., & Chang, Y. H. (2021). *ME-VAE: Multi-Encoder Variational AutoEncoder for Controlling Multiple Transformational Features in Single Cell Image Analysis*. https://doi.org/10.1101/2021.04.22.441005

Tran, D., Tran, B., Nguyen, H., & Nguyen, T. (2022). A novel method for single-cell data imputation using subspace regression. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-06500-4

Tsujikawa, T., Kumar, S., Borkar, R. N., Azimi, V., Thibault, G., Chang, Y. H., Balter, A., Kawashima, R., Choe, G., Sauer, D., El Rassi, E., Clayburgh, D. R., Kulesz-Martin, M. F., Lutz, E. R., Zheng, L., Jaffee, E. M., Leyshock, P., Margolin, A. A., Mori, M., … Coussens, L. M. (2017). Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Reports*, *19*(1), 203–217. https://doi.org/10.1016/j.celrep.2017.03.037

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, *174*(3), 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061

Vinga, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. In *Briefings in Bioinformatics* (Vol. 22, Issue 1, pp. 77–87). Oxford University Press. https://doi.org/10.1093/bib/bbaa122

Werlein, C., Ackermann, M., Stark, H., Shah, H. R., Tzankov, A., Haslbauer, J. D., von Stillfried, S., Bülow, R. D., El-Armouche, A., Kuenzel, S., Robertus, J. L., Reichardt, M., Haverich, A., Höfer, A., Neubert, L., Plucinski, E., Braubach, P., Verleden, S., Salditt, T., … Jonigk, D. (2023). Inflammation and vascular remodeling in COVID-19 hearts. *Angiogenesis*, *26*(2), 233–248. https://doi.org/10.1007/s10456-022-09860-7

Xu, C., Cai, L., & Gao, J. (2021). An efficient scRNA-seq dropout imputation method using graph attention network. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04493-x

Zhai, J., Zhang, S., Chen, J., & He, Q. (2018). Autoencoder and Its Various Variants. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 415–419. https://doi.org/10.1109/SMC.2018.00080

Methods

Experimental Setup

The BOND RX Automated IHC/ISH Stainer was used to bake FFPE slides at 60°C for 30 minutes, to dewax the sections using the Bond Dewax solution at 72°C, and for antigen retrieval using Epitope Retrieval 1 (Leica™) solution at 100°C for 20 minutes. Slides underwent multiple cycles of antibody incubation, imaging, and fluorophore inactivation. All antibodies were incubated overnight at 4°C in the dark. Slides were stained with Hoechst 33342 for 10 minutes at room temperature in the dark following antibody incubation in every cycle. Coverslips were wet-mounted using 200 µL of 10% Glycerol in PBS prior to imaging. Images were acquired using a 20x objective (0.75 NA) on a CyteFinder slide scanning fluorescence microscope (RareCyte Inc. Seattle WA). Fluorophores were inactivated using a 4.5% H2O2, 24 mM NaOH/PBS solution and an LED light source for 1 hour.
The detailed protocol is available in protocols.io (dx.doi.org/10.17504/protocols.io.bjiukkew).

Data Preparation

The original source files include X and Y spatial coordinates and bio-morphological information (orientation, area, extent, etc.) for each cell. These features are removed for the initial imputation experiments, which solely rely on protein information.

To prepare the available data for the machine learning models and deep learning networks, we used Min-Max Scaling to scale features to be in the [0,1] range.

Statistical Validity:
For robust statistical validity, we conducted more than 30 experiments (n > 30) for each protein imputation and each model.

Elastic Net

To setup an experiment using the Elastic Net, the scikit-learn library was used and within this library the ElasticNetCV, which automatically performs cross validation of error values. To support our results with statistical significance as well as a high enough numbers of trials, we performed multiple experiments n > 30. Each run was performed using a different random seed.

Light GBM

To setup a training and evaluation pipeline for our Light GBM (Ke et al., 2017) model, we used the Ludwig (Molino et al., 2019) platform, which enables “End-to-end machine learning pipelines” in a low code environment. To setup a Ludwig network only a config file specifying the features, in this work the proteins, and the target to be imputed is required. We automated this process by using a combination of shell and make scripts. Each run was assigned a different random seed to ensure reproducibility.

Auto Encoder

The Auto Encoder imputation was setup to make use of an iterative approach. As a first step, the preprocessed source data was loaded. To impute a specific protein, the protein values were replaced with a mean value calculated by using all available values for the protein in question. After the dataset was prepared and the protein in question replaced, the data was used as input for the auto encoder. A full encode and decode process was performed (**Fig. 2**), and the output was stored as an

intermediate result. From this intermediate result, the imputed protein values were taken and used as a replacement of the mean protein values created in the preparation step. After this, another round of imputation was performed. This process was preformed 10 times, which resulted in a 10-step iterative imputation process. Reported MAE and RMSE values are calculated by using the last 5 iterative decoding's, calculating the mean of the decoding's for the protein and then calculate the MAE and RMSE.

For each model and network, we performed a multitude of experiments, with a minimum of at least 30 experiments. To create reproducible but different results, each experiment used a unique random seed.

Acknowledgements

This research was supported by the National Cancer Institute (NCI) of the National Institutes of Health grants U24CA231877 and U2CCA233280 and by funding from the Prospect Creek Foundation to the OHSU SMMART (Serial Measurement of Molecular and Architectural Responses to Therapy) Program.

Contributions

R.K and J.G both conceived the study. C.W. and A.C. processed the primary image data to produce the single-cell datasets analyzed in this work. R.K. developed the methods, wrote the code, and performed the analysis. C.W., A.C., and J.G. provided feedback and suggestions on methods, code, and analyses. K.K. designed illustrations used in the figures and provided feedback on the overall figure design. All authors read and approved the final paper.

Ethic Declarations

This research strictly adheres to ethical standards in data collection, processing, and usage. All datasets used for training and testing our machine learning models are publicly available.

Competing interests

The authors declare no competing interests.