

# Automated multi-scale computational pathotyping (AMSCP) of inflamed synovial tissue

---

Received: 26 August 2023

---

Accepted: 26 July 2024
















---

Published online: 29 August 2024

---

 Check for updates

---

Richard D. Bell <sup>1,2,28</sup> ✉, Matthew Brendel <sup>3,28</sup>, Maxwell A. Konnaris<sup>4,5</sup>, Justin Xiang<sup>6</sup>, Miguel Otero <sup>2,5</sup>, Mark A. Fontana <sup>1,3</sup>, Zilong Bai <sup>3</sup>, Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium\*, Daria M. Krenitsky<sup>7</sup>, Nida Meednu <sup>7</sup>, Javier Rangel-Moreno <sup>7</sup>, Dagmar Scheel-Toellner<sup>8</sup>, Hayley Carr<sup>8</sup>, Saba Nayar <sup>8</sup>, Jack McMurray<sup>8</sup>, Edward DiCarlo<sup>9</sup>, Jennifer H. Anolik <sup>7,10</sup>, Laura T. Donlin <sup>1</sup>, Dana E. Orange<sup>1,11</sup>, H. Mark Kenney <sup>10</sup>, Edward M. Schwarz <sup>10</sup>, Andrew Filer <sup>8,29</sup>, Lionel B. Ivashkiv <sup>1,2,29</sup> & Fei Wang <sup>3,29</sup>

Rheumatoid arthritis (RA) is a complex immune-mediated inflammatory disorder in which patients suffer from inflammatory-erosive arthritis. Recent advances on histopathology heterogeneity of RA synovial tissue revealed three distinct phenotypes based on cellular composition (pauci-immune, diffuse and lymphoid), suggesting that distinct etiologies warrant specific targeted therapy which motivates a need for cost effective phenotyping tools in pre-clinical and clinical settings. To this end, we developed an automated multi-scale computational pathotyping (AMSCP) pipeline for both human and mouse synovial tissue with two distinct components that can be leveraged together or independently: (1) segmentation of different tissue types to characterize tissue-level changes, and (2) cell type classification within each tissue compartment that assesses change across disease states. Here, we demonstrate the efficacy, efficiency, and robustness of the AMSCP pipeline as well as the ability to discover novel phenotypes. Taken together, we find AMSCP to be a valuable cost-effective method for both pre-clinical and clinical research.

Disease pathotyping with histopathology, the discovery of disease subtypes using target organ histology, is a critical step in understanding etiology and response to therapy in heterogeneous diseases, like rheumatoid arthritis (RA). Our understanding of RA, which is a chronic, inflammatory joint disease, has greatly benefited from histopathology subtyping because the disease has distinct and disparate etiologies with largely stable pathotypes<sup>1,2</sup> that show differential response to therapy<sup>3-7</sup>. However, the process of pathotyping a patient can be resource intensive involving both basic and immune-

stains requiring a high level of expertise by pathologists to interpret tissue and cellular histologic features, and prone to inter- and intra-observer variation<sup>8,9</sup>. More cost effective and efficient procedures need to be developed in order to incorporate these types of data into a precision medicine decision making process.

Recent work describing RA pathotypes uncover three distinct synovial pathotypes (1) cellular dense, lymphocyte rich (lymphoid), (2) myeloid rich with few lymphocytes (diffuse/myeloid), and (3) fibroblast rich (pauci-immune); which are identifiable through distinct

cellular and tissue level changes within synovial joint biopsies<sup>4,10–13</sup>. These pathotypes also correlate with antibody positivity (i.e. anti-citrullinated peptide antibodies), with the lymphoid type enriched in antibody positive patients whereas both the diffuse/myeloid and pauci-immune types have equal contributions of both antibody positive and negative patients. This aligns with preclinical models that rely on antibody dependent (e.g. Collagen Induced Arthritis and Serum Induced Arthritis) and independent mechanisms (e.g. humanized TNF Transgenic and Zymosan Induced Arthritis) to study disease and are phenotypically similar to these pathotypes<sup>14</sup>. Thus, tools that allow us to study both murine and human pathology would improve our overall understanding of this heterogeneous disease.

Computational tools to study histological changes have been shown to augment pathologist workflows and allow for the identification of disease specific patterns<sup>15</sup>. In particular, machine learning, and specifically deep learning, is a data-driven framework that has recently had success in the automated analysis of musculoskeletal imaging data<sup>16</sup>. Additionally, computational tools can automate and provide a more holistic analysis of a wide variety of histopathologic tissue and cell-level changes to enable a more detailed understanding of disease subtypes. However, an automated and comprehensive tool to study both tissue and cell type specific changes in arthritis that can quantify therapeutic or clinically meaningful differences has not yet been developed<sup>16</sup>.

In this work, we developed an automated multi-scale computational pathotyping (AMSCP) model to analyze tissue and cell-level changes during the progression of inflammatory arthritis. This model can pathotype both mouse and human inflammatory arthritis in therapeutic intervention studies and clinical meaningful scenarios. We leveraged innovative transfer- and active- learning techniques to improve model performance and workload efficiency. Our modeling framework consists of two distinct components that can be utilized together or independently, (1) a method to segment different tissue types to characterize tissue-level changes and (2) a method to classify cell types within each tissue compartment to study how these change across disease states. We utilized two mouse models of inflammatory arthritis to train and validate our computational models with subsequent implementation on additional datasets to measure therapeutic efficacy, known biologic differences and discover novel pathologic changes. Then, we utilized a human synovial biopsy data set from the Accelerating Medicines Partnership Rheumatoid Arthritis (AMP-RA) research consortium to demonstrate our model's utility in a clinical setting by classifying lymphoid pathotypes from diffuse/myeloid pathotype and identifying cell types associated with the pauci-immune pathotype while preserving spatial cell-level data.

## Results

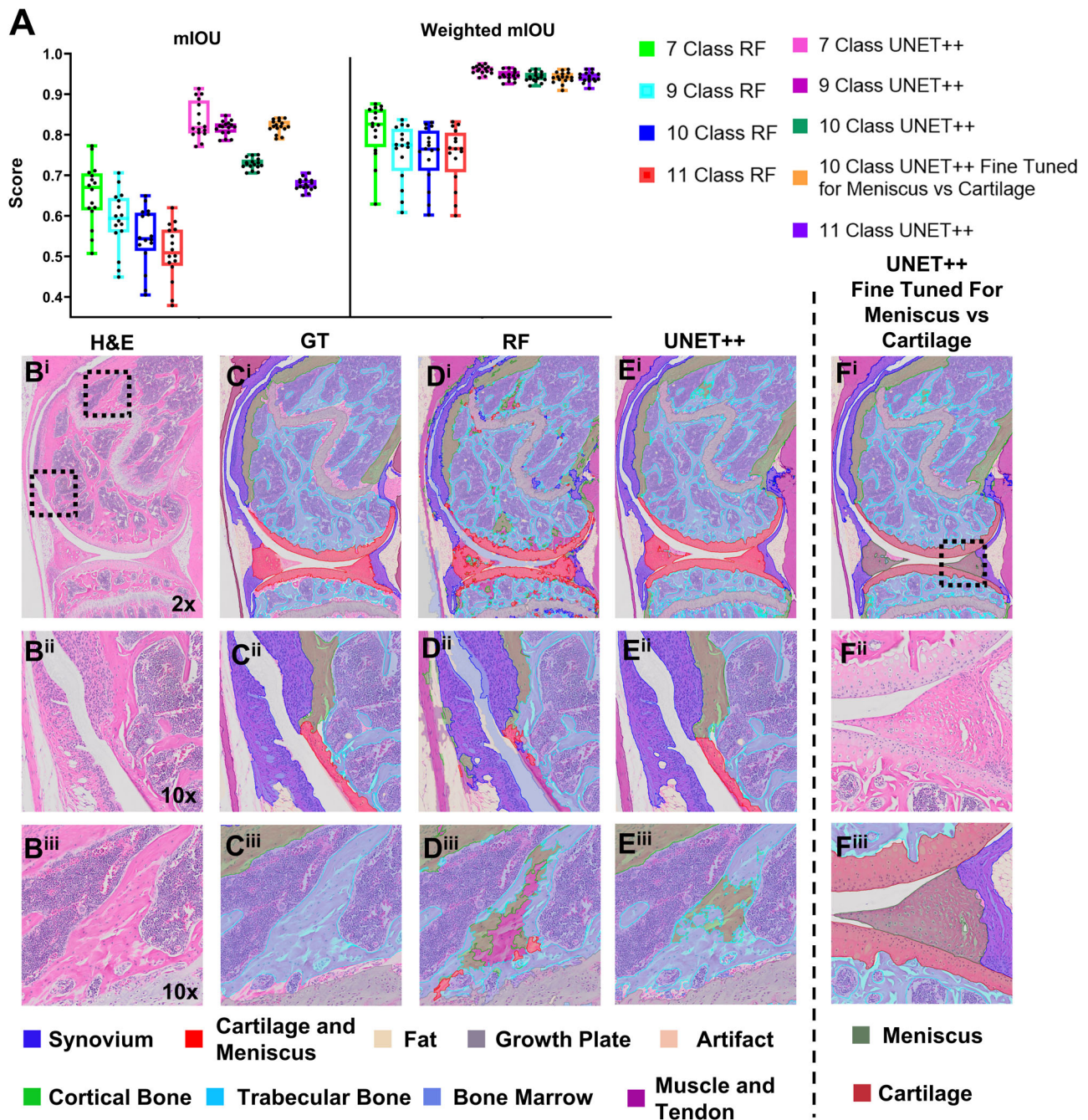
### Deep learning segmentation can identify major tissue types within mouse knee histology and measure therapeutic response

Several model training choices, including patch overlap, training strategy, and use of different amounts of augmentation during training, were empirically derived from initial experiments to inform the final training of the deep learning segmentation model (UNET++)<sup>17,18</sup>. First, we tested if 0%, 50% or 66% patch overlap was more performative and determined that 66% overlap performed the best (0% Overlap fwIOU:  $0.72 \pm 0.04$ ; 50% Overlap fwIOU:  $0.93 \pm 0.02$ ; 66% Overlap fwIOU:  $0.95 \pm 0.01$ ; Supplemental Fig. 4A). Qualitatively, there were less tiling edge artifacts in the 66% overlap vs 50% overlap results (50%—Arrows Supplemental Fig. 4B vs Fig. 1). Second, a mixed training strategy was shown to overcome the large staining batch effect (Supplemental Fig. 5A) commonly seen in histology datasets with all levels of augmentation equally performative (Supplemental Fig. 5B). However, if the data is restricted to a single batch and data augmentation is introduced, model performance

for the single batch training strategy becomes comparable to model performance using the mixed strategy (High Augmentation UNET++:  $0.81 \pm 0.02$  vs  $0.80 \pm 0.06$  mIOU for mixed and single batch respectively), demonstrating the need for image augmentation during training to generalize across batches with this level of heterogeneity in staining and other imaging variations (Supplemental Fig. 5B and C). Thus, we chose to employ a mixed training strategy, with 66% patch overlap and with high augmentation to optimize model performance and generalizability.

Once the training strategy was established, model performance was benchmarked across segmentation tasks at multiple tissue granularities and compared with a standard RF model built-in to QuPath (Fig. 1). As expected, as the number of different tissue types increases, model performance decreases for both the RF and UNET++ model with the DL model outperforming the RF at all levels. Interestingly, the magnitude of decrease is smaller for the UNET++ model compared to the RF model suggesting that it is more robust to increased complexity. When testing the UNET++ model, using the ten-class granularity, model performance drops from  $0.88 \pm 0.06$  mIOU for the cartilage and meniscus class to  $0.83 \pm 0.05$  mIOU for the cartilage class and  $0.0 \pm 0.0$  mIOU for the meniscus class indicating a complete loss of meniscus identification (Supplemental Fig. 6). Because defining the amount of cartilage and meniscus is a very important pathologic readout in inflammatory joint diseases (e.g. pannus invasion at end stage arthritis), we developed a finely tuned two-class model and placed it sequentially after the 9-class model. Predictions from both were incorporated during majority voting process to create a composite 10 class model. Performance jumps for this fine-tuned model from  $0.72 \pm 0.01$  mIOU to  $0.82 \pm 0.02$  mIOU (Fig. 1A, B). Specifically, performance for the cartilage increased from  $0.83 \pm 0.05$  to  $0.90 \pm 0.04$  and for the meniscus from  $0.00 \pm 0.00$  to  $0.92 \pm 0.05$ , a dramatic improvement (Supplemental Fig. 6C). We additionally observe that the worst performing classes are the artifact class and bone marrow fat class, the two most infrequent classes, suggesting that fine tuning may work to improve performance for one or both (Supplemental Fig. 6C, D). Thus, for all subsequent work we either used the 9-class model (termed Original model) for ease of computation or the Fine-Tuned 10 class model for meniscus segmentation because these provided the best predictive performance while encompassing the most amount of tissues.

After Test set validation, we then sought to externally validate the Fine-Tuned 10 class model (i.e., completely independent of the training, testing and validation process above) in two additional ways. (1) We first validated our Fine-Tuned 10 class model directly on previously published data by comparing hand drawn histomorphometry outlines of the synovial tissue previously published in Bell et al.<sup>19</sup> ( $n = 9$ ) to the synovial segmentation predictions that were in the Test set from the TNF-Tg cohort (Batch A). There was a significant positive correlation between the DL segmentation area and the hand drawn area ( $r^2 = 0.96$ , Fig. 2A) which demonstrates the accuracy of our method with hand drawn histomorphometry. However, linear regression and RMSE analysis suggests it is not a perfect fit describing an over segmentation at low areas (healthy samples) and under segmentation at high areas (severely diseased samples). (2) We then validated our model in a real-world setting by collecting 171 slides (64 knees, 2 or 3 histologic levels per knee) from 9.5 month-old male TNF-Tg mice either treated with anti-TNF therapy or placebo for 6 weeks and from 6 month-old TNF-Tg and WT (treatment naïve) mice used as controls<sup>20,21</sup> (experimental design schematic in Fig. 2B). Male TNF-Tg mice display a robust inflammatory arthritis with synovial hyperplasia and pannus invasion of the distal femur and femoral articular cartilage (Fig. 2C, WT vs TNF-Tg,  $p < 0.0001$ ; Fig. 2D, black arrow 2nd panel). Anti-TNF therapy is known to reduce synovitis (Fig. 2D, blue outlined tissue and \* right panel) yet does not alter trabecular bone loss (Fig. 2D, dark teal outlined tissue, red arrows right panels) in mice with

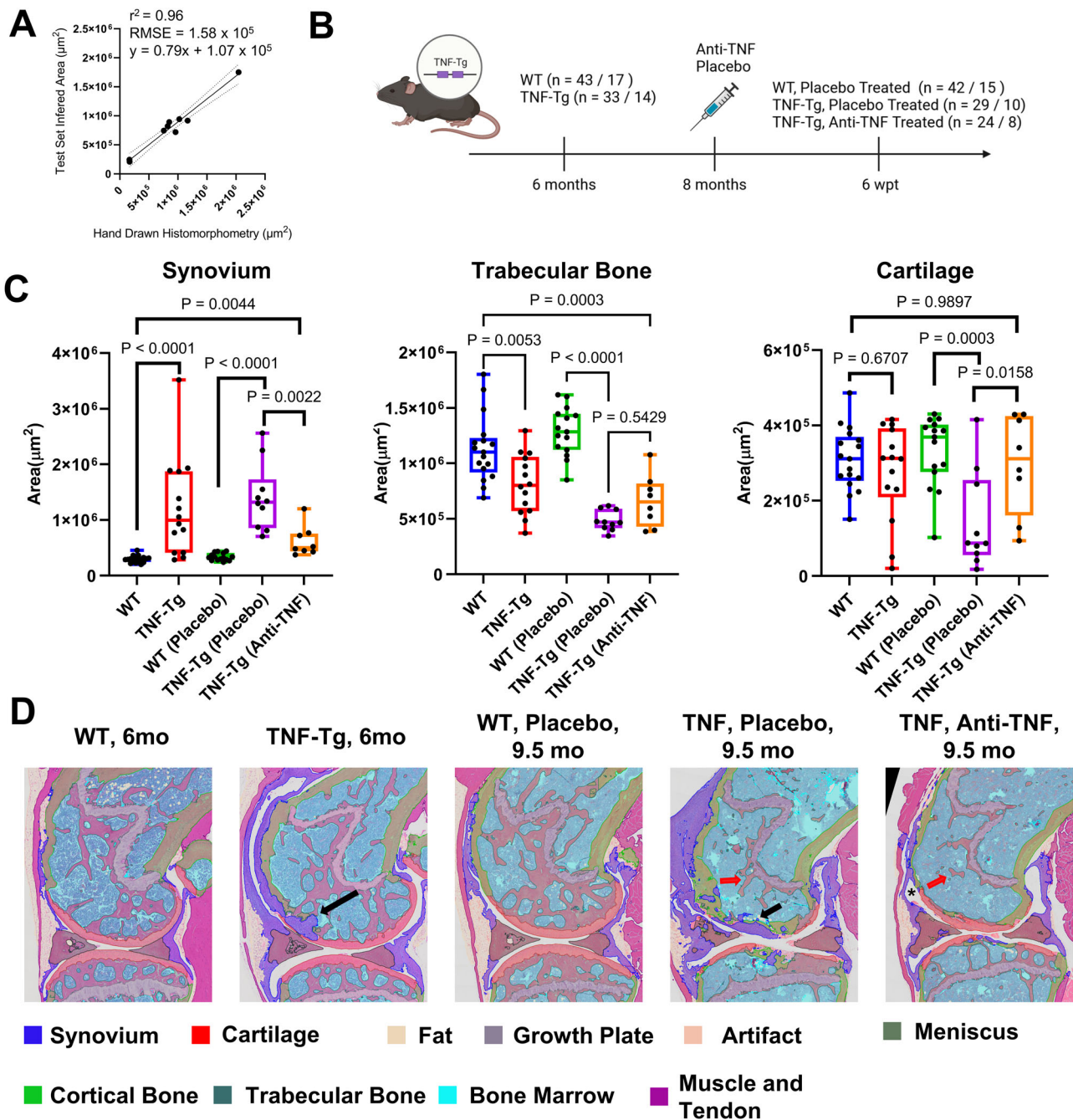


**Fig. 1 | A fine tuned 10 class model can segment relevant tissue in inflammatory arthritis.** **A** Mean Intersection over union (mIOU) and class frequency weighted mIOU statistics from the held-out test set for the RF and DL segmentation models at 4 different tissue granularities. Box and Whisker plots are constructed by showing the Min, 25th percentile, Median, 75th percentile and Max, each dot is one slide,  $n = 16$ . **B–F** Representative images of H&E (**B**) image, with Ground Truth (**C**), RF (**D**), 9 class UNET++ (**E**) and Fine Tuned UNET++ (**F**) tissue overlays from the

Test set. **B<sup>i</sup>–F<sup>i</sup>** 2x magnification of whole the joint. (**B<sup>ii</sup>–E<sup>ii</sup>**) 10x magnification of the anterior femoral condyle depicting synovial pannus encroachment reaching the articular cartilage. **F<sup>ii</sup>** 10x magnification of the posterior articular cartilage and meniscus. (**B<sup>iii</sup>–E<sup>iii</sup>**) 10x magnification of the trabecular bone and bone marrow proximal to the femoral growth plate depicting an area that was difficult to predict for all models (**F<sup>iii</sup>**) 10x magnification of the posterior articular cartilage and meniscus with the Fine Tuned UNET++ tissue prediction overlays.

established disease (>6 months old)<sup>22,23</sup>. Our DL segmentation appropriately modeled these well-established structural changes autonomously (Fig. 2C, D), and it uncovered that cartilage degradation is also moderately reduced when Anti-TNF therapy is provided at 8 months of age for 6 weeks, which is an expected but novel result. Interestingly, trabecular bone area is already decreased in 6 month-old TNF-Tg mice compared to WT counterparts while cartilage area is not, suggesting that trabecular bone loss occurs before cartilage loss. Additionally, multiple other tissue structures can be studied simultaneously

(Supplemental Fig. 7) showing the versatility of studying H&E segmentation models to assess tissue structural changes within the context of mouse disease models. Taken together, these analyses suggest that our model accurately detects relevant and meaningful biologic treatment effects with the potential to discover novel structural changes. It is important to note that some of these slides were stained with a variation of traditional H&E, H&E-Orange-G (see *Methods*) demonstrating that our training strategy choices produced a model that is robust even when introducing additional stain variations.



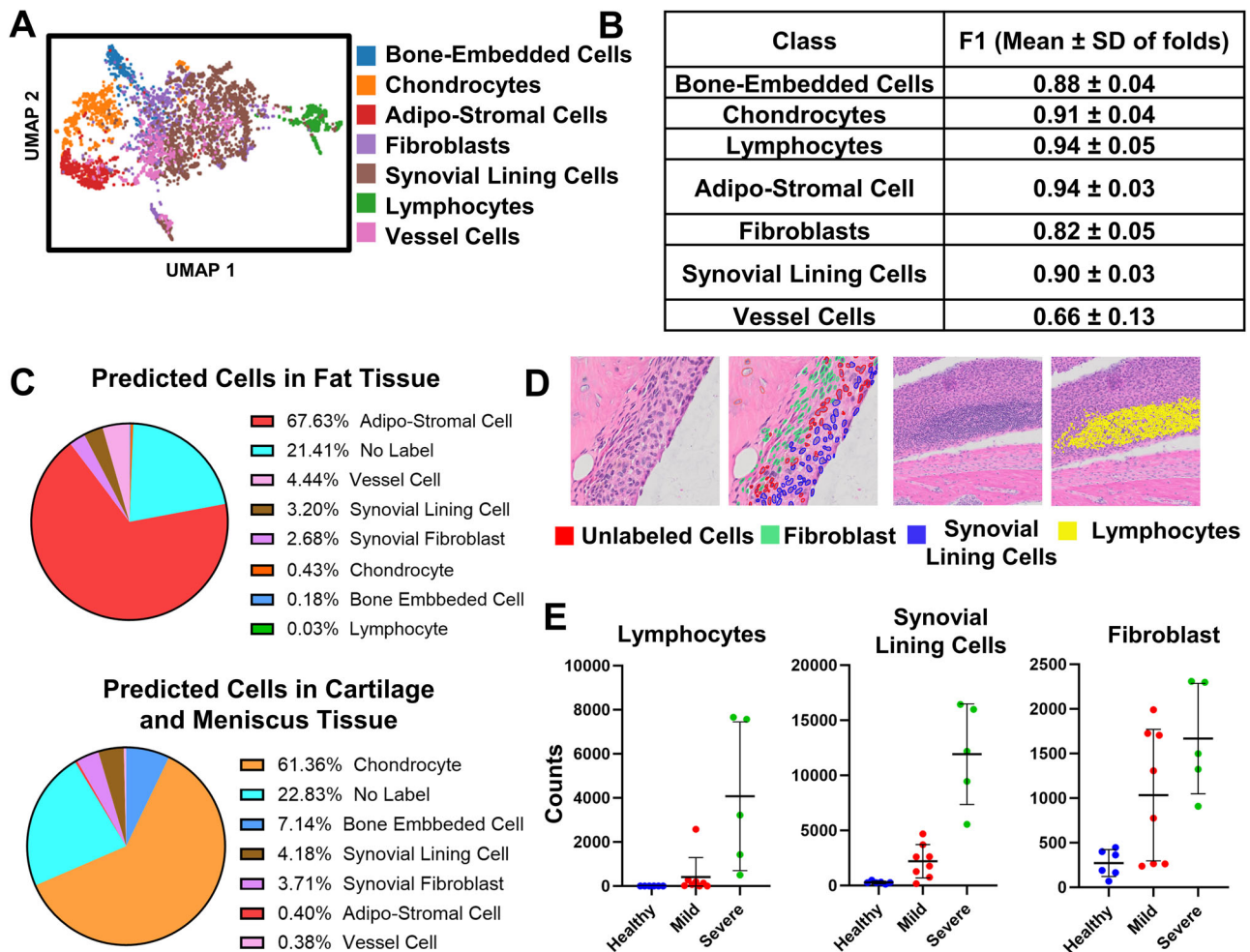
**Fig. 2 | The fine tuned UNET++ model measures treatment response in the TNF-Tg with Anti-TNF therapy.** **A** Inferred synovial area in the held-out test set compared to historical hand drawn synovial histomorphometry area ( $n = 9$ , both Tissue Inferred Area and Hand Drawn Histomorphometry Area are normally distributed, Pearson's Correlation). **B** Anti-TNF therapy Study design. Six-month-old WT and TNF-Tg mice were used as controls. Eight-month-old WT or TNF-Tg mice were treated with Anti-TNF therapy or Placebo control for 6 weeks (weeks post treatment, wpt). Both left and right knees were collected and 2–3 histologic levels per knee were analyzed ( $n = \text{Slides} / \text{Knees}$ ). **C** Using the fine tuned UNET++ model we inferred tissue area of the Synovium (Left), Trabecular Bone (Middle) and Cartilage (Right) for 6mo WT ( $n = 17$ ) and TNF Controls ( $n = 14$ ) as well as Placebo (Irrelevant IgG) treated WT ( $n = 15$ ), Placebo treated TNF-Tg ( $n = 10$ ) and Anti-TNF treated TNF-Tg ( $n = 8$ ). Each dot represents one knee (average of 2–3 histologic levels), Box and

Whisker plots are construct by showing the Min, 25th percentile, Median, 75th percentile and Max. Left Panel: TNF-Tg group was not normally distributed, data was log transformed and a One-Way ANOVA with Tukey's Post-hoc Test was performed. Middle Panel: All data was normally distributed, and a One-Way ANOVA with Tukey's Post-hoc Test was performed. Right Panel: WT (Placebo) and TNF-Tg (Placebo) groups are not normally distributed, data was log transformed and a One-Way ANOVA with Tukey's Post-hoc Test was performed. **D** Representative images (2x magnification) of 6mo WT and TNF Controls as well as 9.5 mo Placebo treated WT, Placebo treated TNF-Tg and Anti-TNF treated TNF-Tg with predicted tissue overlay. Note: black arrows denote pannus invasion of the femoral articular cartilage, the red arrows denote trabecular bone loss, and \* denotes reduction in synovial area.

**Tissue-Specific and Arthritis Effector Cell Types can be Identified with ML in TNF-Tg Mice**

We annotated a total of 4,712 cells across three stages of murine inflammatory arthritis in the TNF-Tg mice ( $n = 6$  healthy,  $n = 8$  mild

disease, and  $n = 5$  severe, Supplemental Fig. 8) to build a cell type classification model that could recognize cells from various disease stages. Cell nuclei were first segmented with HoverNet<sup>24</sup>, nuclei boundaries passed to ImageJ where stain deconvoluted color features



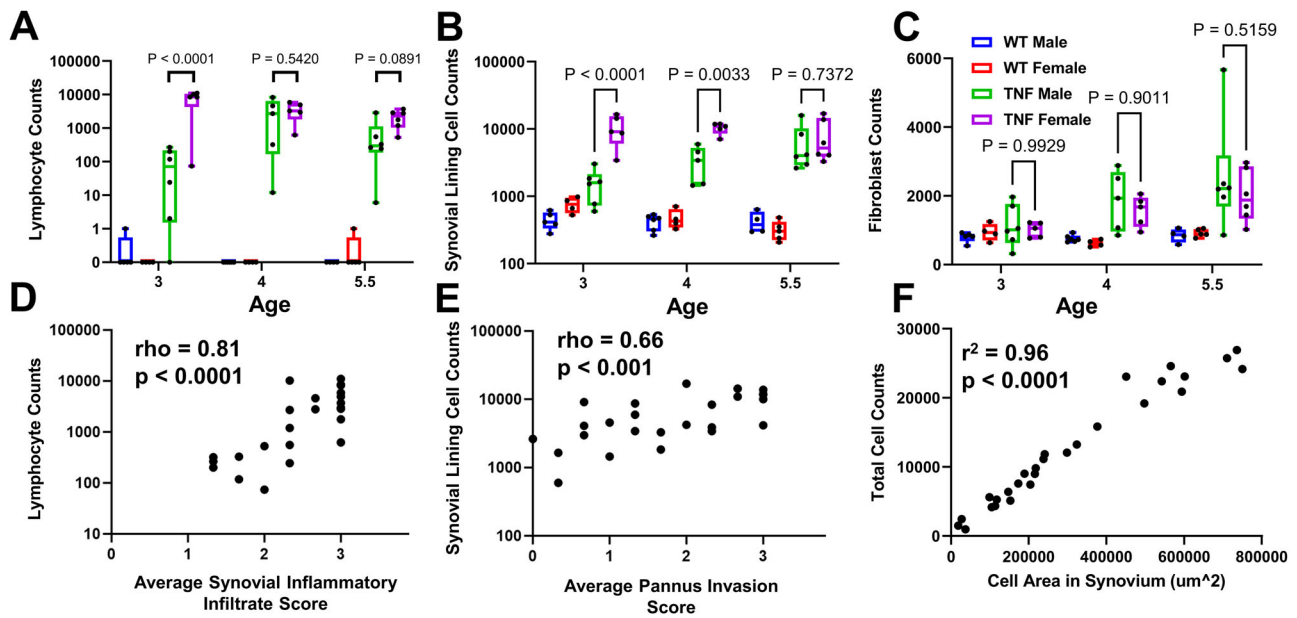
**Fig. 3 | Cell type classification model successfully identifies important cell types in inflammatory arthritis.** **A** Uniform Manifold Approximation and Projection (UMAP) plot after principal component analysis dimensional reduction on 856 cell features of the annotated cells (colored by cell type,  $n = 4,712$ ). **B** A gradient boosted decision tree was trained using a parameter grid search with a nested, stratified, 5-fold cross-validation training strategy. The F1 scores (M  $\pm$  SD) of the five folds for each cell class are presented with the overall weighted F1 of 0.88  $\pm$  0.03 (M  $\pm$  SD). **C–E** On the remaining (not-annotated) cells (~300,000) on the 9 training

slides, the cell class was predicted using the most performant model. Tissue class was also predicted using the Original segmentation model. **C** Predicted cell class is plotted as a percent of total cells within the Fat tissue (Top) and Cartilage and Meniscus tissue (Bottom). **D** Representative images of Synovial tissue cell class predictions within an inflamed synovium. **E** Cell counts from the synovial tissue on the Healthy, Mild disease and Severe disease training slides of lymphocytes (Left), Synovial Lining Cells (Middle) and Fibroblast (Right). Each dot represents one slide, M  $\pm$  SD.

of both the nuclei and cytoplasm as well as nuclei shape parameters were calculated. We next passed these data to our custom feature extraction pipeline leveraging our novel insight that cell type predictive modeling dramatically improves with neighborhood features (Supplemental Table 3) inspired by adjacent work tackling a different classification task<sup>25</sup>. These neighborhood features include standard statistical measure of neighboring cells within a radius to the parent cell like hematoxylin intensity mean or kurtosis, whether the parent cell was located in a dense region of other cells, and the shape (convex hull) of the cells within a radius to the parent cell (Supplemental Table 2). The fidelity of these methods was demonstrated in 2D UMAP space, as most cell types are clearly distinguished (Fig. 3A). We next built a gradient boosted decision tree (GBDT) classification model (Xgboost) using nested stratified 5-fold cross validation training and tested our models' predictions using three methods. (1) we calculated the average ( $\pm$  SD of folds) F1 of each cell class in the test sets (Fig. 3B). Each cell class demonstrated a good F1, between 0.66 for vessel cells and 0.94 for adipo-stromal cells with synovial associated cell classifications among the best (synovial lining cells = 0.90  $\pm$  0.03; synovial fibroblasts = 0.82  $\pm$  0.05; and synovial lymphocytes = 0.94  $\pm$  0.05). Our

next two validation strategies utilized tissue and disease context. (2) Tissue segmentation using the Original model was performed on these training slides and the remaining >300,000 cells cell-type classification was inferred. Tissues with known homogenous cell types, fat tissue and cartilage/meniscus, were investigated and cell types among these tissues were plotted. Within these adipose tissue and cartilage/meniscus, the most predicted cell type were adipo-stromal cells (67%) and chondrocytes (61%), respectively (Fig. 3C). (3) To assess predictions in the context of disease, we utilized the synovial tissue predictions only (as determined by the tissue segmentation model Fig. 1) and stratified by disease severity. As shown by Fig. 3D, E, synovial-specific increases in synovial fibroblasts, synovial lining cells and lymphocytes are seen with increasing disease severity. These results suggest that our cell type model can produce high quality predictions that are sensitive to disease stage.

Given the promising intra-test set performance and tissue- and disease-state specificity of the cell type modeling, we aimed to further validate our model with a larger data-set with more biologic variation. To do this, we utilize the previously described sexually dimorphic synovial pathology in TNF-Tg mice<sup>19</sup> and collected slides from



**Fig. 4 | Computational pathology modeling recapitulates the sexual dimorphism of TNF-Tg inflammatory arthritis.**

An independent set of slides from the training slides were used to validate the cell type prediction model (3 months-old: WT Male  $n = 5$ , WT Female  $n = 4$ , TNF-Tg Male  $n = 6$ , TNF-Tg Female  $n = 5$ ; 4 months-old: WT Male  $n = 6$ , WT Female  $n = 4$ , TNF-Tg Male  $n = 5$ , TNF-Tg Female  $n = 5$ ; 5.5 months-old: WT Male  $n = 4$ , WT Female  $n = 5$ , TNF-Tg Male  $n = 6$ , TNF-Tg Female  $n = 6$ ). Tissue segmentation was first performed to segment the synovium and then cell type predictions were performed within the synovium.

**A** Lymphocytes predictions counts. Each dot is one mouse, Box and Whisker plots are construct by showing the Min, 25th percentile, Median, 75th percentile and Max. Please note the log scale. Lymphocyte counts were found to be lognormal, a log transformation was performed on the data and a Two-Way ANOVA with Tukey's Post-hoc test was conducted. **B** Synovial Lining Cell prediction counts. Each dot is one mouse, Box and Whisker plots are construct by showing the Min, 25th percentile, Median, 75th percentile and Max. Please note the log scale. A Two-Way

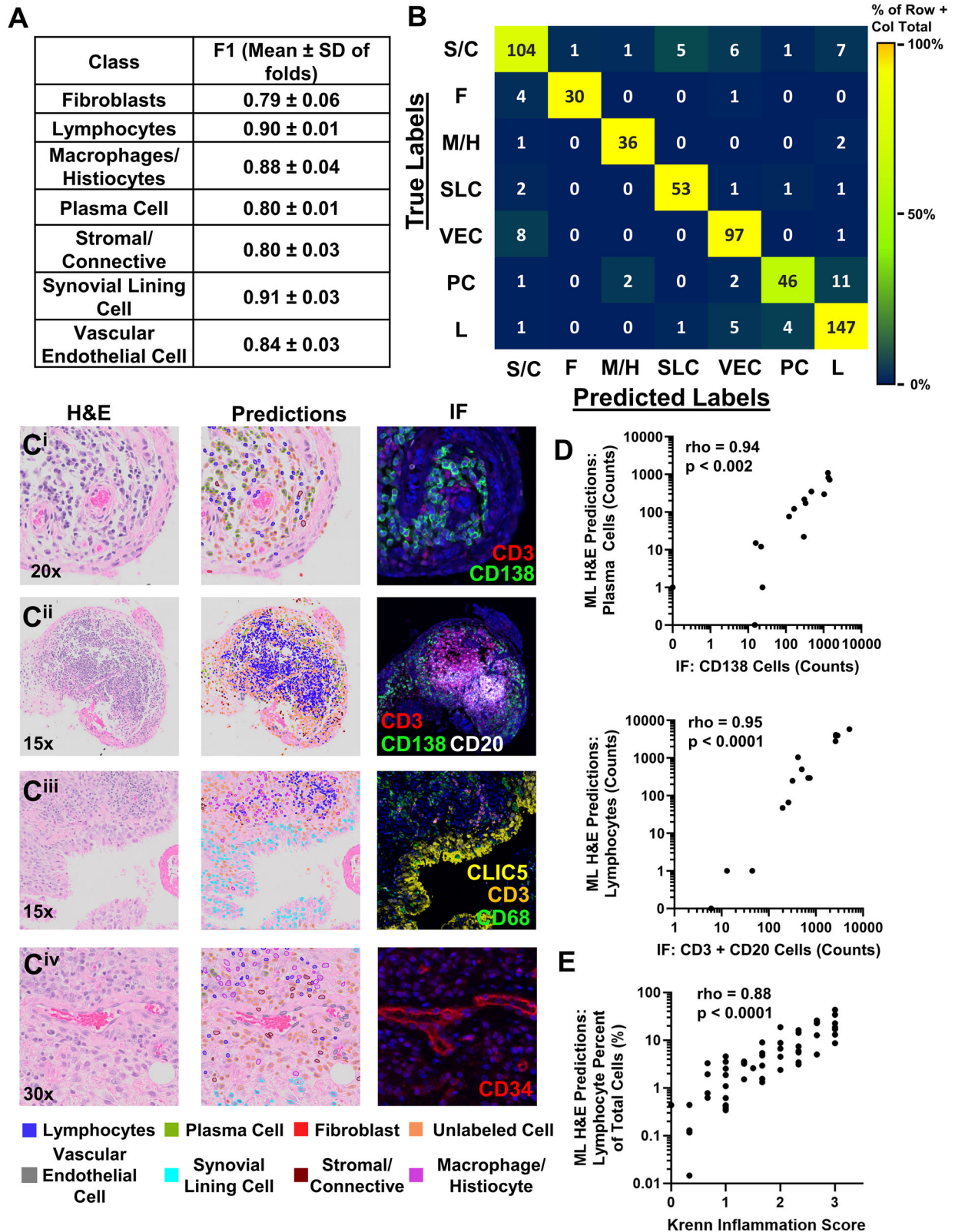
ANOVA with Tukey's Post-hoc test was conducted. Differences only shown between the female and male TNF-Tg mice. **C** Fibroblast predictions counts. Each dot is one mouse, Box and Whisker plots are construct by showing the Min, 25th percentile, Median, 75th percentile and Max. A Two-Way ANOVA with Tukey's Post-hoc test was conducted. Differences only shown between the female and male TNF-Tg mice. **D** Lymphocyte predictions compare to the synovial inflammatory score as previously quantified in Figure 2F of Bell et al.<sup>19</sup>. Lymphocyte counts were not normally distributed. Spearman's correlation, TNF-Tg mice only ( $n = 28$ ). Please note the log scale on the x-axis. **E** Synovial Lining cell counts compared to the pannus invasion score as previously quantified in Supplemental Figure 3I in Bell et al.<sup>19</sup>. Synovial Lining cell counts were not normally distributed. Spearman's correlation, TNF-Tg mice only ( $n = 28$ ). Please note the log scale on the x-axis. **F** Total cell counts in the synovium compared to the cell area as previously quantified in Figure 2D of Bell et al.<sup>19</sup>. Total cell counts and cell area in the synovium were normally distributed. Pearson's correlation, TNF-Tg mice only ( $n = 28$ ).

3–5.5-month-old WT and TNF-Tg-male and female mice. Confirming our previous observation using traditional histologic scoring, we found a significant increase of lymphocytes in female TNF-Tg synovium at 3 months of age with concomitant significant increase in synovial lining cells (Fig. 4A, B). Interestingly, sexual dimorphism was not observed when assessing synovial fibroblasts, which is a novel finding (Fig. 4C). Finally, we found excellent correlations with our computationally derived lymphocyte counts vs expert derived synovial inflammatory score ( $\rho = 0.81$ ,  $p < 0.0001$ ), synovial lining cell counts and pannus score ( $\rho = 0.66$ ,  $p < 0.001$ ), and total cell counts vs total cell area ( $r^2 = 0.96$ ,  $p < 0.0001$ ) (Fig. 4D–F). These data suggest that our mouse cell typing model is sensitive to both subtle and dramatic tissue changes, and importantly recapitulates expert scored data.

To reduce the annotation time, we explored various active learning approaches retrospectively in the murine cell type data set. Model performance using all the active learning strategies with 45% of the total training size was comparable to using the complete dataset with random sampling (0.8188 (0.8157–0.8219) vs 0.8213 (0.8184–0.8243),  $F1 \pm SD$ ). Additionally, with 45% of the data, the model performance using active learning was higher than using a randomly sampled set of examples (0.8188 (0.8157–0.8219) vs 0.8082(0.8053–0.8111)  $F1 \pm SD$ ) (Supplemental Fig. 10). Additionally, the mean 5-fold CV macro F1-score was most different at the 10%–25% of annotated data range, indicating that active learning can drastically improve model performance with fewer examples.

### Cell type modeling on human RA synovial biopsies predicts pathotypes and correlates with clinical outcomes

After validating the active learning strategy in murine tissue, we next applied this approach to generate cell type annotations on human synovial tissue sections, aiming to reduce the overall histopathological evaluation time for our pathologist. We collected a small subset of initial annotations, predicted cell types on new cells and calculated the entropy-based uncertainty, ranking the most uncertain cells for future annotation by the pathologist. After multiple rounds with active learning, a total of 2,341 cells were annotated (Supplemental Fig. 10). Using a GBDT (xgboost) with a nested stratified 5-fold cross-validation training strategy, we achieved good model performance ranging from 0.79–0.91 average F1-scores with the overall weighted F1 of  $0.85 \pm 0.01$  ( $M \pm SD$ , Fig. 5A). The confusion matrix from the best performing fold (Fig. 5B) demonstrates that undifferentiated stromal-connective cells are confused with vascular endothelial cells and synovial lining at the highest frequency. Also, plasma cells are confused with lymphocytes. These results suggest there is very little misclassification between high level cell type (stromal vs lymphoreticular cells) but some misclassification within more specific cell classes. To further validate our human RA cell typing model, we acquired adjacent immunostained sections (Lymphoid,  $n = 7$ ; Diffuse,  $n = 6$ ; Pauci-Immune,  $n = 2$ ) for lymphoid (CD3, CD20 and CD138) and stromal-immune (CD3, CD68, CLIC5 and CD34) markers to qualitatively and quantitatively assess for cell type validity. Our qualitative analysis revealed a remarkable spatial alignment of machine learning predicted plasma cells with CD138+



cells (Fig. 5C<sub>i</sub>, green outline middle panel vs green IF right panel) and lymphocytes with CD3+ or CD20+ cells (Fig. 5C<sub>ii</sub>, blue outline middle panel vs red/white IF right panel), synovial lining cells with CLIC5+ cells (Fig. 5C<sub>iii</sub>, light blue outline middle panel vs yellow IF right panel) and vascular endothelial cells with CD34+ cells (Fig. 5C<sub>iv</sub>, grey outline middle panel vs red IF right panel) in many of the sections. We also observed some alignment of macrophage-histiocytes predictions with

CD68+ cells (Fig. 5C<sub>iii</sub>, purple outline middle panel vs green IF right panel) but these observations were much less consistent suggesting that either the model fails to predict these cells outside of the training data or H&E defined macrophages-histiocytes are not well marked with CD68 in our data. We were unable to acquire immunostains to validate our fibroblast and stromal-connective cell classes. Lymphocyte, and in particular B-cell, infiltration into the synovial tissue is an important

**Fig. 5 | Cell type modeling correctly classifies synovial stromal and immune cells in RA synovial biopsies.** **A** A subset of cells from 13 RA synovial biopsies were annotated ( $n = 2,341$ ) using an active learning strategy. After nuclei detection and custom feature extraction from each cell, a gradient boosted decision tree was trained using a parameter grid search with a nested, stratified, 5-fold cross-validation training strategy. The F1 scores ( $M \pm SD$ ) of the five folds for each cell class are presented with the overall weighted F1 of  $0.85 \pm 0.01$  ( $M \pm SD$ ). **B** The confusion matrix from the most performant model demonstrates the typical misclassification in this dataset (data is cell counts). Stromal cells can be mistaken for other stromal cells (vascular endothelial cells, synovial lining cells, and fibroblast) and lymphocytes can be mistaken for plasma cells. F: Fibroblast, L: Lymphoid, M/H: Macrophage/Histocyte, PC: Plasma Cell, S/C: Stromal/Connective Cell, SLC: Synovial Lining Cell; VEC: Vascular Endothelial Cell. **C, D** The most performant model was used to predict the cell type of the remaining cells from all RA synovial biopsies

( $n = 60$ ). **C** Adjacent sections to the H&E-stained slides were stained with either CD3 (T-Cells), CD20 (B-Cells) and CD138 (Plasma Cells) or CLIC5 (Synovial Lining), CD3 (T-Cells), CD68 (Macrophages), and CD34 (Vascular Endothelial Cells) ( $n = 15$ ). Representative images of plasma cells (C<sup>i</sup>), lymphocytes (C<sup>ii</sup>), synovial lining cells (C<sup>iii</sup>) and vascular endothelial cells (C<sup>iv</sup>) with the original H&E in the left column, prediction overlays in the middle, and adjacent slide IF in the right column. Immunostains and magnification are denoted within the image. **D** Correlation of machine learning predictions with quantitative histomorphometry of IF+ cells from adjacent sections of  $n = 15$  RA synovial biopsy pieces. Top: CD138+ cells vs ML Predictions of Plasmids cells; Bottom: CD3+ and CD20+ cells vs ML Predictions of Lymphocytes; Spearman's Correlations ( $n = 15$ ). Please note the log scale. **E** Correlation of machine learning predictions of lymphocytes (as a percent of total cells) vs the Krenn Inflammation Score, Spearman's Correlations ( $n = 60$ ). Please note the log scale on the x-axis.

pathologic finding to discriminate RA pathotypes<sup>2,11,26</sup> and classify disease severity<sup>27</sup>. Therefore, to quantitatively validate our lymphocyte and plasma cell machine learning predictions vs fiducial protein cell markers on adjacent slides, we performed thresholding histomorphometry analysis to count the number of either DAPI+CD138+, DAPI+CD3+ or DAPI+CD20+ cells within regions of interest. There is an excellent correlation between our machine learning predictions of plasma cells vs DAPI+CD138+ ( $\rho = 0.94$ ,  $p < 0.002$ , Fig. 5D, top), and lymphocytes with vs DAPI+CD3+ or DAPI+CD20+ cells ( $\rho = 0.95$ ,  $p < 0.0001$ , Fig. 5D, bottom). Krenn inflammation score is an important, expert derived histopathology score driven mainly by the number of lymphocytes that are present in the whole synovial biopsy specimen. Thus, we calculated the percent of machine learning predicted lymphocytes to total cells and correlated with the Krenn inflammation score in all samples ( $n = 60$ ). This revealed a high degree of correlation between the predictions and pathologist scores ( $\rho = 0.88$ ,  $p < 0.0001$ , Fig. 5E). Taken together, this analysis demonstrates we have an excellent synovial cell type prediction model that faithfully predicts clinically relevant lymphoid cells.

To further validate model performance, cell type counts were enumerated and proportions of total cells were calculated for the entire synovial biopsy dataset and grouped by pathotype. Importantly, our cell type predictions were consistent with the previously described cellular distribution within each pathotype<sup>4,10,11,13</sup>. Specifically, synovial fibroblast enrichment is found in the pauci-immune pathotype and some diffuse cases, while lymphocytes and plasma cells are found primarily in the lymphoid pathotype (Fig. 6A). Efficient pathotype prediction is a clinically relevant task that we propose will reduce time and cost in RA clinical trials. Area under a receiver operator curve (AUROC) analysis demonstrates that percent plasma cells have a high predictive capability for discriminating between diffuse and lymphoid cases (AUROC =  $0.82 \pm 0.06$ ,  $p < 0.0001$ , Fig. 6B). The optimal simple threshold is 0.82% plasma cells of total can classify 19 out of 24 lymphoid cases and 25 out of 29 diffuse cases correctly. Representative images of pauci-immune, diffuse and lymphoid cases with their respective predictions demonstrate the cellular distributions of these pathotypes (Fig. 6C). This data shows that our model can be used in a clinically meaningful scenario and supports the development of such tools in clinical trials.

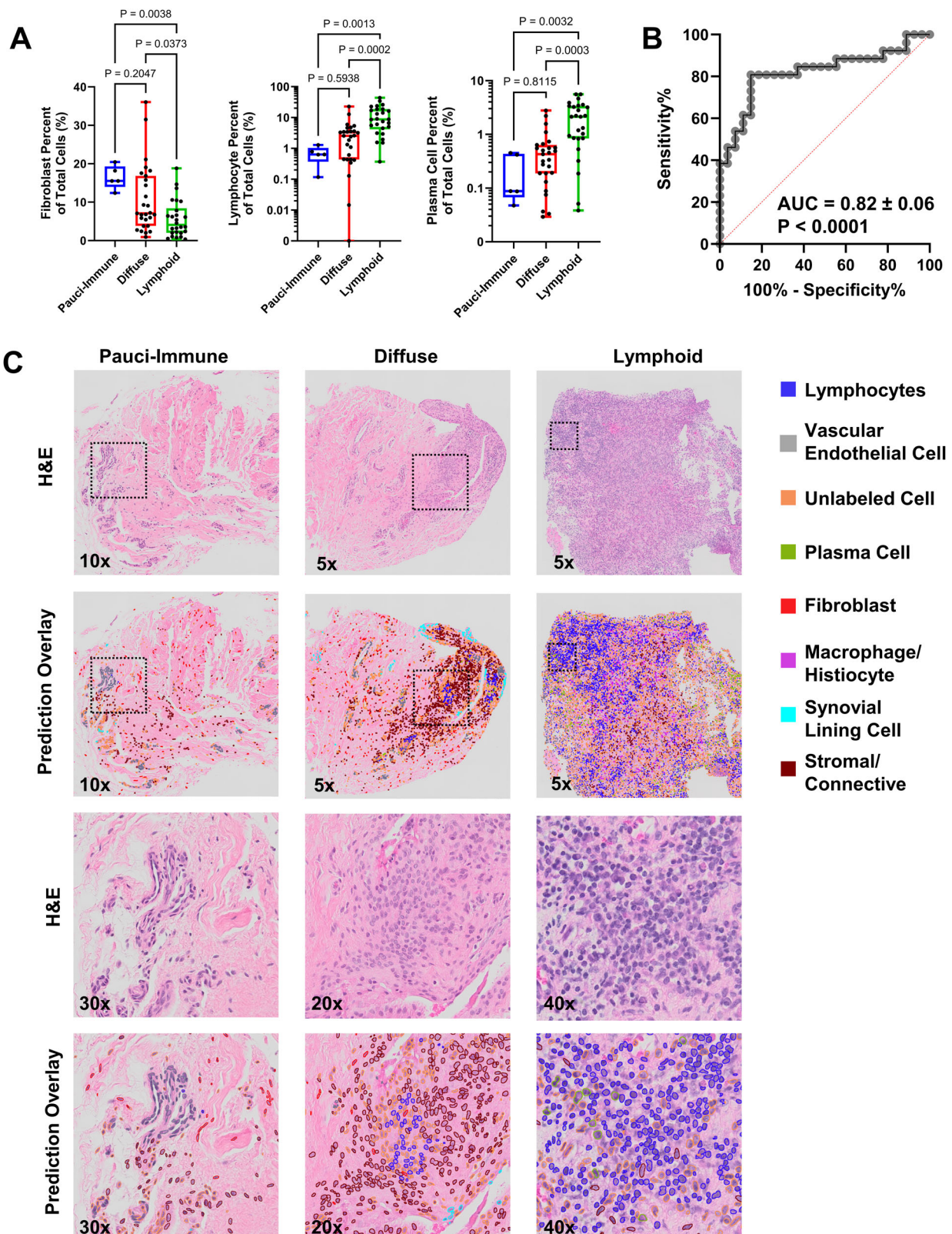
## Discussion

Here, we demonstrate that multi-scale modeling of synovial histopathology can pathotype RA and inflammatory arthritis in clinically meaningful settings, such as treatment response. Our model will reduce the analytical bottleneck associated with histopathology assessment in both the clinical and preclinical settings allowing quicker times to intervention or hypothesis resolution. Furthermore, it will reduce the amount of accessory immunostaining required for pathotyping by using H&E stains to infer cell types, like lymphocytes and plasma cells, which would otherwise need an immunohistochemical

(IHC) stain to confirm specific cell types. This could be very impactful as the diffuse pathotype (plasma cell poor with enrichment in myeloid cells) shows improved response to tocilizumab (anti-IL-6R antibody)<sup>4</sup> while the fibroblast rich pauci-immune pathotype shows inadequate response to anti-TNF therapy<sup>11</sup>. While we did not have enough specimens to build a model to classify the pauci-immune pathotype in the AMP-RA data set, we were able to classify a diffuse vs lymphoid pathotype utilizing a simple threshold on plasma cell percent with an AUROC of  $0.82 \pm 0.06$ .

Computational approaches to understanding tissue and cellular information from histology slides have greatly improved in recent years. Specifically, tools to segment or classify malignant tissues and cells from biopsy specimens<sup>28</sup>, and cell type classification on cytology blood smears<sup>29</sup> have seen the largest amount of development with many applications acquiring FDA approval<sup>30</sup>. These tools have dramatically increased the throughput of clinical histologic analysis by, for example, red flagging potential malignant cases with  $>0.95$  AUC for further review<sup>31</sup>, classifying and counting cells on cytology smears or slides for detection of malignant cells and other pathologies<sup>24,29</sup>. However, there is a dearth of computational pathology tools outside of cancer or outside of diagnostic settings and within the space of musculoskeletal pathologies<sup>32</sup>. Our work represents the first set of comprehensive tissue and cellular analysis tools for both pre-clinical and clinical phenotyping (in this case pathotyping) in inflammatory arthritis. In adjacent work in the oncology field, Pati and colleagues utilized hierarchical graph convolutional neural networks to integrate information from both the cellular and tissue levels on 2048 px x 1536 px sized images ( $0.42 \mu\text{m}^2$  pixels)<sup>33</sup>. This model was largely effective at detecting cancerous images with an F1 of  $84.9\% \pm 0.8\%$ , however did not perform well for non-cancerous, pre-cancerous or normal images with F1s of  $56.6 \pm 1.7\%$ ,  $66.1 \pm 3.7\%$ , and  $66.2 \pm 1.7\%$ , respectively. In other work, HoVer-Net is a state-of-the-art nucleus segmentation and cell type classification deep learning model for H&E-stained tissues<sup>24</sup>. HoVer-Net outperformed all other models in nuclei segmentation, however, only achieved cell classification F1s from 0.30–0.68. Thus, we utilized the nuclear segmentation portion of the model and transferred this knowledge into our custom feature extraction pipeline for classification of cells relevant in RA synovial pathology. In designing our feature extraction pipeline, we were inspired by Wang et al.<sup>25</sup> who used cell intrinsic features, contextual features about surrounding nuclei, density features and spatial arrangement features generated from a H&E tumor biopsy in a graph neural net framework to predict if patients would benefit from checkpoint inhibitor therapy. We extend this feature modeling framework to include both hematoxylin and eosin specific features, as well as cytoplasm features with statistical calculations of neighborhood features that empirically show improved that cell classification performance (Cell Intrinsic Features:  $0.75 \pm 0.01$  vs Cell Intrinsic Features plus 150 px distance features:  $0.86 \pm 0.02$  vs All Features:  $0.88 \pm 0.03$ , Supplemental Table 3 and Fig. 3). To the authors' knowledge, our work is the





**Fig. 6 | Cell type modeling can differentiate diffuse vs lymphoid cases with plasma cell counts alone.** **A** Cell type predictions were made on 58 RA biopsy specimens ( $n=5$  Pauci-Immune,  $n=27$  Diffuse,  $n=26$  Lymphoid) and plots of the Synovial Fibroblasts, Lymphocytes, and Plasma Cells percent of total cells demonstrate the known clinical differences among these pathotypes. Each dot is one human biopsy specimen, Box and Whisker plots are construct by showing the Min, 25th percentile, Median, 75th percentile and Max. All data was not normally

distributed. Kruskal–Wallis tests with Dunn’s post hoc were performed. **B** Using plasma cell counts alone, we can discriminate between diffuse and lymphoid cases with a ROC-AUC of  $0.82 \pm 0.06$  ( $n=53$ ). The optimal threshold is 0.82% plasma cells of total cells. **C** Representative H&E and cell type prediction overlays with low magnification and high magnification images of a pauci-immune case, diffuse case and lymphoid case.

first to utilize this type of framework in a cell type classification task. Independent validation of our human cell typing model is needed to formalize general applicability. However, we were able to utilize an independent cohort of mouse inflammatory arthritis slides (Fig. 4) which formally demonstrated the utility of such a model to recapitulate historical scoring measures and discover novel phenotypes.

Some computational approaches have already been successfully applied to H&E slides of synovial biopsies of RA patients to quantify cellular changes, such as nuclei density and its association with clinical inflammatory measures<sup>34</sup>, and simple counting of CD3+ T-cells or CD68+ macrophages on IHC<sup>35–37</sup>. Further, pathologist scores of specific cells types have been associated with quantitative inflammatory gene expression changes in the RA synovium<sup>38</sup>. Our approach aimed to incorporate the important cell types from this previous work while producing models that only require H&E-stained tissues. However, RA has been shown to be a complex polygenic autoimmune disorder with various environmental risk factors contributing to multiple etiologies<sup>4,39,40</sup>. As a consequence of this complexity, many RA patients are refractory to existing approved therapies<sup>41–44</sup>. This highlights the need for a personalized medicine approach to improve clinical trial design and treatment allocation, as done in recent trials like the R4RA<sup>6</sup> and PEAC<sup>26</sup> that utilize ultrasound guided synovial biopsies and pathology evaluation to stratify patients. Our models may be able to improve the workflow of these types of clinical trials and reduce the overall cost. To realize the full potential of our human models and understand unbiased performance, a formal independent validation cohort needs to be developed.

One major benefit of our model is the fact that we used H&E-stained tissues. These tissues are routinely and easily collected and represent a large proportion of historical datasets allowing for larger retrospective studies. While some of the computational pathology tools utilize H&E-stained slides, many have utilized special stains geared towards tissue- or cell type- specific classification<sup>32</sup>. This limits the overall throughput and utility of such pipelines by adding additional steps and costs that may be prohibitive. In addition, using a common stain facilitates transfer learning applications to other musculoskeletal pathologies, like osteoarthritis, bone fracture or disc degeneration, with different etiologies but similar tissue involvement and histopathology requirements. However, this fundamentally limits the specificity of some cell classes. For example, not all fibroblasts are long, thin and spindly; not all macrophages are plump, granulated cells; and not all plasma cells have a red/pinkish cytoplasm or clockface appearance; all H&E histologic features which our pathologist utilized during cell annotation (Supplemental Fig. 10). We address some of these concerns by staining adjacent slides with fiducial protein markers (CD3, CD20, and CD138) that largely validate our lymphocyte and plasma cell predictions. However, we found less consistency with macrophage identification, likely due to the heterogeneity of CD68+ cells. Until high dimensional biochemical methods for cell labeling that also allow subsequent high-quality H&E staining and imaging are developed, like spatial transcriptomics, we will be fundamentally limited when reaching for multi-class cell typing.

Consensus scoring that utilize Likert scales<sup>45</sup> is the gold standard analysis method for histopathology. For RA, various types of assessment, including Krenn lining and inflammation scores, rely on a consensus grading system to summarize high level pathologic features of the tissue (e.g. percent of area effected) and cells (e.g. ranges of quantity within a region)<sup>46,47</sup>. This reduces the challenge of quantifying complex and heterogeneous disease states. This approach also has an added benefit of often measuring large differences which usually correspond to clinically meaningful differences. However, this reduction in complexity may remove vital biologic information about treatment response or disease heterogeneity. For example, in our previous work we used a Likert scale to quantify synovial inflammatory infiltrate in which a score of 3 corresponded to “>30 inflammatory cells

thick”. If we had utilized this system in an interventional study which aimed to achieve a 50% reduction in synovial infiltrates but the number of cells across the synovium went from 100 cells thick to 50 cells thick, there would be no difference in the histology score despite a quite large treatment effect.

In the current work, we attempt to quantify this complexity with tissue segmentation and cell typing computational approaches. To demonstrate the benefit of these approaches, we can calculate the effect size (Glasse’s delta) between male and female TNF-Tg, 3 month-old synovium comparing the historical synovial inflammatory infiltrate scoring system (Bell et al., Fig. 2F<sup>19</sup>) with our machine learning predictions of lymphocyte counts within the synovium (Fig. 4A) of the exact same slides. This analysis reveals that our computational methods are orders of magnitude more sensitive to biologic differences than histologic scoring (Histologic Scoring: 2.79 Glasse’s delta vs Lymphocyte Counts: 74.82 Glasse’s delta). These differences in quantification method represent ~25 fold increase in measured effect size using the computational approach. This increased data sensitivity does, however, place an additional burden on the investigator to fully realize if a measured difference is clinically meaningful. Previously, this burden was partly shouldered by the scoring mechanism. In addition to more granular quantification of pathology these analyses are also more efficient. For example, to generate the annotations to build our segmentation model we spent 200+ hours drawing annotations on the 94 slides. However, to infer the tissue segment on the 174 slides in Fig. 2 the model took ~30 h of hands-off compute time with only 2–3 h of labor to visualize the results, representing a ~120-fold increase in efficiency.

Lastly, utilizing computational tools that quantify multiple tissues and cell types improves the ability to find novel phenomena. For example, while our primary focus is on the synovial pathology in inflammatory arthritis, having a model that measures cartilage, meniscus, and bone pathology provides a more comprehensive picture of disease. This allows easier detection of off-target or unexpected therapeutic effects with a singular methodology. However, reliance solely on computational modeling may increase false positives and expert-level quality control is advised for high impact results.

In conclusion, we have developed a set of models that can characterize tissue and cellular pathology in pre-clinical and clinical inflammatory arthritis settings. These models can be leveraged to better understand disease mechanisms in pre-clinical settings and be used in a precision medicine pipeline to improve patients’ health.

## Methods

### Dataset description

All mouse work was approved by the University Committee on Animal Resources at the University of Rochester Medical Center and the Institutional Animal Care and Use Committee at the Hospital for Special Surgery. Whole slide images (WSIs) of sagittal mouse knee sections were taken from two different mouse models of inflammatory arthritis and the accompanying controls for segmentation experiments (Supplemental Fig. 1A). Batch A consisted of male and female TNF transgenic mice (TNF-Tg,  $n = 47$ ) and wild-type littermates (WT,  $n = 15$ ) used in previous publications<sup>19,20</sup>. Batch B consisted of previously unpublished male and female knees that received intra-articular injections of 180  $\mu\text{g}$  of Zymosan to induce Zymosan Induced Arthritis (ZIA,  $n = 24$ ) and control contralateral limbs (Control,  $n = 8$ )<sup>48</sup> that were euthanized on Day 7 after injection. Different batches were used to test model generalizability across different biological mechanisms of arthritis development, differences in H&E staining protocols and slide scanners used to digitally capture slides at 40 $\times$  magnification (Batch A: VS120 Olympus, 0.173  $\mu\text{m}$  per pixel; Batch B: CS2 Aperio Leica, 0.253  $\mu\text{m}$  per pixel).

To further test model generalizability, 2 different independent holdout datasets were used to validate the model: (1) the remaining

H&E-stained sagittal knee WSIs from Bell et al.<sup>19,20</sup> that were not annotated or used in model training and (2) Orange G-H&E stained sagittal knee WSIs from Kenney et al.<sup>21</sup>. Slides from Bell et al. were ensured to not have been used in the initial model training, internal validation, or testing. These included slides from 6 month-old male TNF-Tg ( $n = 33$  slides from 14 knees) and WT littermates ( $n = 43$  slides from 17 knees), and slides from 9.5 month old male TNF-Tg mice (anti-TNF:  $n = 24$  slides from 8 knees; Placebo:  $n = 29$  slides from 10 knees) and WT littermates (Placebo:  $n = 42$  slides from 15 knees) either treated with a 6 week course of anti-TNF antibodies or placebo control (irrelevant IgG). To generate downstream measurements of tissues of interest, a region of interest (ROI) was drawn from the tibial growth plate to femoral growth plate including the anterior and posterior extra articular tissue.

WSIs of H&E-stained human synovial biopsies were collected from the Accelerating Medicines Partnership Rheumatoid Arthritis (AMP-RA) Phase II consortium<sup>47</sup>. In short, synovial tissue biopsies were acquired from RA patients at 13 different clinical sites in the United States and 2 in the United Kingdom from October 2016 to February 2020. The study was performed in accordance with protocols approved by the institutional review board at each site. The tissue was paraffin embedded, stained with H&E and imaged on a VS120 Olympus. Three pathologists independently determined Krenn lining and inflammatory infiltrates scores (0–3 each) for each tissue sample<sup>27</sup>, and the mode of the three scores was used for further analysis. To classify the cases into H&E based pathotypes, the UK Birmingham group developed consensus semiquantitative four point scores for infiltrate density and aggregate radial size on a per fragment basis with a custom atlas using a test set of tissues from the Birmingham Early Arthritis Cohort<sup>49</sup>, scored by three pathologists. Aggregate grade was derived as follows: Grade 3; high  $\geq 20$  radial count. Grade 2: medium 10–19 radial count. Grade 1: low 6–9 radial cell count. Grade 0: No aggregates. This approach was validated by scoring tissues from the first AMP RA cohort<sup>46</sup> and original data is presented from the second AMP RA cohort<sup>47</sup>. These semiquantitative scores were then used to classify the cases into three pathotypes, either lymphoid ( $n = 27$ ), diffuse ( $n = 26$ ) or pauci-immune ( $n = 5$ ) according to the following rules: Lymphoid: The presence of  $\geq 1$  grade 1 aggregate in at least two fragments, or any grade 2 aggregate, or any grade 3 aggregate. Diffuse: Does not meet lymphoid criteria but with a mean fragment density score  $\geq 1$ . Pauci-immune: Does not meet lymphoid criteria, mean fragment infiltrate density score  $< 1$ .

### Semantic segmentation annotation and preprocessing

Manual annotations were performed within QuPath<sup>50</sup> to assign labels for WSIs. To test model performance across tissue types at different granularity, multiple different class structures were tested (Supplemental Fig. 2). Eleven different classes were manually annotated, including synovium, muscle and tendon, growth plate, bone marrow, cortical bone, trabecular bone, articular cartilage, meniscus, fat, bone marrow fat, and histology artifact (i.e., out of focus). A seven-class, nine-class, and ten-class segmentation task was generated by merging histologically similar tissues, such as merging cartilage and meniscus into the same class (Supplemental Fig. 2). Overall, we estimate about 250 hours were spent annotating the 11 tissue classes on 94 WSIs.

Due to the gigapixel nature of WSIs, the entire slide cannot be fed directly into a deep learning model. Previous work has shown that WSIs can be broken into patches, in this case 512 pixel  $\times$  512 pixel, to perform downstream learning tasks<sup>31</sup>. For semantic segmentation, a custom QuPath script was used to export patches at a 4x downsample while filtering out regions of the scanned slide that lacked annotations or without tissue. Additionally, images were normalized to mean 0 and standard deviation 1 by sampling a subset of patches to get mean and standard deviation RGB statistics.

### Semantic segmentation models and training strategy

Initially, a stratified random sampling method was used to randomize the 94 annotated WSIs into a Training, Validation, and Test split, using 70%, 15%, and 15% for each split, respectively (3 splits total, Supplemental Fig. 1B). We stratified by batch (e.g. staining and site differences) and disease type (e.g. healthy and disease) to ensure even allocation of data variation into each set. Randomization occurred at the slide level, as opposed to the patch level, to ensure no data leakage across splits. During our initial qualitative explorations of models and hyperparameters (detailed below), we only utilized Training and Validation sets; and calculated the Dice score on the validation set to measure performance. Dice score was calculated as

$$\text{Dice} = \frac{2 * \text{intersection}}{\text{union} + \text{intersection}}$$

These initial experiments included variations in SLIC feature extraction, model selection (UNET vs UNET++), efficient-net backbone size (B0, B2, B5), loss metrics (weight loss vs unweighted), and learning rate parameters as described below. Once these items were tuned, quantitative experiments (details below) were performed varying the tissue segmentation number (7, 9, 10, or 11), image augmentation (None, Low, Medium, or High), or patch overlap percentage (0%, 50%, 66%). These experiments were performed by training the models with the Training and Validation sets, freezing the models' weights, and then inferring segmentation on the Test set. These inferences were then compared to the ground truth labels to calculate the mean Intersection over Union (mIOU) or the frequency weighted mIOU (fwIOU). The fwIOU is the class frequency weighted sum of the mIOU.

In addition, to assess how site-specific differences in histology slides may impact model performance<sup>51</sup>, we performed a single batch training method while varying the image augmentation style. In this training strategy, Batch A ( $n = 62$ ) was used for the Training (45%) and Validation (20%) sets, and Batch B ( $n = 32$ , 35%) was used as the held-out Test set. These experiments were performed to assess if image augmentation could overcome batch related staining differences that are seen in the real world.

Data augmentation strategies were also tested to assess their impact across the different training strategies (Supplemental Fig. 3). We had three different levels of augmentation tested, (1) None, (2) Low, (3) Medium, and (4) High. The python package imgaug<sup>52</sup> as used to implement augmentation. Augmentation was applied in the following way: (1) None had zero augmentation, (2) Low had 2–4 augmentation process applied 25% of the time, (3) Medium had 3–7 augmentations applied 50% of the time and (4) High augmentation had 5–11 augmentation process applied 50% of the time during training. Augmentation was randomly selected from 11 different types of augmentations including, horizontal flip ( $p = 0.5$ ), coarse dropout or pixel dropout from 0.2x the original image resolution ( $p = 0.1$ ), one of three different rotation types at 90°, 180°, and 270°, additive gaussian noise sampled from a normal distribution with mean 0 and variance 0.2\*255, blur using gaussian kernel with sigma of 1.5, hue modification using addition (-30,10) and saturation modification using multiplication (0.5,1.5) and linear contrast (0.5,2), brightness adjustment both add(-30,30) and multiply (0.5,1.5), and color change adjustment (3000, 8000). Data augmentation was performed only during the training of the models, not during testing or inference.

To compare segmentation performance using both conventional machine learning and deep learning methods, we tested two different model architectures. A Random Forest (RF) model implemented in QuPath with OpenCV<sup>53</sup> and an U-Net++<sup>54</sup> deep learning model implemented in PyTorch (1.8.0)<sup>55</sup>. We qualitatively assessed both pixel level segmentation and super-pixel level segmentation in QuPath and

determined super-pixel segmentation performed better. To generate super-pixels, we applied a Simple Linear Iterative Clustering (SLIC) algorithm ( $\sigma = 5$ , spacing = 20  $\mu\text{m}$ , Max Iterations = 1, Regularization = 0.01) in which each over segmented area was considered a super-pixel. We qualitatively assessed SLIC feature extraction variations and chose to extract RGB, estimated Hematoxylin, Eosin and residual stain means, standard deviations, min, max, and median values from each SLIC super pixel. We also calculated the Haralick value using a distance of 1 and bin of 32. We next calculated the average features of super-pixels within 40 and 80  $\mu\text{m}$ s. These features were then used in the RF, which was implemented fully in QuPath using their “Train Object Classifier” GUI with max depth = 20, min sample count = 10, Active variable count = 0, maximum trees = 50 and termination epsilon = 0. Four models were built, one for each of the 7 class, 9 class, 10 class and 11 class segmentation tasks. After the models were trained, they were used to infer on the Test set and performance metrics calculated.

For our deep learning pipeline, we utilized the segmentation\_models\_pytorch<sup>56</sup> package for the UNET++ implementation with an ImageNet pre-trained EfficientNet-B5 backbone for the encoder to improve training time and computational efficiency<sup>57</sup>. The decoder was left unchanged from the native UNET++ architecture except for the final layer which was changed to match the number of tissue types being segmented. We also initially explored using the UNET architecture and other efficient net backbones, however UNET++ with a B5 backbone was found to be most performant. We initially explored both class frequency weighted loss and un-weighted loss and the un-weighted loss demonstrated improved performance. We utilized a combo loss calculation, which was the arithmetic average of the dice loss and binary cross entropy (BCE) loss during model training<sup>58</sup>. We explored a few hyperparameter variations for the learning rate scheduler including step size = [1,2], gamma = [0.25, 0.5] and learning rate start = [0.01, 0.02, 0.025]. From initial explorations we found that a step size of 2, a gamma of 0.5 and a learning rate start of 0.025 resulted in the best performance. We used a stochastic gradient descent (SGD) optimizer with momentum (0.9) and regularization of  $1 \times 10^{-4}$  and the models were trained for 10 epochs at a batch size of 40.

When jointly training across 10 different classes, this model provided no prediction for the meniscus class resulting in a mean Intersection Over Union (mIOU) of  $0 \pm 0$  (Supplemental Fig. 6), likely due to insufficient training examples (Supplemental Fig. 2; 0.9% frequency overall). As the meniscus was important for downstream biological analyses, we developed a strategy to improve predictions for this class. A second UNET++ model was fine-tuned from the nine-class model (i.e. the model that has cartilage and meniscus merged into one class) by changing the prediction (final) layer specifically to predict between cartilage and meniscus. We then re-trained the full model only on images and Ground Truth annotations (GTs) that contained either cartilage or meniscus within the mixed training set.

Previous work had suggested that UNET based semantic segmentation can have image boundary level artifacts<sup>17</sup>. Therefore, we assessed how including patch level overlap for prediction can improve model performance. We included experiments using no overlap, 50% overlap and 66% overlap between them with images from both batches (Supplemental Fig. 4). To analyze the results, we looped through all the predictions ( $N$ ) for the entire WSI and calculated a majority vote for each pixel after thresholding to remove low confidence predictions (pixel value of 75).  $N$  can be variable depending upon the region, for example if it is on border, but typically is between 4 and 9.

### Semantic segmentation model evaluation, inference on external validation set and statistical evaluation

To evaluate the semantic segmentation model performance for the Validation and Test set slides with ground-truth labels, we calculated the mean Intersection over Union (mIOU) and frequency-weighted mIOU (fwIOU) to prevent very rare classes from drastically impacting

overall model performance<sup>59</sup>. All model hyperparameter optimization was performed on the Validation set, and once the above parameters were chosen the models were used to infer segmentation on the Test set and the mIOUs were calculated.

We used the optimal settings from the training/validation process to evaluate the model performance on the held-out datasets. Pearson's correlation was used to compare the hand drawn synovial tissue area reported in Bell et al.<sup>19</sup> with model classified synovial tissue area. The fine-tuned 10-class model was used to infer tissue segmentation on the held-out data<sup>21,60</sup>. Specifically, the UNET++ 9 class model was used to predict tissues classes on each patch, and if the nine-class model had a prediction output for the combined Cartilage-Meniscus class on a patch, then the patch was passed through the fine-tuned 2 class model to assign to either the meniscus or cartilage class. Predictions were merged by only allowing the fine tuned predictions to be within the predictions from the combined Cartilage-Meniscus class. Once the inference was complete, a ROI was drawn on each slide from femoral growth plate to tibial growth plate around the joint to restrict the downstream analysis to the joint space, subchondral bone, and synovial adjacent tissue. Tissue area was calculated for each slide and averaged for each knee then a One-Way ANOVA with Tukey's post-hoc adjustment was used to detect significant differences.

### Cell type classification framework and preprocessing

For cell type classification, a combination of transfer learning and active learning was used to identify several different cell types that exist within the joint tissue. Cell type classification can be broken into a two-step process, (1) segmentation and (2) classification. For cell segmentation, transfer learning was used by leveraging a deep learning model, HoVer-Net<sup>24</sup>, pretrained on the PanNuke dataset<sup>61</sup>, to extract nuclei regions. Image patches (1024  $\times$  1024 pixels) at 40x magnification were given as inputs, and ROI contours of nuclei were obtained to perform feature extraction upon. These nuclei with their features and labels (detailed below) were then leveraged in a gradient boosted decision tree (GBDT, XGBoost (<https://xgboost.readthedocs.io/en/stable/>)) implemented within the ScikitLearn package (<https://scikit-learn.org/stable/>)) model to classify cells.

The input for the classifier was derived from features extracted from each ROI generated by HoVer-Net. Specifically, every nucleus from the json file output of HoVer-Net was converted into a ROI (.roi) file to be read into FIJI/ImageJ<sup>62</sup> for feature extraction using a custom script. Detailed workflow for the ImageJ/FIJI analysis is described as follows. Each image was split using the built-in color deconvolution<sup>63</sup> algorithm in FIJI into hematoxylin and eosin color channels. For each channel the nuclei were measured for several different parameters, including morphological quantities (area, perimeter, circularity, feret's diameter, feret angle, aspect ratio, roundness, and solidity) and staining color quantities (mean, mode, min, max, standard deviation, skewness, median, and kurtosis). The nuclei ROI was then enlarged by 20 px, the original nuclei masked out, and the data from this surround 20 px was used to calculate cell specific cytoplasmic H&E color information for each cell. These features are called cell intrinsic features (Supplemental Table 1). Neighborhood characteristics of cells at several different distance ranges (150 px and 300 px) were used to include local cell and tissue level context into cell type classification (Supplemental Table 2)<sup>25</sup>. Simply, a neighborhood is determined as all the cells within a certain distance to the parent cell. These neighborhood characteristics included the average, standard deviation, skew, kurtosis, Z-score, interquartile range, standard error of the mean and entropy the cell intrinsic features of cells in the neighborhood. We also calculated shape characteristics of the neighborhood including the average distance of the neighboring cells, the linear correlation coefficient of the cells within the neighborhood, a string of up to 30 cells linear correlation coefficient, the straight line distance of a string of up to 30 cells, a scored measure of density of the cells and the number of

cells within the distance measure. A final total of 854 features were extracted for the downstream analysis.

Known healthy and pathologic cell types that contribute to inflammatory arthritis were then annotated on both mouse and human tissues (details below, Supplemental Figs. 8 and 10). The mouse proof-of-concept classification task consisted of bone-embedded cells, blood vessel cells, adipo-stromal cells (both adipose and stromal cells within fatty tissue), synovial fibroblasts (healthy and pathologic), chondrocytes, lymphocytes, and other synovial lining cells (healthy and pathologic) as detailed in Supplemental Fig. 8; annotated by subject-matter experts familiar with histologic analysis of these cell types. These cells were annotated on six healthy, eight mild disease and five severely diseased TNF-Tg knee sections. For human samples, a clinically meaningful set of cell types were labeled by a senior pathologist, following a standard cell type hierarchy (Supplemental Fig. 10). These included stromal/connective tissue cells, synovial lining cells, synovial fibroblasts, vascular endothelial cells, tissue macrophages/histocytes, lymphocytes, and plasma cells. These cells were labeled on five lymphoid, five diffuse and three pauci-immune cases. Nuclei were then mapped to manually annotated nuclei by checking if a nuclei's centroid (as determined by Hover-Net) was within an annotation mask.

### Mouse cell type classification model

A total of 4,712 cells were annotated for mouse cell type classification from seven different classes (Supplemental Fig. 8). Cells were labeled from a total of 19 different slides. A GBDT model was trained for cell type classification (XGBoost (<https://xgboost.readthedocs.io/en/stable/>) implemented within the Scikit-learn package (<https://scikit-learn.org/stable/>)), using stratified nested 5-fold cross validation with grid search to select the best models. In order to minimize the influence of annotations from any one slide, we enforced an even sampling method to ensure approximately equal numbers of cells from each slide appeared in all folds (sklearn.model\_selection.StratifiedKFold). To tune the parameters of the GBDT, we performed a grid search (sklearn.model\_selection.GridSearchCV) on the inner CV for learning rate = [0.05, 0.1, 0.2], colsample\_bytree = [0.6, 0.8, 1.0], subsample = [0.25, 0.5], max\_depth [6,12], n\_estimators = [10, 100, 200, 400], gamma = [0, 0.1, 0.3], and min\_child\_weight = [1,5,10]. To evaluate model performance, F1 statistics were calculated as the average of the 5 external folds and then the best performing models was used to infer cell type in two different biological settings, (1) to identify cell composition changes across different disease severities on the remaining cells (>300,000) on the 19 slides, and (2) identify differences between male and female mice in the context of disease progression in a held-out dataset<sup>19</sup>. Finally, average synovial inflammatory infiltrate scores and average pannus invasion scores were correlated with lymphocyte and synovial lining cell counts respectively (Spearman's Correlation).

### Feature ablation studies

To demonstrate the performance improvements of the distance features, we performed a feature ablation study within mouse cohort in which no distance features were utilized, features at distance 150 px and all features (cell intrinsic features, 150 px features and 300 px features) in our modeling framework (detailed above). To evaluate model performance, F1 statistics were calculated as the average of the 5 external folds.

### Active learning implementation

Human annotation was the time-consuming step for the cell type classification pipeline. Therefore, we applied the active learning strategy to improve the annotation efficiency for cell annotation of human samples. To develop this strategy, we tested a proof-of-concept active learning strategy using labeled data from the mouse H&E slides (Supplemental Fig. 9). Active learning is an iterative process that consists of three main steps, (1) annotation, (2) model training, and (3)

sample selection for further annotations. Its goal is to select the samples that can lead to the largest model performance improvement when adding to the training data after annotation. To validate the strategy, 100 different rounds of 5-fold cross-validation were performed. Average F1-scores were reported for each class and a macro-F1 score was additionally reported. 25 runs of 5-fold cross validation were removed due to cells from a single class not being present in both the training and testing sets. For the training dataset for each split, 5% of cells were first randomly selected as the first set of cells selected as being labeled annotated. Subsequently, the GBDT classifier was trained using this randomly selected data. Several different metrics for determining cells for annotation and subsequent model finetuning, including smallest margin uncertainty<sup>64</sup>, least confidence uncertainty<sup>65</sup> and entropy-based uncertainty<sup>64</sup> were assessed. The top 5% of cells were added to the training dataset and the cycle of model training and evaluation and new cell annotations continued until the entire training dataset was used. A random selection of cells after shuffling was also tested to compare model performance to the various active learning strategies. The package modal<sup>66</sup> was leveraged in our implementation. Mean and 95% confidence intervals are reports for each subset across the 75 different runs of 5-fold cross validation.

### Cell classification model evaluations

Confusion matrices were generated for model prediction along with F1-scores calculated as  $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$  where  $\text{precision} = \frac{TP}{TP+FP}$  and  $\text{recall} = \frac{TP}{TP+FN}$ , where TP, FP, FN stand for true positives, false positives, false negatives. Models were tested using known cell types within specific tissue types to evaluate the model qualitatively.

### Human synovial biopsy cell type modeling

Active learning was then leveraged for human cell type classification using H&E-stained slides of human synovial biopsy tissue of RA patients from the AMP consortium as described above. A subset of slides was selected to be annotated (Lymphoid,  $n = 5$ ; Diffuse,  $n = 5$ ; Pauci-Immune,  $n = 3$ ) that represent the diversity of specimens within this cohort. Multiple rounds of cell type labeling were performed with the assistance of active learning, to obtain a total of 2,639 cells grouped in seven different cell types, detailed in Supplemental Fig. 10 (stromal/connective tissue cells  $n = 597$ , synovial lining cells  $n = 309$ , synovial fibroblasts  $n = 189$ , vascular endothelial cells  $n = 486$ , Tissue Macrophages/Histocytes  $n = 201$ , lymphocytes  $n = 826$ , and plasma cells  $n = 310$ ). A cell type classification model using GBDT was trained using a stratified nested 5-fold cross validation with grid search strategy (as described above) to select the best models. F1 statistics were calculated as the average of the 5 external folds. The best performing model was used to infer cell types all cells on the slides within this patient cohort ( $n = 58$  subjects; 2,976,535 total cells). Summary cell type quantification (total cell counts and percent of total) was then assessed for each patient. Two analyses were performed using the derived cell types from the cell classification model. First, cell type counts and proportions were correlated with either immunofluorescent stained adjacent sections (described below) or with a pathologist-derived, and clinically relevant Krenn inflammation scores. As these data were non-normal, we utilized a Spearman's correlation. Second, we assessed the frequency of cell types across pathotypes. Specifically, statistical significance testing using lymphocyte, plasma cell, and fibroblast slide proportions were evaluated across pathotypes. Additionally, we performed a receiver-operator curve analysis of plasma cell frequency of total to predict if a biopsy was a lymphoid or diffuse case ( $n = 53$ ).

### Immunofluorescence (IF) and histomorphometry

Adjacent sections from 15 of the RA synovial biopsies were stained in batches with either CD3 (T-Cells), CD20 (B-Cells) and CD138 (Plasma Cells) or CLIC5 (Synovial Lining), CD3 (T-Cells), CD68 (Macrophages),

and CD34 (Vascular Endothelial Cells) antibodies; and counter stained with DAPI. In depth staining procedures are described in the original work<sup>47</sup>. All IF images were imported into QuPath to perform histomorphometry. All biopsies were evaluated for tissue morphology similarity to the adjacent H&E to ensure as little physical distance between the sections as possible. To count IF+ cell, DAPI+ cells were first segmented with a watershed algorithm in QuPath (cell detection) and then mean CD138, CD3, and CD20 IF intensity for each cell was calculated. Staining batch specific thresholds for each channel were used to count positive cells.

### Visualization of data

Uniform Manifold Approximation and Projection<sup>67</sup> visualization was used for feature representations between batches and cell type framework features. Both tissue segmentation masks and cell type masks for each class were reimported into QuPath<sup>50</sup> for visualization purposes.

### Statistical approach and implementation

All graphing and hypothesis testing statistics were performed in Prism (10.0, Graph Pad, Boston, MA). For all continuous variables, a Shapiro-Wilks Normality test was performed to assess normality. If the test determined the specific distribution to be non-normal, the equivalent non-parametric test was utilized to test for significance or correlation. If an ordinal variable was being associated with a continuous variable a non-parametric Spearman's correlation was chosen. Otherwise, One-Way, Two-Way and Two-Way Repeated Measures ANOVAs with Tukey's Post-Hoc tests were used to test for significant main effects, interaction effects and post-hoc pairwise differences. All pairwise tests are two-tailed. Specific test information including sample size for each figure is provided in the "Supplemental Statistical Information Pertaining to Data Presented in Figures" document.

### Software and hardware

QuPath (0.3.2 or later) was used to visualize WSIs and annotate tissues or cells as well as perform some image processing (detailed above). All other machine learning or deep learning techniques were performed in Python (3.8.1) as described above. Primary machine learning libraries include PyTorch (1.8.0), segmentation\_models\_pytorch (0.1.3)<sup>56</sup> Sklearn (scikit-learn, 1.3.2), and xgboost (2.0.2). Deep learning segmentation training was performed on four Nvidia V100's GPUs with 16 gb of RAM in parallel with a CUDA implementation (11.6.2). Segmentation inference was performed on either a Nvidia 3070 or 3090. Cell type classification was performed on an Nvidia 3070.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All data supporting the findings described in this manuscript are available in the article and in the Supplementary Information. Source data are provided with this paper as a Source Data file. AMP-RA data can be accessed via the source work<sup>47</sup>. Source data are provided with this paper.

### Code availability

All analysis scripts and models are provided at [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath). Full environment dependencies are in the spec-files in the Segmentation or Cell Classification folders.

### References

1. Firestein, G. S. The disease formerly known as rheumatoid arthritis. *Arthritis Res. Ther.* **16**, 114 (2014).
2. Lewis, M. J. et al. Molecular portraits of early rheumatoid arthritis identify clinical and treatment response phenotypes. *Cell Rep.* **28**, 2455–2470.e5 (2019).
3. Wang, J. et al. Synovial inflammatory pathways characterize anti-TNF-responsive rheumatoid arthritis patients. *Arthritis Rheumatol.* **74**, 1916–1927 (2022).
4. Rivellese, F. et al. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat. Med.* **28**, 1256–1268 (2022).
5. Micheroli, R. et al. Role of synovial fibroblast subsets across synovial pathotypes in rheumatoid arthritis: a deconvolution analysis. *RMD Open* **8**, e001949 (2022).
6. Humby, F. et al. Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16 week outcomes of a stratified, biopsy-driven, multicentre, open-label, phase 4 randomised controlled trial. *Lancet* **397**, 305–317 (2021).
7. Nerviani, A. et al. A pauci-immune synovial pathotype predicts inadequate response to TNF $\alpha$ -blockade in rheumatoid arthritis patients. *Front Immunol.* **11**, 845 (2020).
8. Elmore, J. G. et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama* **313**, 1122–1132 (2015).
9. Elmore, J. G. et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *bmj* **357**, j2813 (2017).
10. Pitzalis, C., Kelly, S. & Humby, F. New learnings on the pathophysiology of RA from synovial biopsies. *Curr. Opin. Rheumatol.* **25**, 334–344 (2013).
11. Nerviani, A. et al. A pauci-immune synovial pathotype predicts inadequate response to TNF $\alpha$ -blockade in rheumatoid arthritis patients. *Front Immunol.* **11**, 845 (2020).
12. Humby, F. et al. Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. *Ann. Rheum. Dis.* **78**, 761–772 (2019).
13. Manzo, A. et al. Histopathology of the synovial tissue: perspectives for biomarker development in chronic inflammatory arthritides. *Reumatismo* **70**, 121–132 (2018).
14. Chang, M. H. & Nigrovic, P. A. Antibody-dependent and -independent mechanisms of inflammatory arthritis. *JCI Insight.* **4**, e125278 (2019).
15. Raciti, P. et al. Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. *Arch. Pathol. Lab. Med.* **147**, 1178–118 (2022).
16. Konnaris, M. A. et al. Computational pathology for musculoskeletal conditions using machine learning: advances, trends, and challenges. *Arthritis Res. Ther.* **24**, 1–15 (2022).
17. Chan, L. et al. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proc. IEEE/CVF International Conference on Computer Vision*. 10661–10670 (IEEE, 2019).
18. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. image Anal.* **58**, 101544 (2019).
19. Bell, R. D. et al. Selective sexual dimorphisms in musculoskeletal and cardiopulmonary pathologic manifestations and mortality incidence in the tumor necrosis factor-transgenic mouse model of rheumatoid arthritis. *Arthritis Rheumatol.* **71**, 1512–1523 (2019).
20. Bell, R. D. et al. iNOS dependent and independent phases of lymph node expansion in mice with TNF-induced inflammatory-erosive arthritis. *Arthritis Res Ther.* **21**, 240 (2019).
21. Kenney, H. M. et al. Persistent popliteal lymphatic muscle cell coverage defects despite amelioration of arthritis and recovery of popliteal lymphatic vessel function in TNF-Tg mice following anti-TNF therapy. *Sci. Rep.* **12**, 12751 (2022).

22. Yi, X. et al. TNF-polarized macrophages produce insulin-like 6 peptide to stimulate bone formation in rheumatoid arthritis in mice. *J. Bone Min. Res.* **36**, 2426–2439 (2021).
23. Shealy, D. J. et al. Anti-TNF-alpha antibody allows healing of joint damage in polyarthritic transgenic mice. *Arthritis Res.* **4**, R7 (2002).
24. Graham, S. et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal.* **58**, 101563 (2019).
25. Wang, X. et al. Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors. *Sci. Adv.* **8**, eabn3966 (2022).
26. Lliso-Ribera, G. et al. Synovial tissue signatures enhance clinical classification and prognostic/treatment response algorithms in early inflammatory arthritis and predict requirement for subsequent biological therapy: results from the pathobiology of early arthritis cohort (PEAC). *Ann. Rheum. Dis.* **78**, 1642–1652 (2019).
27. Krenn, V. et al. Grading of chronic synovitis—a histopathological grading system for molecular and diagnostic pathology. *Pathol. Res. Pr.* **198**, 317–325 (2002).
28. Cifci, D. et al. AI in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? *Ann. Rev. Cancer Biol.* **7**, 57–71 (2023).
29. Jiang, H. et al. Deep learning for computational cytology: a survey. *Med. Image Anal.* **84**, 102691 (2023).
30. Muehlethaler, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health* **3**, e195–e203 (2021).
31. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
32. Konnaris, M. A. et al. Computational pathology for musculoskeletal conditions using machine learning: advances, trends, and challenges. *Arthritis Res. Ther.* **24**, 68 (2022).
33. Pati, P. et al. Hierarchical graph representations in digital pathology. *Med Image Anal.* **75**, 102264 (2022).
34. Guan, S. et al. Rheumatoid arthritis synovial inflammation quantification using computer vision. *ACR Open Rheumatol.* **4**, 322–331 (2022).
35. Kraan, M. C. et al. Quantification of the cell infiltrate in synovial tissue by digital image analysis. *Rheumatol. (Oxf.)* **39**, 43–49 (2000).
36. Haringman, J. J. et al. Synovial tissue macrophages: a sensitive biomarker for response to treatment in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **64**, 834–838 (2005).
37. Rooney, T. et al. Microscopic measurement of inflammation in synovial tissue: inter-observer agreement for manual quantitative, semiquantitative and computerised digital image analysis. *Ann. Rheum. Dis.* **66**, 1656–1660 (2007).
38. Orange, D. E. et al. Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histological features and RNA sequencing data. *Arthritis Rheumatol.* **70**, 690–701 (2018).
39. Firestein, G. S. The disease formerly known as rheumatoid arthritis. *Arthritis Res. Ther.* **16**, 1–3 (2014).
40. Mizoguchi, F. et al. Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis. *Nat. Commun.* **9**, 789 (2018).
41. Buch, M. H. Defining refractory rheumatoid arthritis. *Ann. Rheum. Dis.* **77**, 966–969 (2018).
42. Smolen, J. S. et al. Rheumatoid arthritis. *Nat. Rev. Dis. Prim.* **4**, 18001 (2018).
43. Fraenkel, L. et al. 2021 American college of rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis Rheumatol.* **73**, 1108–1123 (2021).
44. Donlin, L. T. Inching closer to precision treatment for rheumatoid arthritis. *Nat. Med.* **28**, 1129–1131 (2022).
45. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **22**, 55–55 (1932). **140**.
46. Zhang, F. et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.* **20**, 928–942 (2019).
47. Zhang, F. et al. Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature* **623**, 616–624 (2023).
48. Frasnelli, M. E. et al. TLR2 modulates inflammation in zymosan-induced arthritis in mice. *Arthritis Res. Ther.* **7**, 1–10 (2005).
49. Choi, I. Y. et al. Stromal cell markers are differentially expressed in the synovial tissue of patients with early arthritis. *PLoS One* **12**, e0182751 (2017).
50. Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 1–7 (2017).
51. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 1–13 (2021).
52. imgaug 0.4.0. *imgaug: A Library For Image Augmentation in Machine Learning Experiments.* <https://pypi.org/project/imgaug/> (2023).
53. Bradski, G. *The OpenCV Library.* <https://opencv.org/> (2000).
54. Zhou, Z. et al. U-net ++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018* (eds. Stoyanov, D. et al.) 11045 (Springer, Cham, 2018).
55. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Info. Process. Syst.* **32**, 8026–8037 (2019).
56. Iakubovskii, P. *Segmentation Models Pytorch.* [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch) (2019).
57. Tan, M. & Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *Int. Conf. Mach. Learn.* <https://doi.org/10.48550/arXiv.1905.11946> (2019).
58. Taghanaki, S. A. et al. Combo loss: handling input and output imbalance in multi-organ segmentation. *Comput. Med. Imaging Graph.* **75**, 24–33 (2019).
59. Thoma, M. A survey of semantic segmentation. *arXiv* <https://doi.org/10.48550/arXiv.1602.06541> (2016).
60. Stokbro, K. et al. Does mandible-first sequencing increase maxillary surgical accuracy in bimaxillary procedures? *J. Oral. Maxillofac. Surg.* **77**, 1882–1893 (2019).
61. Gamper, J. et al. Pannuke dataset extension, insights and baselines. *arXiv* <https://doi.org/10.48550/arXiv:2003.10778> (2020).
62. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. methods* **9**, 676–682 (2012).
63. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
64. Joshi, A. J., Porikli, F. & Papanikolopoulos, N. Multi-class active learning for image classification. In *2009 IEEE Conf. Computer Vision and Pattern Recognition.* 2372–2379 (IEEE, 2009).
65. Lewis, D. D. and J. Catlett, Heterogeneous uncertainty sampling for supervised learning. *Mach. Learn. Proc.* **1994**, 148–156 (1994).
66. Tivadar Danka, P. H. modAL: A modular active learning framework for pythonmodular active learning framework for {Python}. *arXiv* <https://doi.org/10.48550/arXiv.1805.00979> (2018).
67. McInnes, L., Healy, J., Saul, N. & GroBberger, L. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).

## Acknowledgements

The authors would like to sincerely thank the QuPath user community and development team at <https://forum.image.sc/tag/qupath>. Some

figures were made with BioRender.com (Supplemental Fig. 1A, B, Fig. 2B). This work was supported by Research into Inflammatory Arthritis Centre Versus Arthritis funding (to A.F.); National Institutes of Health (NIH) grants R21-AR071670, R01-AI175212, UH2-AR067690, and UC2-AR081025 (to J.H.A.); NIH grants F30-AG076326 and T32-GMO07356 (to H.M.K.); NIH grants R01-AR046713 and R01-AR050401 (to L.B.I.); NIH grants R01-AR078268, UC2-AR081025, and UL1-TR001866 (to D.E.O.); NIH grant R01-AR056702 (to E.M.S.); and National Science Foundation (NSF) grants 1750326 and 2212175 (to F.W.).

## Author contributions

R.D.B. conceived and designed all the experiments; acquired, analyzed, and interpreted the data; created new software; drafted and edited the manuscript; and approved the final version. M.B. designed the experiments, analyzed, and interpreted the data; created new software; drafted and edited the manuscript; and approved the final version. M.K., J.X., E.D., D.K., N.M., J.R.M., D.S.T., H.C., S.N., J.M. and H.M.K. acquired data, drafted and edited the manuscript; and approved the final version. M.O., M.A.F., J.A., L.D., D.O., A.F., and E.M.S. conceived and designed some the experiments; drafted and edited the manuscript; and approved the final version. Z.B. created new software; drafted and edited the manuscript; and approved the final version. L.B.I. and F.W. oversaw the work, conceived and designed some of the experiments; drafted and edited the manuscript; and approved the final version.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-51012-6>.

**Correspondence** and requests for materials should be addressed to Richard D. Bell.

**Peer review information** *Nature Communications* thanks Myles Lewis, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

<sup>1</sup>Arthritis and Tissue Degeneration Program and Research Institute, Hospital for Special Surgery, New York, NY, USA. <sup>2</sup>Weill Cornell Medical College, New York, NY, USA. <sup>3</sup>Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, USA. <sup>4</sup>Huck Institute of the Life Sciences, Pennsylvania State University, State College, University Park, PA, USA. <sup>5</sup>Orthopedic Soft Tissue Research Program, Hospital for Special Surgery, New York, NY, USA. <sup>6</sup>Horace Greely High School, Chappaqua, NY, USA. <sup>7</sup>Allergy, Immunology and Rheumatology Division, Department of Medicine, University of Rochester Medical Center, Rochester, NY, USA. <sup>8</sup>Rheumatology Research Group, Institute for Inflammation and Ageing, University of Birmingham, NIHR Birmingham Biomedical Research Center and Clinical Research Facility, University of Birmingham, Queen Elizabeth Hospital, Birmingham, UK. <sup>9</sup>Department of Pathology and Laboratory Medicine, Hospital for Special Surgery, New York, NY, USA. <sup>10</sup>Center for Musculoskeletal Research, University of Rochester Medical Center, Rochester, NY, USA. <sup>11</sup>The Rockefeller University, New York, NY, USA. <sup>28</sup>These authors contributed equally: Richard D. Bell, Matthew Brendel. <sup>29</sup>These authors jointly supervised this work: Andrew Filer, Lionel B. Ivashkiv, Fei Wang. ✉e-mail: [bellr@hss.edu](mailto:bellr@hss.edu)

## Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium

Jennifer Albrecht<sup>12</sup>, William Apruzzese<sup>13</sup>, Brendan F. Boyce<sup>12</sup>, David L. Boyle<sup>14</sup>, Michael B. Brenner<sup>13</sup>, S. Louis Bridges Jr.<sup>15</sup>, Christopher D. Buckley<sup>16</sup>, Jane H. Buckner<sup>17</sup>, Vivian P. Bykerk<sup>12,14</sup>, James Dolan<sup>13</sup>, Thomas M. Eisenhaure<sup>18</sup>, Andrew Filer<sup>16</sup>, Gary S. Firestein<sup>14</sup>, Chamith Y. Fonseka<sup>13,18</sup>, Ellen M. Gravallese<sup>19</sup>, Peter K. Gregersen<sup>20</sup>, Joel M. Guthridge<sup>21</sup>, Maria Gutierrez-Arcelus<sup>13,18</sup>, Nir Hacohen<sup>18</sup>, V. Michael Holers<sup>19</sup>, Laura B. Hughes<sup>15</sup>, Eddie A. James<sup>17</sup>, Judith A. James<sup>21</sup>, A. Helena Jonsson<sup>13</sup>, Josh Keegan<sup>13</sup>, Stephen Kelly<sup>22</sup>, James A. Lederer<sup>13</sup>, Yvonne C. Lee<sup>23</sup>, David J. Lieb<sup>18</sup>, Arthur M. Mandelin II<sup>23</sup>, Mandy J. McGeachy<sup>13</sup>, Michael A. McNamara<sup>12,14</sup>, Joseph R. Mears<sup>13,18</sup>, Fumitaka Mizoguchi<sup>13,24</sup>, Larry Moreland<sup>25</sup>, Jennifer P. Nguyen<sup>13</sup>, Akiko Noma<sup>18</sup>, Chad Nusbaum<sup>18</sup>, Harris Perlman<sup>23</sup>, Christopher T. Ritchlin<sup>12</sup>, William H. Robinson<sup>25</sup>, Mina Rohani-Pichavant<sup>25</sup>, Cristina Roza<sup>14</sup>, Karen Salomon-Escoto<sup>19</sup>, Jennifer Seifert<sup>19</sup>, Anupamaa Seshadri<sup>13</sup>, Kamil Slowikowski<sup>13,18</sup>, Danielle Sutherby<sup>18</sup>, Darren Tabechian<sup>12</sup>, Jason D. Turner<sup>16</sup>, Paul J. Utz<sup>26</sup>, Gerald F. M. Watts<sup>13</sup>, Kevin Wei<sup>13</sup>, Costantino Pitzalis<sup>27</sup>, Deepak A. Rao<sup>13</sup> & Soumya Raychaudhuri<sup>13</sup>

<sup>12</sup>University of Rochester Medical Center, Rochester, NY, USA. <sup>13</sup>Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. <sup>14</sup>University of California, San Diego, La Jolla, CA, USA. <sup>15</sup>University of Alabama at Birmingham, Birmingham, AL, USA. <sup>16</sup>University Hospitals Birmingham NHS Foundation



Trust and University of Birmingham, Birmingham, UK. <sup>17</sup>Benaroya Research Institute at Virginia Mason, Seattle, WA, USA. <sup>18</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>19</sup>University of Massachusetts Medical School, Worcester, MA, USA. <sup>20</sup>Feinstein Institute for Medical Research, Northwell Health, Manhasset, New York, NY, USA. <sup>21</sup>Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. <sup>22</sup>Barts Health NHS Trust, London, UK. <sup>23</sup>Northwestern University Feinberg School of Medicine, Chicago, IL, USA. <sup>24</sup>Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. <sup>25</sup>University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. <sup>26</sup>Stanford University School of Medicine, Palo Alto, CA, USA. <sup>27</sup>Centre for Experimental Medicine & Rheumatology, William Harvey Research Institute, Queen Mary University of London, London, UK.