Transformer-based artificial intelligence on single-cell clinical data for homeostatic mechanism inference and rational biomarker discovery

Authors
Veronica Tozzo[1,2,3 *], Lily H. Zhang[4, *], Rajesh Ranganath[4,5], John M. Higgins[1,2]

*These authors contributed equally.

Affiliation
[1] Department of Pathology and Center for Systems Biology, Massachusetts General Hospital, Boston, MA, USA

[2]Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[3]Department of Computational Medicine, UCLA, Los Angeles, CA, USA

[4]Center for Data Science, New York University, New York, USA

[5]Department of Computer Science, New York University, New York, USA

Corresponding authors
John Higgins
185 Cambridge Street, Suite 5.226
Boston, MA 02114
617-643-6129
higgins.john@mgh.harvard.edu

Veronica Tozzo
Gonda Center, room 2506B
695 Charles E. Young Drive South
Los Angeles, CA 90095
vtozzo@mednet.ucla.edu

Abstract word count
Word count
Number of tables
Number of figures

Abstract

Artificial intelligence (AI) applied to single-cell data has the potential to transform our understanding of biological systems by revealing patterns and mechanisms that simpler traditional methods miss. Here, we develop a general-purpose, interpretable AI pipeline consisting of two deep learning models: the Multi-Input Set Transformer++ (MIST) model for prediction and the single-cell FastShap model for interpretability. We apply this pipeline to a large set of routine clinical data containing single-cell measurements of circulating red blood cells (RBC), white blood cells (WBC), and platelets (PLT) to study population fluxes and homeostatic hematological mechanisms. We find that MIST can use these single-cell measurements to explain 70-82% of the variation in blood cell population sizes among patients (RBC count, PLT count, WBC count), compared to 5-20% explained with current approaches. MIST's accuracy implies that substantial information on cellular production and clearance is present in the single-cell measurements. MIST identified substantial crosstalk among RBC, WBC, and PLT populations, suggesting co-regulatory relationships that we validated and investigated using interpretability maps generated by single-cell FastShap. The maps identify granular single-cell subgroups most important for each population's size, enabling generation of evidence-based hypotheses for co-regulatory mechanisms. The interpretability maps also enable rational discovery of a single-WBC biomarker, "Down Shift", that complements an existing marker of inflammation and strengthens diagnostic associations with diseases including sepsis, heart disease, and diabetes. This study illustrates how single-cell data can be leveraged for mechanistic inference with potential clinical relevance and how this AI pipeline can be applied to power scientific discovery.

Introduction

Homeostatic processes regulating blood cell production and clearance are crucial for maintaining physiologic balance and achieve remarkable stability within individuals[1]. These processes are typically assessed at an aggregate level using complete blood count (CBC) parameters, which define the population sizes of red blood cells (RBC), white blood cells (WBC), and platelets (PLT). These population sizes are derived from over 100,000 single-cell cytometric measurements, which not only quantify the number of circulating cells but also capture cellular characteristics such as volume, protein content, or refractive index.[2,3] While some of these characteristics are used to derive markers providing more granular insights into homeostatic dynamics,[4–6] the full extent to which this single-cell data reflects underlying processes remains unknown. If more detailed signatures of homeostatic processes were present in single-cell characteristics, this information would likely yield greater in-depth understanding of pathophysiology as well as help discover novel biomarkers.

To investigate whether the distributions of single-cell characteristics across RBC, WBC, and PLT contain information about homeostatic regulation, we sought to predict population sizes from these distributions. In general, it is not obvious that one should be able to infer population size given a probability distribution of quantitative features of the members of that population. For instance, knowing the age or height distribution of a group of people does not enable accurate prediction of the number of people in the group.

To facilitate both accurate predictions from single-cell data and biological interpretability of the predictive models, we developed a transformer-based artificial intelligence (AI) pipeline for single-cell analysis. The pipeline consists of two models: Multi-Input Set Transformer++ (MIST), a prediction model that extends

state-of-the-art approaches[7], and single-cell FastShap, which identifies the most important cells for prediction by integrating MIST with efficient Shapley value computation[8]. The goal of predicting population sizes from single-cell characteristics tests a clear biological hypothesis and, as a self-supervised learning task, leverages all available data.

We find that MIST predicts blood cell population sizes from single-cell data with high accuracy. Ablation analysis reveals substantial crosstalk between populations, consistent with the presence of co-regulatory processes that link single-cell characteristics of one population (e.g. WBC) to the sizes of the others (e.g. RBC). Using single-cell FastShap, we quantified the effect of individual cells on predictions, allowing interpretation of the physiologic basis for MIST's accuracy. These findings generated data-driven hypotheses for co-regulatory mechanisms, shedding light on the clinical relevance of existing single-cell markers and enabling the discovery of a novel marker associated with future inflammatory states and major diseases including sepsis, heart attack, stroke, and diabetes.

Results

A transformer-based general-purpose deep learning pipeline for interpretable inference from single-cell data

We developed a pipeline composed of two novel transformer-based AI models. The first, multi-input Set Transformer++ (MIST), is a prediction model for sets as input that extends state-of-the-art permutation-invariant neural network architectures[7] to accept any number of single-cell distributions. MIST encodes each single-cell distribution separately and then merges these encoded representations to predict the desired target (**Figure 1** panel A). This flexible architecture allows the model to learn separate functions for each cell type while also modeling interactions between their encoded representations. The second model in the pipeline, single-cell FastShap (**Figure 1** panel B), quantifies the contribution of each single cell to the predicted population size. Single-cell FastShap combines MIST with recent advances in efficient amortized computation of Shapley values[8]. Shapley values provide rigorous grounding for interpretability in terms of game theory[12] but are expensive to compute.[13] Single-cell FastShap efficiently estimates Shapley values with a single forward pass of a neural network while satisfying desirable theoretical properties of interpretability methods[8] (see **Extended Figure 1** and **Methods** for more detail). Single-cell Shapley values correspond to each cell's marginal contribution to the average prediction across all combinations of cells. For our choice of a value function, a positive Shapley value indicates that inclusion of the cell increases the prediction on average, while a negative Shapley value indicates a decrease.

CBC measurements of cell population sizes and single-cell data

CBCs were measured on Advia 2120i automated hematology analyzers for 402,490 individuals treated at Massachusetts General Hospital between 2006-2012 with no exclusion criteria (see **Methods** and **Table S1**). CBCs measure the sizes of the RBC, WBC, and PLT populations in a microliter of blood, with size quantified as simple cell counts (N_{RBC}, N_{WBC}, and N_{PLT}), and the RBC population size also quantified in terms of the fraction of volume it occupies (N_{HCT}) and the total hemoglobin mass it contains (N_{HGB}). On the same blood sample, CBCs also measure optical scatter and fluorescence properties for >50,000 individual RBCs, WBCs, and PLTs. The single-cell data consists of four two-dimensional distributions measured on subsets of cells under different staining or surfactant conditions: $\boldsymbol{P_{RBC+PLT}}$ contains both RBCs and PLTs, $\boldsymbol{P_{RBC}}$ contains only RBCs, and $\boldsymbol{P_{WBC\_BASOS}}$ and $\boldsymbol{P_{WBC\_PEROX}}$ contain only WBCs (**Figure 1** panel C and **Methods**). When referring to the sub-region in $\boldsymbol{P_{RBC+PLT}}$ specific of platelets we will denote it $\boldsymbol{P_{PLT}}$. For analysis, we randomly

subsampled 1,000 cells from each distribution to ensure that no information on population size is present in the single-cell data.

MIST explains 70%-82% of the variance in cell population size

We trained five different MIST models to predict the cell populations sizes (N_{HCT}, N_{HGB}, N_{RBC}, N_{WBC}, or N_{PLT}) given all available single-cell data ($\boldsymbol{P_{RBC+PLT}}$, $\boldsymbol{P_{RBC}}$, $\boldsymbol{P_{WBC\text{-}BASOS}}$, and $\boldsymbol{P_{WBC\_PEROX}}$). **Figure 1** and **Table S2** show that MIST predictions explained at least 70% and up to 82% of the baseline variance in population sizes among individuals. Standard automated hematology analyzers are highly accurate but have some measurement noise (see **Table S3** and **Methods**). MIST predictions were within measurement noise for 21% of N_{HCT} , 16% of N_{HGB}, 36% of N_{RBC}, 36% of N_{PLT}, and 14% of $N_{WBC.}$ MIST's root mean squared error (RMSE) was 1.9x measurement noise for N_{HCT}, 2.7x for N_{HGB}, 1.1x for N_{RBC}, 1.1x for N_{PLT}, and 3.5x for N_{WBC}. RMSE was lower for counts within or below standard clinical reference intervals (**Figure S1)** and showed differences based on reported sex and age (**Figure S1, S2, S3**). However, the addition of sex and age as covariates did not improve prediction accuracy (**Table S2)**. MIST's high accuracy provides evidence that the single-cell data contains information on cell population sizes and that further investigation may yield insights into homeostatic processes.

Single-cell data has up to 15x more information on population size than standard CBC parameters

One reasonable null hypothesis for MIST's accuracy would be that the simple statistics of single-cell measurements (e.g. their average) are associated with cell population size. To investigate this, we tried to predict N_{HCT}, N_{HGB}, N_{RBC}, N_{WBC}, or N_{PLT} from standard CBC parameters: the mean volumes of RBCs and PLTs, the mean hemoglobin mass and hemoglobin concentration of RBCs, as well as the other population sizes (e.g. N_{WBC} for the prediction of N_{RBC}). Gradient-boosted tree models trained on these parameters were able to explain less than 20% of the variance, leaving an average of 59% of the variance unexplained compared to MIST (see **Figure 2, Methods,** and **Table S2** for details). Next, we expanded the inputs of the gradient-boosted tree models to include the first four moments of each of the four single-cell marginal distributions (mean, variance, skewness, kurtosis) as well as all 10 distribution deciles along each of the 8 single-cell dimensions (112 inputs total). Performance improved but still failed to explain at least 10% of the variance in population sizes explained by MIST. We also investigated the effect of the transformer-based architecture by comparing MIST to a non-transformer deep learning model for sets, Deep Sets++[7], and found that MIST explained more variance except for N_{WBC} where the methods were comparable (see **Table S2** and **Methods**). Even when population sizes for the other cell types were added as inputs, along with other standard reported CBC parameters, MIST only explained an additional ~4% of the variance compared to single-cell input only. These results imply that single-cell data contains a substantial amount of information on the dynamic processes regulating cell population size and that transformer-based methods can capture more of this information than other simpler approaches.

Single-cell data contains cross-population information on size

Prior studies have identified evidence for correlation among N_{RBC}, N_{WBC}, and N_{PLT} both at steady state[1] and in response to multiple disease processes[14–16]. We therefore investigated whether single-cell data from one cell population (e.g. $\boldsymbol{P_{RBC}}$) was being used by MIST to infer the sizes of other cell populations (e.g. N_{WBC}). We performed ablation analysis comparing MIST's prediction accuracy using all permutations of single-cell distributions as input (**Figure 3** panel A). Accuracy was generally lowest for predictions that used only one single-cell distribution at a time and typically improved as other single-cell distributions

were added. A notable exception to this pattern was N_{PLT} prediction performance where $\boldsymbol{P_{RBC+PLT}}$ on its own yielded significantly greater accuracy than the other three single-cell distributions combined. For all predictions, MIST was most accurate using all four single-cell distributions. This crosstalk and the greater accuracy of MIST compared to simpler methods together suggest that simple physiologic interpretations of the relationships between single-cell distributions and cell population sizes might not be adequate. We therefore applied single-cell FastShap to help interpret the relationships between single-cell data and each cell population size in physiologic terms (**Figure 3** panel B).

PLT population size is associated with single-cell data for PLTs but not RBCs or WBCs

Ablation analysis showed that $\boldsymbol{P_{RBC+PLT}}$ enabled accurate prediction of N_{PLT} on its own, with an RMSE of 45.90 (3.40) 10^3/μL similar to that achieved with all inputs 39.71 (0.31) 10^3/μL. $\boldsymbol{P_{RBC}}$, $\boldsymbol{P_{WBC\text{-}BASOS}}$, and $\boldsymbol{P_{WBC\_PEROX}}$ had negligible effects (**Figure 3** panel A). Single-cell FastShap analysis was consistent with the ablation analysis, showing high Shapley values concentrated in the extreme bottom left region of $\boldsymbol{P_{RBC+PLT}}$ where PLTs are located, and low signal in $\boldsymbol{P_{WBC\text{-}BASOS}}$ and negligible signal in $\boldsymbol{P_{RBC}}$ and $\boldsymbol{P_{WBC\_PEROX}}$ (**Figure** 3 panel B, **Extended Figure S2).** These results suggest that changes in the single-PLT distribution are systematically associated with changes in N_{PLT}. Little's Law for stationary systems establishes that cell population size (N) is equal to the product of birth rate (b) and mean lifespan (L): $N = b \cdot L$.[17] Therefore we can hypothesize that a higher density in the $\boldsymbol{P_{PLT}}$ distribution is associated with either elevated average PLT production rate or elevated PLT lifespan, and that the single-RBC and single-WBC distributions do not change when PLT production rate or PLT lifespan are modulated.

WBC population size is most strongly associated with single-WBC and single-PLT data

Ablation analysis showed that $\boldsymbol{P_{WBC\_PEROX}}$ alone predicted N_{WBC} with an RMSE of 3.35 (0.43) 10^3/μL, which fell within confidence interval of that of all-input prediction 2.66 (0.08) (**Figure 3** panel A and **Table S2**). Adding $\boldsymbol{P_{RBC+PLT}}$ to $\boldsymbol{P_{WBC\_PEROX}}$ improved accuracy slightly with a RMSE of 2.91 (0.42) 10^3/μL while $\boldsymbol{P_{RBC}}$ provided the least amount of information. Single-cell FastShap found high positive Shapley values in the region of $\boldsymbol{P_{WBC\_PEROX}}$ containing mostly lymphocytes[3] (top left boundary) in contrast to the region containing mostly neutrophils[3] (top right) which had moderately negative values (**Figure 3** panel B, **Extended Figure S3**). These results suggest that patients with higher N_{WBC} tend to have more lymphocytes and fewer neutrophils. Negative Shapley values were found in $\boldsymbol{P_{WBC\text{-}BASOS}}$, with stronger values in the left side of the distribution. $\boldsymbol{P_{RBC+PLT}}$ had very negative Shapley values in the bottom left region, suggesting a negative correlation between $\boldsymbol{P_{PLT}}$ and N_{WBC}. The Shapley values for $\boldsymbol{P_{RBC}}$ show areas with moderate influence on N_{WBC} in both directions with positive Shapley values near the top right of $\boldsymbol{P_{RBC}}$ and the lower left tail. Overall, this analysis finds evidence that changes to the compositions of the single-WBC and PLT distributions are associated with changes to N_{WBC} and suggestions that the composition of the single-RBC distribution is linked to N_{WBC} as well.

RBC population size is associated with all single-cell distributions

Ablation analysis (**Figure 2** panel A) shows generally steady improvement in MIST accuracy for estimating the RBC population sizes (N_{HCT}, N_{HGB}, N_{RBC}) as additional single-cell distributions are included as inputs. The $\boldsymbol{P_{RBC+PLT}}$ and $\boldsymbol{P_{WBC\text{-}BASOS}}$ distributions alone yielded a RMSE of 3.07 (0.18) % for N_{HCT} , 1.03 (0.09) g/dL for N_{HGB},, and 0.35 (0.01) 10^6/μL for N_{RBC}, all of which were within one standard error of the full input prediction performance (**Table S2**). The $\boldsymbol{P_{WBC\_PEROX}}$ distribution had the smallest effect on accuracy, and when excluded, the residual accuracy was close to that for all inputs (RMSE 2.86 (0.03) % for N_{HCT}, 0.98 (0.08) g/dL for N_{HGB}, and 0.32 (0.01) 10^3/μL for N_{RBC}. Single cell Fastshap interpretability maps have some similarities (**Figure 3** panel B, **Extended Figures 4-6**): the PLT area of $\boldsymbol{P_{RBC+PLT}}$ has negative Shapley values, and the left side of $\boldsymbol{P_{WBC\text{-}BASOS}}$ has high positive Shapely values at the top (top left) and moderately negative

at the bottom (bottom left). The Shapley values for $\boldsymbol{P_{RBC}}$ are consistently moderately positive in the top right where reticulocytes appear[4]. The presence of more reticulocytes may be associated with a faster RBC birth rate, and N_{HCT}, N_{HGB}, and N_{RBC} would then be expected to increase if mean lifespan and other cellular characteristics remain stable. We also observe different shapes in the reticulocytes' region for the three population sizes, indicating that different regulation patterns might be present. Single-cell FastShap analysis thus finds evidence that all single-cell distributions systematically associated with changes in N_{HCT}, N_{HGB}, and N_{RBC}, suggesting that on average across individuals RBC, WBC, and PLT population compositions are altered in the course of regulating RBC population sizes.

Rational discovery of single-cell markers of cellular kinetics

Single-cell FastShap finds high Shapley values in the left side region of $\boldsymbol{P_{WBC\text{-}BASOS}}$ which is used to define the clinical "left shift" (LS) marker (see **Supplementary Methods**)[18]. It also finds a novel connection between this area and N_{HCT}, N_{HGB}, and N_{RBC} with a clear demarcation separating positive and negative Shapley values. The juxtaposition of positive and negative associations suggests that a shift of white blood cells from the top left of $\boldsymbol{P_{WBC\text{-}BASOS}}$ distribution to the bottom left may be strongly associated with a decrease in the size of N_{HCT}, N_{HGB}, and N_{RBC}. We tested this hypothesis by quantifying the "down shift" (DS) of probability density in terms of the y-coordinates of the centroids in the leftmost portion of $\boldsymbol{P_{WBC\text{-}BASOS}}$ (**Figure 4** panel A and **Methods**). **Figure 4** panel B shows that the presence of DS is associated with a significant decrease in N_{HCT} -1.98 % (as well a N_{HGB} significant decrease of -0.67 g/dL and N_{RBC} -0.20 10^3/μL) and a significantly higher average N_{WBC} in our study cohort. The presence of LS is associated with a smaller but significant decrease in N_{HCT}, -1.54 %, and a higher N_{WBC} of 2.54 10^3/μL. All differences were tested a two-sided independent t-test and had a p-value of less than 0.0001. The existing clinical LS flag reflects WBC dynamics, and because DS seems to reflect RBC dynamics, DS may complement LS in some clinical situations where changes in both WBC and RBC populations occur.

Down Shift complements Left Shift to improve risk stratification for multiple major diseases

We analyzed diagnostic associations for LS, DS, and their combination in our study cohort. Individuals were divided into four groups based on whether they had LS, DS, or both (LS+DS). DS was associated with a larger change in N_{WBC} and N_{HCT} in those with and without LS (**Figure 5** panel A). We also tested their association with other markers of inflammation: erythrocyte sedimentation rate (ESR)[19] and C-reactive protein (CRP)[20]. **Figure 5** panel B shows that both LS and DS were significantly associated with elevated ESR with DS and LS+DS having higher odds ratio. DS and LS were also significantly associated with elevated CRP, with DS having a slightly higher odds ratio. ESR had a prevalence of 54 in the analyzed cohort and the positive predictive value (PPV) for elevated ESR was 65, 68 and 72 for LS, DS, and LS+DS respectively. CRP had a prevalence of 15 and PPVs for elevated CRP were 18, 18 and 19 (see **Table S4**). Next, we investigated the associations of LS, DS, and LS+DS with new diagnoses made within [0, 30] days after the CBC measurement, adjusting for age, sex, and N_{WBC} (see **Methods**). **Figure 5** panel C and **Table S5** show that LS and DS were both significantly associated with new diagnosis with similar signal, but LS+DS was more strongly associated with many new diagnoses than LS or DS alone suggesting that DS complements LS in capturing inflammatory processes.

Discussion

This study demonstrates that routinely available single-cell blood data contains enough information for a near-complete characterization of the dynamic processes regulating cell population size, which can be

leveraged to gather new insights into co-regulatory mechanisms and discover new biomarkers. It also shows that interpretable transformer-based deep learning pipeline can increase information extraction and can be used to power scientific discovery.

We find that MIST, a transformer-based permutation-invariant deep learning model[7], applied on single-cell blood data (***P*****$_{RBC+PLT}$**, ***P*****$_{RBC}$**, ***P*****$_{WBC-BASOS}$**, and ***P*****$_{WBC\_PEROX}$**) can explain the majority of inter-patient variation in the sizes of circulating RBCs, WBCs, and PLTs populations (N_{HCT}, N_{HGB}, N_{RBC}, N_{WBC}, or N_{PLT}). MIST's accuracy implies that single-cell data encodes enough information about the dynamic and physiologic processes such as production and clearance regulating blood cell populations. Previous studies have shown that single-cell data contains information on RBC or WBC population dynamics.[21–25] However, these studies focus on one blood cell population at a time using simple mechanistic models[2] which lack the flexibility to scale to large datasets and may fail to capture complexity such as crosstalk. Our findings provide substantial proof of concept that single-cell blood data can be leveraged beyond its current clinical use, where only few parameters are calculated before the rest is discarded. Since CBC parameters are essential diagnostic markers to assess an individual's hematologic, immunologic, and hemodynamic states, any additional information supplementing these parameters could enhance health monitoring. For instance, recent research shows that small differences in steady-state CBC parameters among apparently healthy patients are associated with major diseases and mortality[1]. If single-cell data can provide finer-grained insights into these changes, for example by describing these small changes as summation of production it could significantly improve risk assessment.

Ablation analysis and single-cell FastShap, a permutation-invariant deep learning interpretability model, provided evidence of co-regulatory mechanisms, with modulation of all three single-cell distributions associated with N_{HCT}, N_{HGB}, N_{RBC}, and N_{WB}, and modulations of ***P*****$_{PLT}$** associated with changes in N_{PLT}. The causality of these associations cannot be established without follow-up studies, but results are consistent with the hypothesis that RBC and WBC regulation pre-empt the processes governing ***P*****$_{PLT}$** while the opposite may not be true. Consistent co-regulatory interactions have been found in studies evaluating co-regulatory connections at the population level[1,14,16] and found in studies in presence of pathologies [cits].

Single-cell Shapley values found interesting patterns for the prediction of N_{WBC}, with the density in the ***P*****$_{WBC\_PEROX}$** area corresponding to neutrophils and lymphocytes associated with N_{WBC} negatively and positively respectively consistent with a recent study[1] and with lymphocytes generally having a longer mean lifespan than neutrophils[26,27]. These inter-patient associations are the opposite of what would be expected for intra-patient associations, where acute inflammatory responses are largely driven by increasing neutrophils[14]. The left side of ***P*****$_{WBC-BASOS}$** -- which contains monocytes, lymphocytes, and immature neutrophils -- had negative Shapley values for prediction of N_{WBC}. Monocytes are generally reported to have the shortest mean lifespan of those three types[26,27], and an increase in their fraction would therefore be consistent with lower N_{WBC}. Negative correlations between N_{WBC} and ***P*****$_{PLT}$** are corroborated by correlations found during acute inflammatory responses[14,15], but at steady state, it is not clear that there is a significant correlation between N_{WBC} and N_{PLT} in either direction[1]. Lastly, in the **P$_{RBC}$** distribution negative and positive Shapley values were found in the reticulocytes region and lower left tail of the distribution which may reflect delayed RBC turnover or increased RBC lifespan.[22–24]

We found that N_{HCT}, N_{HGB}, N_{RBC} were more strongly associated with changes in $\boldsymbol{P_{PLT}}$ and $\boldsymbol{P_{WBC\text{-}BASOS}}$ distributions than changes in the $\boldsymbol{P_{RBC}}$ distribution, with overall an inverse correlation. Our findings are corroborated by previous work showing that and red cells are co-regulated with white cells in presence of inflammation[28] and of thrombocytosis in presence of anemia[16,29].The positive Shapley values found in the top right region of $\boldsymbol{P_{RBC}}$ for the prediction of N_{HCT}, N_{HGB}, and N_{RBC} are consistent with the reticulocyte count, a well-established clinical marker of RBC population kinetics[4], with higher reticulocyte signifying higher rate of RBC production. Given that different regions are highlighted for N_{HCT}, N_{HGB}, and N_{RBC}, it is possible that this single-cell FastShap analysis could be used to identify variants of the reticulocyte count specific to each of these population sizes. The evidence found by single-cell FastShap in the left region of $\boldsymbol{P_{WBC\text{-}BASOS}}$ is relevant for a commonly-available single-cell clinical marker, the "left shift" flag (LS), which identifies the presence of an unusually large number of immature neutrophils, as can be seen in acute bacterial infections[18,30,31].

The split between negative and positive association in the left region of $\boldsymbol{P_{WBC\text{-}BASOS}}$ was a completely novel finding, which allowed us to define a novel marker of inflammation, Down Shift (DS), that it appears to be specific of changes in both WBC and RBC populations. DS was significantly associated with future markers of inflammation and diagnosis and complemented clinically used WBC marker to improve diagnostic accuracy for several important diseases and pathologic processes. Future work is required to validate prospectively the associations found for DS, and clinical application of this marker will require to expand its definition to other hematology instruments.

Our interpretable transformer-based single-cell pipeline thus allowed us to not only to prove that single-cell blood data contains relevant homeostatic dynamics information, but it also led us to discover novel co-regulatory mechanisms as well as finding a translational biomarker which is novel in the AI domain where often methods are applied as a black box and do not lead to translational biological or clinical insights. Our pipeline is general-purpose and can be applied to other types of single cell data including other flow cytometry measurements and single cell omics and it also offers an important advance in foundational models for single-cell analysis by addressing one of the major challenges in the field: interpretability[32–34]. Single-cell FastShap enables precise identification of which cells contribute to predictions in a theoretically grounded and computationally efficient manner enabling generation of evidence-based hypotheses guiding the prediction.

Our AI pipeline and blood single-cell data can be leveraged for wide range of future works. First, the same approach used for the rational development of Down Shift could be used to detect disease-specific single-cell biomarkers by directly predicting clinical diagnosis. In this context, our pre-trained models could be used as foundational models in presence of small sample sizes or biased labels. Second, MIST can be used to detect deviation from steady state. These estimates might be less bias than those obtained by CBC parameters due to summation of processes such as production and clearance as well as physiologically changes such as hypovolemia and hypervolemia. Third, MIST could be used to provide estimates of current age distribution of circulating RBCs, WBCs, and PLTs, as well as more accurate estimates of their production rates from a single blood test[1] with applications in many clinical contexts including but not limited to the diagnosis and management of diabetes.[43]

Methods

Data collection

Data was collected at Massachusetts General Hospital in the period between 2006 and 2012 for a total of 402,490 individuals and ~3 million CBC blood tests. Demographics of the analyzed cohort can be found in **Table S1**. We consider only one laboratory test per individual, selected at random. Clinical information was obtained retrospectively from medical records. The Mass General Brigham Institutional Review Board approved the study and waived the requirement for informed consent. There were no exclusion criteria, and all available individuals were included in the analysis.

CBCs were measured on Advia 2120i blood analyzers[47] which utilize flow cytometric principles to measure light scatter properties of individual cells. Three single-cell distributions are generated: one that measures red blood cells and platelets, one that measures white blood cell counts based on nuclear density and produces the common white blood cell count ($\boldsymbol{P_{WBC\text{-}BASOS}}$), and one that measures peroxidase activity in the cytoplasm and provides leukocyte differential counts such as neutrophils and lymphocytes ($\boldsymbol{P_{WBC\_PEROX}}$). For the red blood cells and platelets distribution, two versions were considered: the original distribution containing platelet measurements ($\boldsymbol{P_{RBC+PLT}}$), and a processed version where platelets are removed and measurements are projected using the Mie Transform[48] to produce single red blood cell volume and hemoglobin content ($\boldsymbol{P_{RBC}}$). We chose to not extract the platelet distribution as it is hard to obtain a proper division between platelets and red cells sedimentation. Additional details on flow cytometry data are available in the **Supplementary Methods**. Each of the generated distributions is two-dimensional and contains approximately 50,000 cells (visualization in **Figure 1 panel C**). Each of these distributions was subsampled to 1,000 samples, removing information about relative counts and to make model training computationally manageable with deep learning on a single 24GB consumer GPU while retaining cell population statistics with minimal variance. Each distribution is then normalized by the mean and standard deviation across the population. From the ADVIA blood analyzer, we also collected the following blood cell clinical indices which were used as dependent variables: hematocrit, hemoglobin, red cell count, platelet count, and white cell count. During the period considered, three different ADVIA blood analyzers were used in the laboratories and information on each different analyzer was collected from raw data associated with the CBC laboratory test.

Multi-input Set Transformer++ (MIST) for single cell data

A key property of single-cell data that differentiates it from tabular, image, or sequential data types is the lack of ordering between cells in a sample. This physiological property can be encoded computationally through models that respect permutation invariance, i.e., any reshuffling of the input produces the same output. Each input distribution ($\boldsymbol{P_{RBC+PLT}}$, $\boldsymbol{P_{RBC}}$, $\boldsymbol{P_{WBC\text{-}BASOS}}$, and $\boldsymbol{P_{WBC\_PEROX}}$) is represented as matrix $X \in R^{N \times 2}$ where N is the number of cells. For each individual $i \in [1, \dots, M]$, the full data input is represented by $\boldsymbol{X^i} = \{X^i_{RBC}, X^i_{RBC+PLT}, X^i_{WBC\_BASOS}, X^i_{WBC\_PEROX}, X^i_{cov}\}$ where $X^i_{cov} \in R^d$ is a vector of covariates containing age, year of measurement, and the identity of the blood analyzer used for the test.

MIST respects the permutation invariance property within each distribution of cells per individual. Specifically, let H be the dimension of the latent space, $\phi: R^{N \times 2} \rightarrow R^{N \times H}$ be equivariant encoder, meaning any reordering of cell for a given input results in the corresponding ordering of the output, and $\varphi: R^{N \times H} \rightarrow R^{H}$ be a permutation-invariant aggregation function that maps each cell type to a latent vector. Then, each input single cell distribution can be encoded as $\varphi_t(\phi_t(X_t)) \in R^H$ for $t \in \{RBC, RBC + PLT, BASOS, PEROX\}$. The encoded distributions in the latent space can be used optionally alongside the additional covariates vector X_{cov}. The concatenated vector is then passed into a decoder

function ρ: $R^{(4\times H+d)} \rightarrow R$ that produces the output. Overall, a multi-input single cell deep architecture can be defined as

$$f(\boldsymbol{X};\ \theta) = \rho\left(\left[\left\{\varphi_t\left(\phi_t(X_t)\right) \mid t = [RBC, RBC+PLT, BASOS, PEROX\};\ X_{dem}\right]\right) \quad (1)$$

where θ denotes parameters of overall architecture. We instantiate ϕ, φ, and ρ to follow the Set Transformer++ architecture[7].

We created three datasets with similar demographics (**Table S6**) of N=268326 training samples and N=134164 test samples such that the test splits of each dataset are non-overlapping. The validation set was selected to be a random 10% subset of the training set. For each dependent variable (e.g. hematocrit, white cell count), a separate model was trained for each split for a total of 15 models, or three models per output. The models were trained for 30 epochs with a batch size of 64, with the final model chosen based on validation loss. Models were trained on a NVIDIA TITAN RTX 24GB or Tesla V100-PCIE-32GB GPU. The number of model parameters and running times are available in **Table S7**.

To understand the role of demographic and machine information on the prediction task, we train MIST both with and without the additional covariates vector. We also train a Gradient Boosted Tree (GBT) model to predict CBC values using this covariates vector alone. The MIST model performs similarly both with and without the additional covariates vector, suggesting that the additional covariates do not add additional information over the single cell data; we verified this by predicting the counts using this vector alone obtaining only minimally better performances than a constant baseline prediction and significantly worse than prediction with the single cell data (**Table S2**). Given the lack of additional information provided by the available covariates the main results in the text are presented for a MIST model without them.

Prediction with other CBC parameters

The CBC laboratory test produces ten CBC parameters; in addition to the five used as main outputs, the test also provides mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), red cell distribution width (RDW), mean corpuscular hemoglobin concentration (MCHC), and mean platelet volume (MPV). To evaluate the efficacy of single cell data compared to the existing ten CBC parameters we train a gradient-boosted decision tree (GBT) to predict the five main outputs using every other CBC parameter. Due to the correlation between hematocrit, hemoglobin, and red cell count, we do not use any of these indices as input when predicting any of these indices as output. Concretely, we use platelet count, white cell count, MCV, MCH, RDW, MCHC, and MPV as input for prediction of hematocrit, hemoglobin and red cell count; hemoglobin, hematocrit, red cell count, white cell count, MCV, MCH, RDW, MCHC, and MPV for the prediction of platelet count; and hemoglobin, hematocrit, red cell count, platelet count, MCV, MCH, RDW, MCHC, and MPV for the prediction of white cell count. We use a GBT model with ten estimators and sweep hyperparameters including learning rate (0.001, 0.01, 0.1) and minimum number of samples allowed for a split (2, 4, 8).

MIST comparison with gradient-boosted tree and Deep Sets++

To confirm the need to use a transformer-based pipeline we compare MIST with two baselines. The first is a multi-input variation of Deep Sets++ (DS) using the same single-cell data used as input for MIST. The DS model consists of a 2-layer multi-layer perceptron (MLP) encoder, a sum aggregation, and a 3-layer MLP decoder, all with He-style residual connections and set norm[7]. This model was used to evaluate the utility of a transformer-based architecture. The second comparison was done with a gradient-boosted trees (GBT) model trained on hand-engineered statistics from the single-cell distributions. This comparison was performed to verify that simple methods on easily interpretable hand-engineered

features capture less information than our complex pipeline. The GBT model sees up to the fourth-order moment (mean, standard deviation, skew, kurtosis) and quantiles (every 10th) of each distribution along both dimension (i.e. cell size and content). We perform cross-validation grid search over the number of trees (100 or 200), learning rate (.001, .01, .1), and minimum number of samples allowed for a split (2, 4, 8).

Comparison with measurement noise

Measurement noise denoted as a coefficient of variation (CV) was collected from available literature[49] and is available in **Table S3**. For comparison with population-wide RMSE we calculated the measurement noise at the average value of each output in the analyzed cohort. For example, in the case of HCT the CV was 1.8 and measurement noise was calculated as $\frac{1.8 \times \frac{1}{\mathrm{N}} \sum \mathrm{HCT_i}}{100}$ where i denotes a different individual. To assess how many predictions fell within measurement noise we performed the calculation at the individual level. Therefore, if HCT_i is the measured hematocrit level for the individual i and $\widehat{HCT_i}$ is MIST prediction, we consider the prediction to fall within measurement error if $|\widehat{HCT_i} - HCT_i| < \frac{1.8 \times HCT_i}{100}$.

Single-cell Shapley values

To evaluate the role of each cell in the predictions, we estimate their Shapley values. Shapley values are defined based on a particular choice of value function dictating the worth of an input subset, and Shapley values attribute the contribution of the overall value across different players (in this case, cells) according to the desirable theoretical properties of efficiency, symmetry, and additivity.

The Shapley value $\varphi_c(v)$ of a cell c depends on the choice of value function $v(c)$ and is calculated as the marginal contribution of the cell c averaged over all possible subsets of cells. The marginal contribution is computed by comparing the value of a randomly sampled subset s with and without cell c. The subsets s will have different sizes to account for interactions and complementary effects among the cells. Let $n \sim Unif(N)$ denote that integer n is sampled uniformly from 0 to $N-1$, and let $s \sim Unif(P_{X\setminus\{c\}}(n))$ denote that subset s sampled from a uniform distribution over cell subsets of size n that do not include cell c. Then, the Shapley value φ_c is defined formally as

$$\varphi_c(v) = \mathbb{E}_{n \sim Unif(N)} \mathbb{E}_{s \sim Unif(P_{X\setminus\{c\}}(n))} [v(s \cup \{c\}) - v(s)]$$

Our choice of value function is the conditional expectation (CE) of the output given the subset as input: $v_{X,CE}(s) = \mathbb{E}[p(y|s)]$. Given this value function, a positive Shapley value for a cell means that the inclusion of the cell increases the predicted output on average over all possible input cell subsets. On the other hand, a negative Shapley value means that the cell decreases the predicted output on average over all possible input cell subsets.

In addition to the advantageous properties of Shapley values, single-cell FastShap provides non-encoding explanations only[50], due to the fact that the model sees only the selected cells without information about the selection mask. This means that there is no possibility that the explanation leaks information about the label beyond the information contained in the cells selected by the explanation, meaning that the predictive accuracy measures the quality of the explanation accurately.

Single-cell FastShap training

Since the Shapley value for a given cell is an average over all possible subsets, it is prohibitively expensive to compute directly for each cell for each patient. Instead, we train a Single-cell FastShap model to predict the outcome of this computation directly. At a high-level, training proceeds as follows (**Extended Figure 1**):

1. We train a surrogate model to predict the output y from a random subset of cells. This model has the same architecture of MIST;
2. We train a Single-cell FastShap model following the training objective of the original FastShap, which utilizes the prediction of the surrogate model to compute the FastShap loss[8]. The model architecture matches that of the MIST encoder.

Surrogate model objective function

The surrogate model f_E is trained to predict the output given randomly masked inputs via the following objective, following Jethani et al[51]:

$$f_E(y|x;\beta) = \max_{\beta} \sum_{i=1}^{n} \mathbb{E}_{\prod_{j=1}^{m} r_j \sim B} [\log p_\theta(y_i \,|mask(r_j, x_{i,j})_{j=1}^{m}],$$

where β are the parameters of the surrogate model and B is a distribution over input masks chosen to favor contiguous regions rather than uniformly sampled cells. In practice, for a given individual, we produce new sets of input distributions by selecting five ellipsoids across all available data types (see **Supplementary methods** – **Ellipsoids for data augmentation**). Then, the surrogate model is trained to predict the output of interest using as input only the cells that fall within the random ellipsoid regions. In other words, the training process for the surrogate model matches that of the full MIST model except that the surrogate model sees an augmented input dataset containing both the full 1000-cell distributions as well as arbitrary subsets of the single cells randomly sampled. Performance of the surrogate model on the full input is marginally worse than that of the full model but is mostly within a 95% confidence interval (**Table S8**), implying that the surrogate is a good approximation.

Single cell FastShap objective function

The optimization of the single-cell FastShap model involves four steps (Extended **Figure 1**): (1) for an individual i and their input single cell distributions, the FastShap model predicts single-cell Shapley values, one value for each cell; (2) a random mask k is sampled to produce a subset of cells to pass to the surrogate model to get a predicted output y_{SURR}^{k}; (3) the same random mask is applied to the single cell Shapley values which are then summed up to produce an estimate of the predicted output y_{SHAP}^{k}; (4) the loss is computed as the mean squared error between y_{SURR}^{k} and y_{SHAP}^{k}.

Formally, given a subset of cells s and a boolean mask m_s for the subset s, the FastShap model $\phi_{fastshap}$: $\mathbb{R}^{Nx2} \longrightarrow \mathbb{R}^{N}$ with parameters η is trained via the following objective:

$$\phi_{fastshap}(X;\eta) = \operatorname*{argmin}_{v_X(s)} \mathbb{E}_{p(X)} \mathbb{E}_{p(s)} (\, v_X(s) - \, m_s^T \phi(X;\, \eta) - \, v_X(\emptyset))^2,$$

where $p(X)$ is the distribution of inputs in the training data and $p(s)$ is the Shapley kernel[52] $p(s) \propto \frac{N-1}{\binom{d}{|s|}|s|(d-|s|)}$. This approach takes advantage of the weighted least squares characterization of the Shapley values[52] by directly optimizing for the model to predict the value of an input subset, $v_X(s)$, minus the value of the empty set as input, $v_X(\emptyset)$. We also ensure that the Shapley values sum up to the correct total for each individual by forcing the FastShap model's predictions to satisfy the efficiency constraint $1^T\phi(X;\, \eta) = \, v_X(X) - \, v_X(\emptyset)$ via additive efficient normalization[8,]. In practice, during inference, for each cell we add $\frac{1}{N}\left(v_X(X) - \, v_X(\emptyset) - \, 1^T\phi(X;\, \eta)\right)$ to the model's Shapley value prediction.

We consider two different training approaches for the single cell FastShap model: the first, default approach samples a subset S for each example from the Shapley kernel and computes the FastShap objective for each; the second, paired sampling approach samples a subset S for a given and then computes the FastShap objective for both S and its complement and averages the result. The latter has been shown to reduce gradient variance which can improve optimization efficacy[51]. We report the results from the paired sampling approach in the main paper and the results from default, nonpaired approach in the Supplement.

Compared to simply retraining a separate model for different ablations of entire distributions, this interpretability pipeline offers more fine-grained insights with a comparable amount of total computational effort (see **Table S7**).

To test the robustness of the FastShap interpretability procedure, we plot the inclusion plots and compare the interpretability maps for both sampling schemes (default and paired), noting that results look similar (**Figures S4-5**). We also see that the surrogate model achieves good performance relative to the original MIST predictor as well as a baseline (**Table S8**), and that compare the performance of the FastShap model to outperforms a baseline that assumes all cells for all individuals are equally important for prediction and find that the FastShap model is significantly better at matching the subset-based marker outputs (**Table S9**). For details see **Supplementary Methods**.

Interpretability maps

To visualize the single cell Shapley values at a population level, we randomly sub-sampled 3000 individuals for each of the predicted markers: 1000 falling below clinical range, 1000 falling within, and 1000 falling above (for clinical ranges see **Table S10**). We then computed the Shapley values for their input distributions and considered both their signed and absolute version We computed a 100 x 100 mesh for each individual and normalized them in an interval [-100, 100] for signed analysis and [0, 100] for absolute analysis (**Extended Figures 2-6**) to account for differences in population sizes across individuals. We then averaged the Shapley value for the cells falling within each mesh boundary and. The parts of the mesh that had more than 80% missing data (i.e., more than 80% of the individuals considered had no cells falling in that 2D space) were given a value of zero and not shown.

Down shift definition

Down shift was conceived through visual inspection of the interpretability map in **Figure 3** for the prediction of hematocrit, hemoglobin and red cell count. Given the split behavior on the left side of the BASOS distribution, we decided to isolate the left-most subpopulation of cells in the basophil distribution. To do this, we fit a mixture of gaussian with 2 components using full covariance matrices with 1e-6 regularization added to the diagonals to ensure matrices are positive semi-definite. We performed Expectation-Maximization on five separate random initializations and chose the best result based on likelihood. We collect the y-coordinate of these left centroids across the population of individuals. Individuals with a y-coordinate below the 25th percentile across the population were assigned a Down Shift. We evaluated the alternative of just considering the y-axis mean, eliminating the need for clustering, and results were consistent, we therefore proceeded with this approach as it is closer to biological differences of the cells (see **Supplementary Methods – Difference between Left Shift and Down Shift**)

Down Shift association analysis

We evaluated the clinical utility of Down Shift compared with Left Shift alone, as well as the combined presence of both Left Shift and Down Shift. Left Shift is returned from the hematological analyzer as 0 if absent or as 1,2,3 depending on the strength; in the analysis we considered it as a binary variable where any strength was considered as a Left Shift. We analyzed the association of these three exposures with two markers of inflammation, Erythrocyte Sedimentation Rate (ESR) and C-Reactive Protein (CRP), as well as future diagnosis. The analysis was performed with a logistic regression model adjusted by age, sex and white cell count. We collected all the diagnosis for the individuals analyzed from Electronic Health Records and mapped them to PheCodes[54], considering only those diagnoses that were within 30 days from the CBC measurement, were not diagnosed in the year prior to CBC and had at least 1% prevalence in the examined cohort for a total of 26 diagnosis. Significance was evaluated at an alpha of 0.05 with Bonferroni adjustment for multiple testing (p=0.0056 for ESR and CPR and p=0.0002 for diagnosis).

Acknowledgments

This work was supported by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, a DeepMind Fellowship, and NIH Awards R01HL148248, R01DK123330, and R01HD104756. The authors thank Siemens for instrument reagents and Fred Stelling and Val Jones for technical support with Advia data formats. The funders played no role in the analysis or the decision to publish.

Author Contributions

V.T, L.H.Z, R.R, and J.M.H conceptualized, designed, and conducted the study as well as interpreted the results. V.T, C.M., H.R.P performed data collection. V.T, L.H.Z developed the code and ran experiments. V.T, L.H.Z, R.R, and J.M.H wrote the first draft of the manuscript, and all authors edited, reviewed, and approved the final version of the manuscript.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Veronica Tozzo and John M. Higgins.

References

1. Foy, B. H. *et al.* Haematological setpoints are a stable and patient-specific deep phenotype. *Nature* **637**, 430–438 (2025).
2. Horton, S. *et al.* The Top 25 Laboratory Tests by Volume and Revenue in Five Different Countries. *Am. J. Clin. Pathol.* **151**, 446–451 (2019).
3. Harris, N., Kunicka, J. & Kratz, A. The ADVIA 2120 hematology system: flow cytometry-based analysis of blood and body fluids in the routine hematology laboratory. *Lab Hematol* **11**, 47–61 (2005).
4. Brugnara, C. Reticulocyte cellular indices: A new approach in the diagnosis of anemias and monitoring of erythropoietic function. *Crit. Rev. Clin. Lab. Sci.* **37**, 93–130 (2000).

5. Seebach, J. D., Morant, R., Rüegg, R., Seifert, B. & Fehr, J. The Diagnostic Value of the Neutrophil Left Shift in Predicting Inflammatory and Infectious Disease. *Am. J. Clin. Pathol.* **107**, 582–591 (1997).
6. Handtke, S. & Thiele, T. Large and small platelets—(When) do they differ? *J. Thromb. Haemost.* **18**, 1256–1267 (2020).
7. Zhang, L., Tozzo, V., Higgins, J. & Ranganath, R. Set Norm and Equivariant Skip Connections: Putting the Deep in Deep Sets. in *Proceedings of the 39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) vol. 162 26559–26574 (PMLR, 2022).
8. Jethani, N., Sudarshan, M., Covert, I. C., Lee, S.-I. & Ranganath, R. Fastshap: Real-time shapley value estimation. in (2021).
9. Sendak, M. P. *et al.* A Path for Translation of Machine Learning Products into Healthcare Delivery. *EMJ Innov.* **10**, 19–00172 (2020).
10. Javaid, M., Haleem, A., Pratap Singh, R., Suman, R. & Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. *Int. J. Intell. Netw.* **3**, 58–73 (2022).
11. Vis, J. Y. & Huisman, A. Verification and quality control of routine hematology analyzers. *Int. J. Lab. Hematol.* **38**, 100–109 (2016).
12. SHAPLEY L. S. A value for n-person games. *Contrib. Theory Games* **0**, 307–317 (1953).
13. Christoph Molnar. *Interpretable Machine Learning*. (2022).
14. Foy, B. H., Sundt, T. M., Carlson, J. C. T., Aguirre, A. D. & Higgins, J. M. Human acute inflammatory recovery is defined by co-regulatory dynamics of white blood cell and platelet populations. *Nat. Commun.* **13**, 4705 (2022).

15. Foy, B. H., Carlson, J. C. T., Aguirre, A. D. & Higgins, J. M. Platelet-white cell ratio is more strongly associated with mortality than other common risk ratios derived from complete blood counts. *Nat. Commun.* **16**, 1113 (2025).
16. Xavier-Ferrucio, J. *et al.* Low iron promotes megakaryocytic commitment of megakaryocytic-erythroid progenitors in humans and mice. *Blood* **134**, 1547–1557 (2019).
17. Little, J. D. C. A Proof for the Queuing Formula: L = λW. *Oper. Res.* **9**, 383–387 (1961).
18. Honda, T., Uehara, T., Matsumoto, G., Arai, S. & Sugano, M. Neutrophil left shift and white blood cell count as markers of bacterial infection. *Clin. Chim. Acta* **457**, 46–53 (2016).
19. Tishkowski K & Gupta V. Erythrocyte Sedimentation Rate. *Treasure Isl. FL StatPearls Publ.* (2024).
20. Nehring, S. M., Goyal, A. & Patel, B. C. C Reactive Protein. *Treasure Isl. FL StatPearls Publ.*
21. Chaudhury, A., Noiret, L. & Higgins, J. M. White blood cell population dynamics for risk stratification of acute coronary syndrome. *Proc. Natl. Acad. Sci.* **114**, 12344–12349 (2017).
22. Higgins, J. M. & Mahadevan, L. Physiological and pathological population dynamics of circulating human red blood cells. *Proc. Natl. Acad. Sci.* **107**, 20587–20592 (2010).
23. Chaudhury, A., Miller, G. D., Eichner, D. & Higgins, J. M. Single-cell modeling of routine clinical blood tests reveals transient dynamics of human response to blood loss. *eLife* (2019) doi:10.7554/eLife.48590.
24. Patel, H. H., Patel, H. R. & Higgins, J. M. Modulation of red blood cell population dynamics is a fundamental homeostatic response to disease. *Am. J. Hematol.* **90**, 422–428 (2015).
25. Golub, M. S., Hogrefe, C. E., Malka, R. & Higgins, J. M. Developmental plasticity of red blood cell homeostasis. *Am. J. Hematol.* **89**, 459–466 (2014).

26. Borghans, J. A. M., Tesselaar, K. & de Boer, R. J. Current best estimates for the average lifespans of mouse and human leukocytes: reviewing two decades of deuterium-labeling experiments. *Immunol. Rev.* **285**, 233–248 (2018).
27. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–48 (2021).
28. Straat, M., van Bruggen, R., de Korte, D. & Juffermans, N. P. Red Blood Cell Clearance in Inflammation. *Transfus. Med. Hemotherapy* **39**, 353–360 (2012).
29. Evstatiev, R. *et al.* Iron deficiency alters megakaryopoiesis and platelet phenotype independent of thrombopoietin. *Am. J. Hematol.* **89**, 524–529 (2014).
30. Ishimine, N. *et al.* Combination of White Blood Cell Count and Left Shift Level Real-Timely Reflects a Course of Bacterial Infection. *J. Clin. Lab. Anal.* **27**, 407–411 (2013).
31. Riley, L. K. & Rupert, J. Evaluation of Patients with Leukocytosis. *Am Fam Physician* **92**, 1004–1011 (2015).
32. Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. *Nat. Rev. Mol. Cell Biol.* **23**, 303–304 (2022).
33. Wagle, M. M., Long, S., Chen, C., Liu, C. & Yang, P. Interpretable deep learning in single-cell omics. *Bioinformatics* **40**, btae374 (2024).
34. Szałata, A. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
35. Montero, D., Diaz-Canestro, C., Oberholzer, L. & Lundby, C. The role of blood volume in cardiac dysfunction and reduced exercise tolerance in patients with diabetes. *Lancet Diabetes Endocrinol.* **7**, 807–816 (2019).

36. An, Y. *et al.* Blood flow characteristics of diabetic patients with complications detected by optical measurement. *Biomed. Eng. OnLine* **17**, 25 (2018).

37. Patel, K. P. Volume reflex in diabetes. *Cardiovasc. Res.* **34**, 81–90 (1997).

38. Safar, M., London, G., Weiss, Y. & Milliez, P. Altered blood volume regulation in sustained essential hypertension: a hemodynamic study. *Kidney Int* **8**, 42–47 (1975).

39. Weir, M. R. Hypervolemia and Blood Pressure. *Hypertension* **56**, 341–343 (2010).

40. Tarazi, R. C., Dustan, H. P., Frohlich, E. D., Gifford, R. W., Jr. & Hoffman, G. C. Plasma Volume and Chronic Hypertension: Relationship to Arterial Pressure Levels in Different Hypertensive Diseases. *Arch. Intern. Med.* **125**, 835–842 (1970).

41. Cullinane, E. M., Yurgalevitch, S. M., Saritelli, A. L., Herbert, P. N. & Thompson, P. D. Variations in plasma volume affect total and low-density lipoprotein cholesterol concentrations during the menstrual cycle. *Metab. - Clin. Exp.* **44**, 965–971 (1995).

42. Singh, A. *et al.* Hyperlipidemia and Platelet Parameters: Two Sides of the Same Coin. *Cureus* **14**, e25884 (2022).

43. Cohen, R. M. *et al.* Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA1c. *Blood* **112**, 4284–4291 (2008).

44. Chaudhury, A., Noiret, L. & Higgins, J. M. White blood cell population dynamics for risk stratification of acute coronary syndrome. *Proc. Natl. Acad. Sci.* **114**, 12344–12349 (2017).

45. Weiss Guenter & Goodnough Lawrence T. Anemia of Chronic Disease. *N. Engl. J. Med.* **352**, 1011–1023.

46. Patel, N. J., Tozzo, V., Higgins, J. M. & Stone, J. H. The Effects of Daily Prednisone and Tocilizumab on Hemoglobin A1c During the Treatment of Giant Cell Arteritis. *Arthritis Rheumatol.* **75**, 586–594 (2023).

47. Harris, N., Kunicka, J. & Kratz, A. The ADVIA 2120 hematology system: flow cytometry-based analysis of blood and body fluids in the routine hematology laboratory. *Lab Hematol* **11**, 47–61 (2005).

48. Tycko, D. H., Metz, M. H., Epstein, E. A. & Grinbaum, A. Flow-cytometric light scattering measurement of red blood cell volume and hemoglobin concentration. *Appl. Opt.* **24**, 1355–1365 (1985).

49. Vis, J. Y. & Huisman, A. Verification and quality control of routine hematology analyzers. *Int. J. Lab. Hematol.* **38**, 100–109 (2016).

50. Puli, A. M., Nguyen, N. & Ranganath, R. Explanations that reveal all through the definition of encoding. in *Advances in Neural Information Processing Systems* (eds. Globerson, A. et al.) vol. 37 99965–100006 (Curran Associates, Inc., 2024).

51. Jethani, N., Sudarshan, M., Aphinyanaphongs, Y. & Ranganath, R. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (eds. Banerjee, A. & Fukumizu, K.) vol. 130 1459–1467 (PMLR, 2021).

52. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).

53. Ruiz, L. M., Valenciano, F. & Zarzuelo, J. M. The Family of Least Square Values for Transferable Utility Games. *Games Econ. Behav.* **24**, 109–130 (1998).

54. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017).

55. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc. Natl. Acad. Sci.* **117**, 21373–21380 (2020).

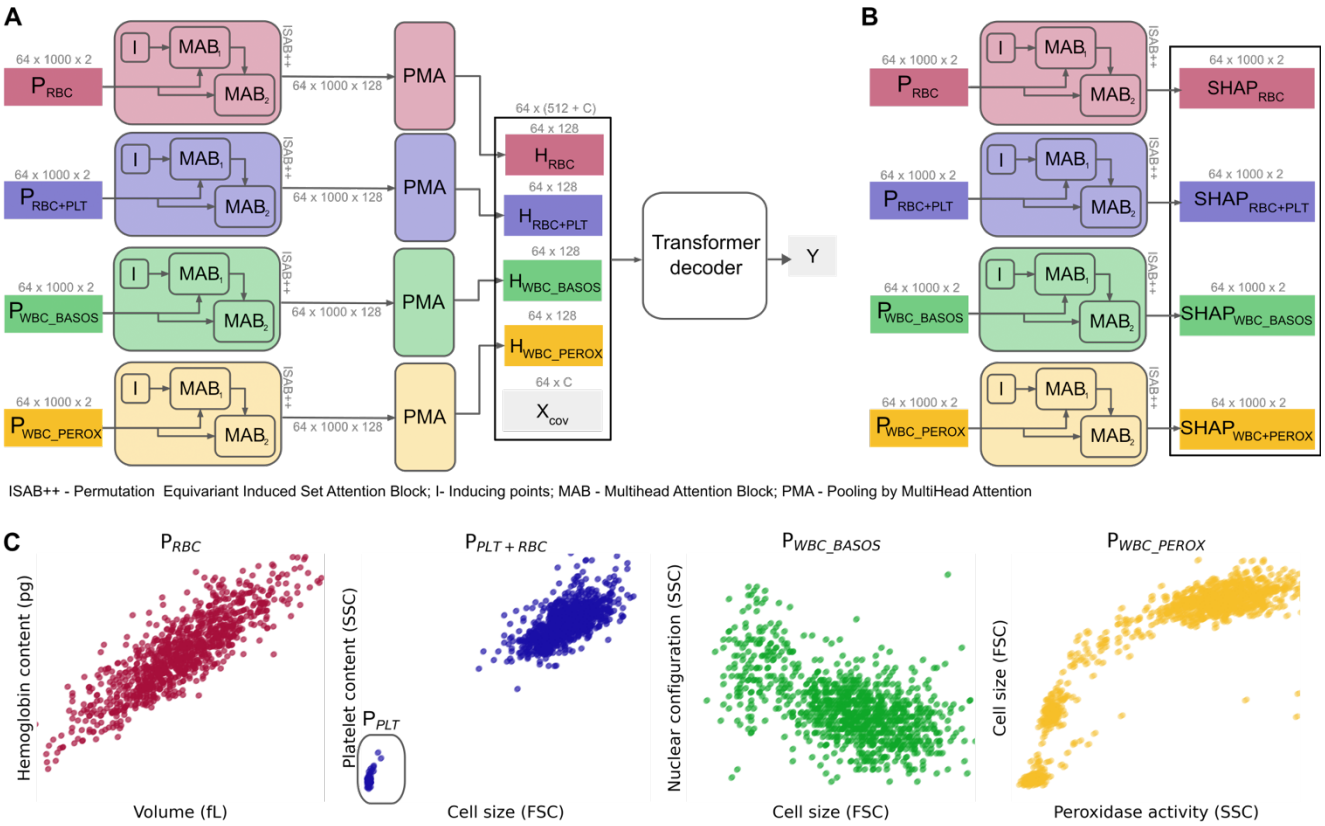


Figure 1**. Interpretable transformer-based models and single cell cytometric data.** Our interpretable transformer-based pipeline for single cell data is composed of Multi-Input Set Transformer++ (MIST) (panel A) a permutation invariant model that encodes each input single cell set independently to predict the output. This model is coupled with single-cell FastShap (panel B), a permutation equivariant model who provides single cell Shapley values for a given prediction task. We train a separate MIST and single-cell FastShap for each prediction task. We apply this pipeline on single cell cytometric data of blood cells (panel C) which measure red blood cells (RBC), platelets and red blood cells (PLT+RBC), and white blood cells in two different ways (BASOS and PEROX).

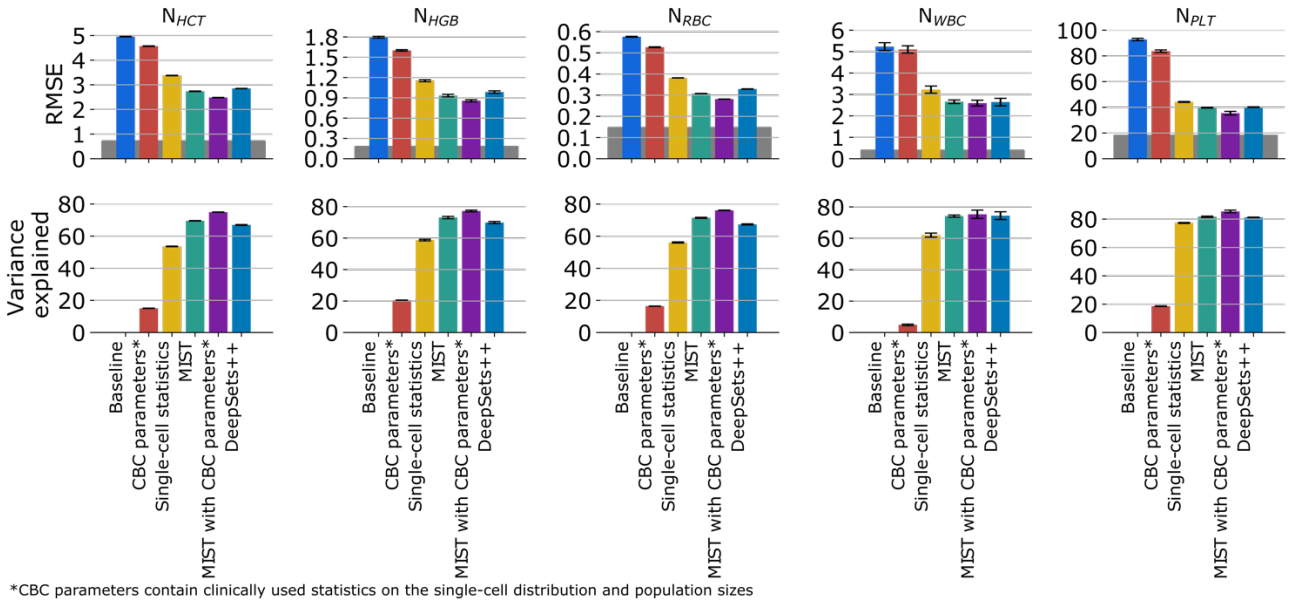


Figure 2. Prediction results of clinical markers from single-cell data. Error bars denote standard deviation across 3-fold cross validation with N=268326 training samples and N=134164 test samples in each split. Gray area denotes measurement error calculated as the mean of the value at the population level multiplied by the coefficient of variation (see Supplementary Table 2)

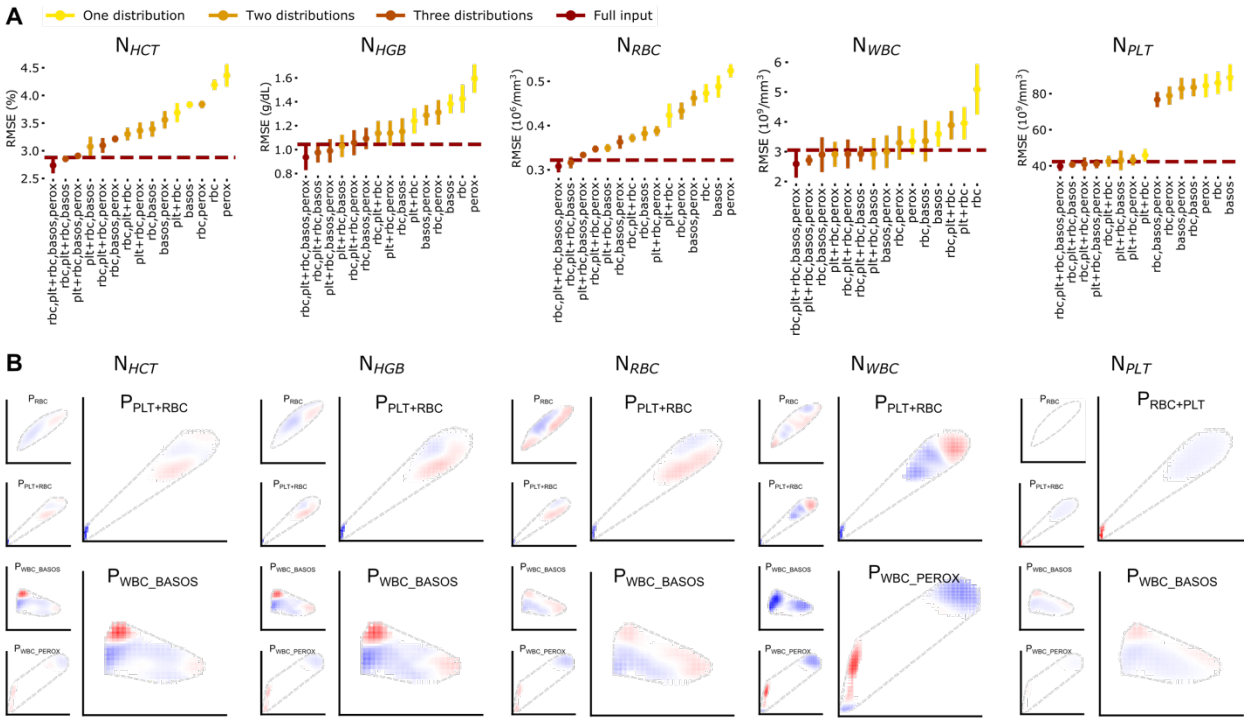


Figure 3. Interpretability analysis shows that the platelets distribution is co-regulated with all CBC indices and the basophil distribution is important for the prediction of red cell markers. (A) Full distribution ablations obtained by training a separate model on all possible permutations of the four distributions as input with n<4. (B) The Shapley values heatmaps for the prediction of the five CBC laboratory values (columns). A lighter color corresponds to a less important region for the prediction of the index, and a red region is positively correlated with the prediction while a blue region is negatively correlated with the prediction. The bigger heatmaps correspond to the two distributions that show a higher contribution to the prediction performance based on ablations in panel B. More detailed figures are available in **Extended Figures**. The dashed line represents the convex hull of the distributions of cells independent of their Shapley value.

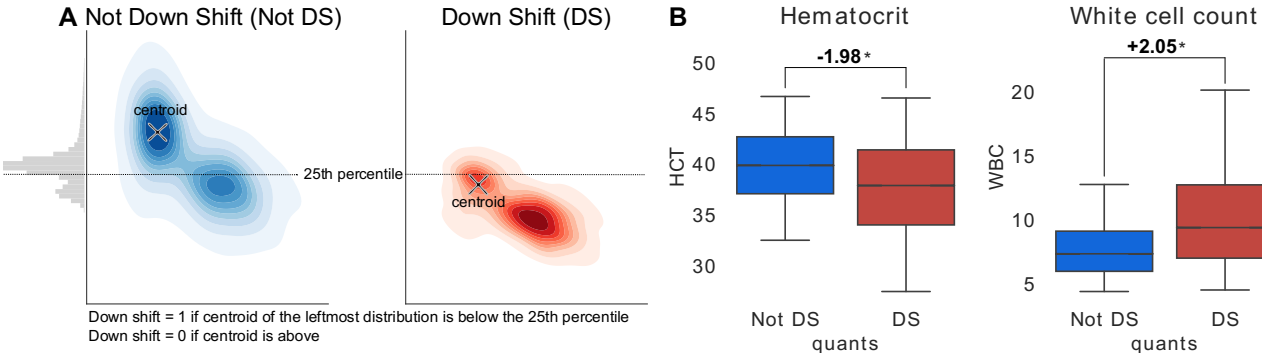


Figure 3. **Down Shift in basophil distribution is associated with lower hematocrit and higher white cell count.** Panel A is a cartoon of how the Down Shift (SD) marker, defined by the y-axis position of the leftmost cluster centroid in the basophil distribution. Panel B shows the distribution of hematocrit and white cell count for the individuals with and without the Down Shift marker. The star (*) denotes that the means are statistically significantly different using a two-sided independent t-test.

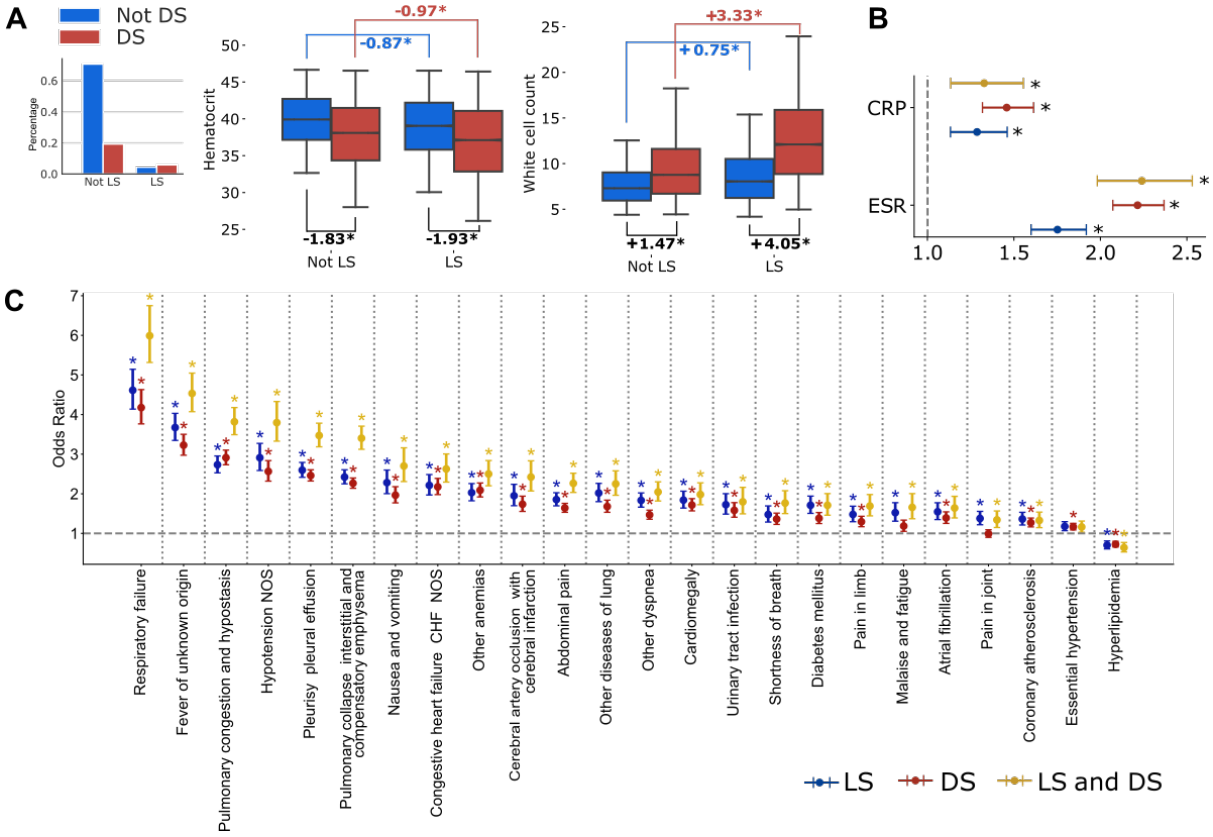
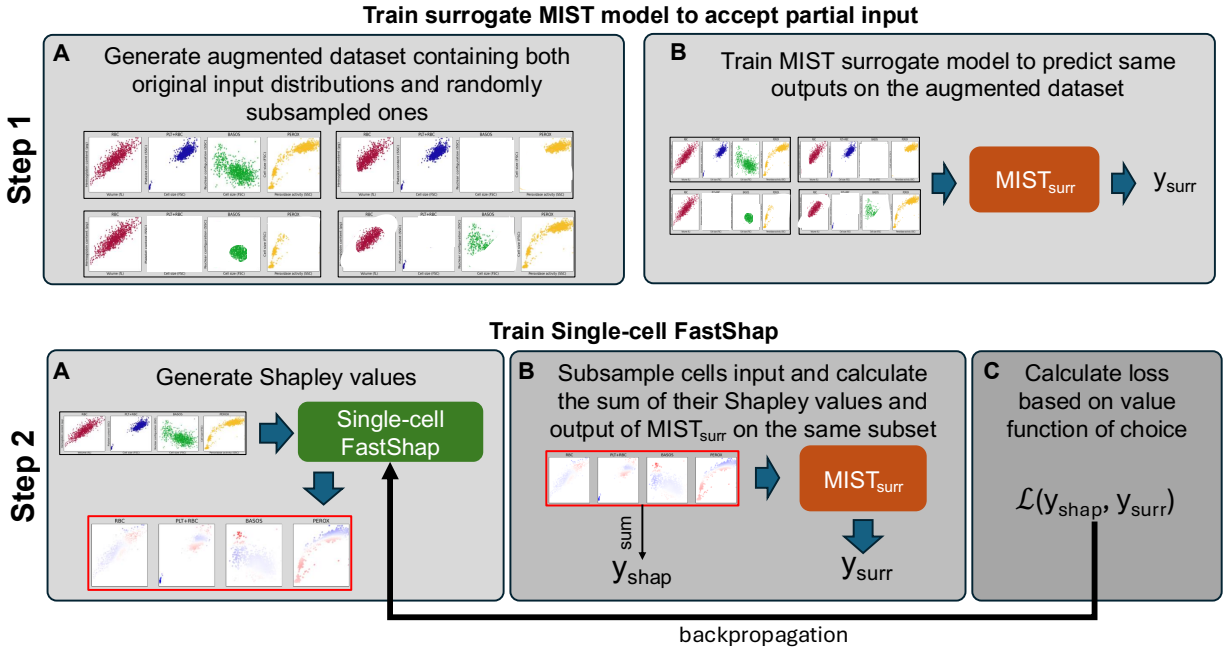
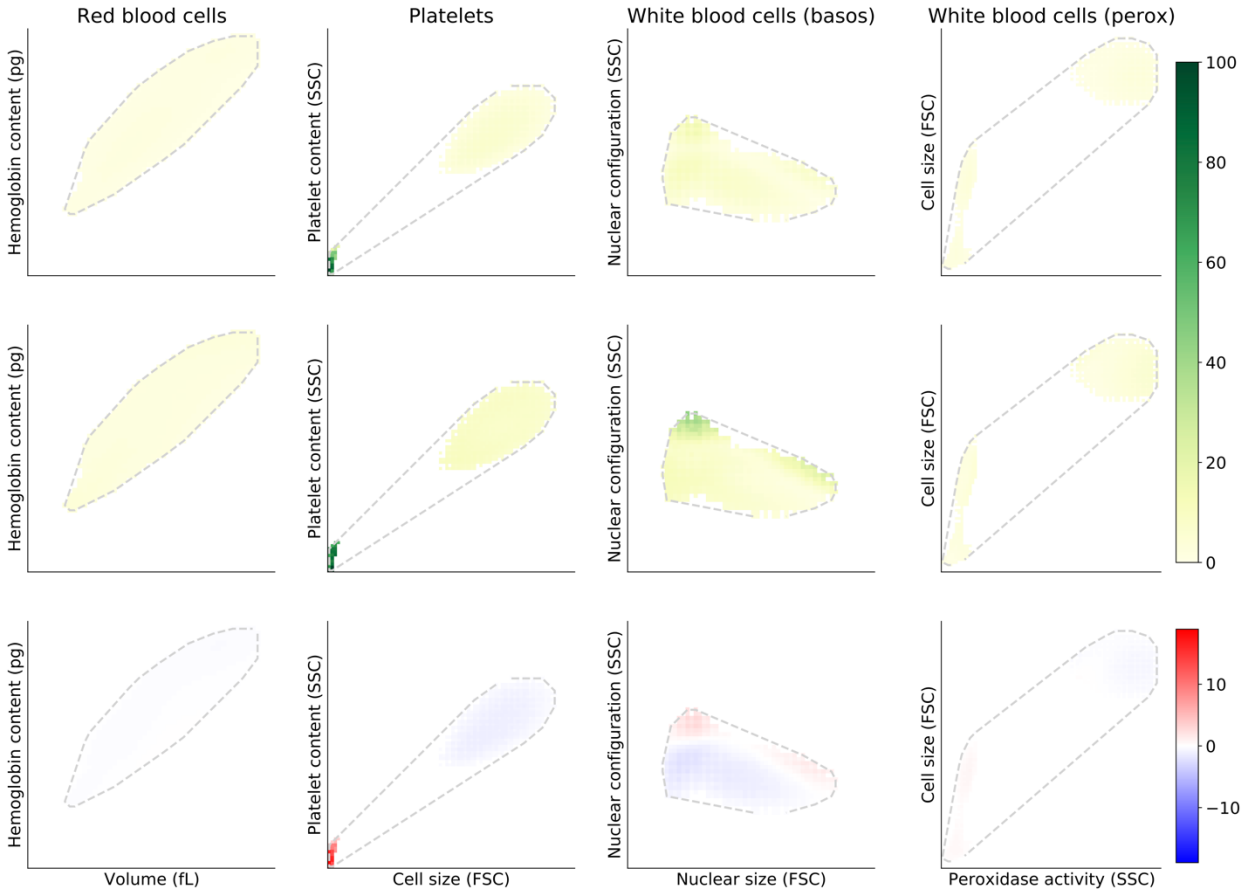


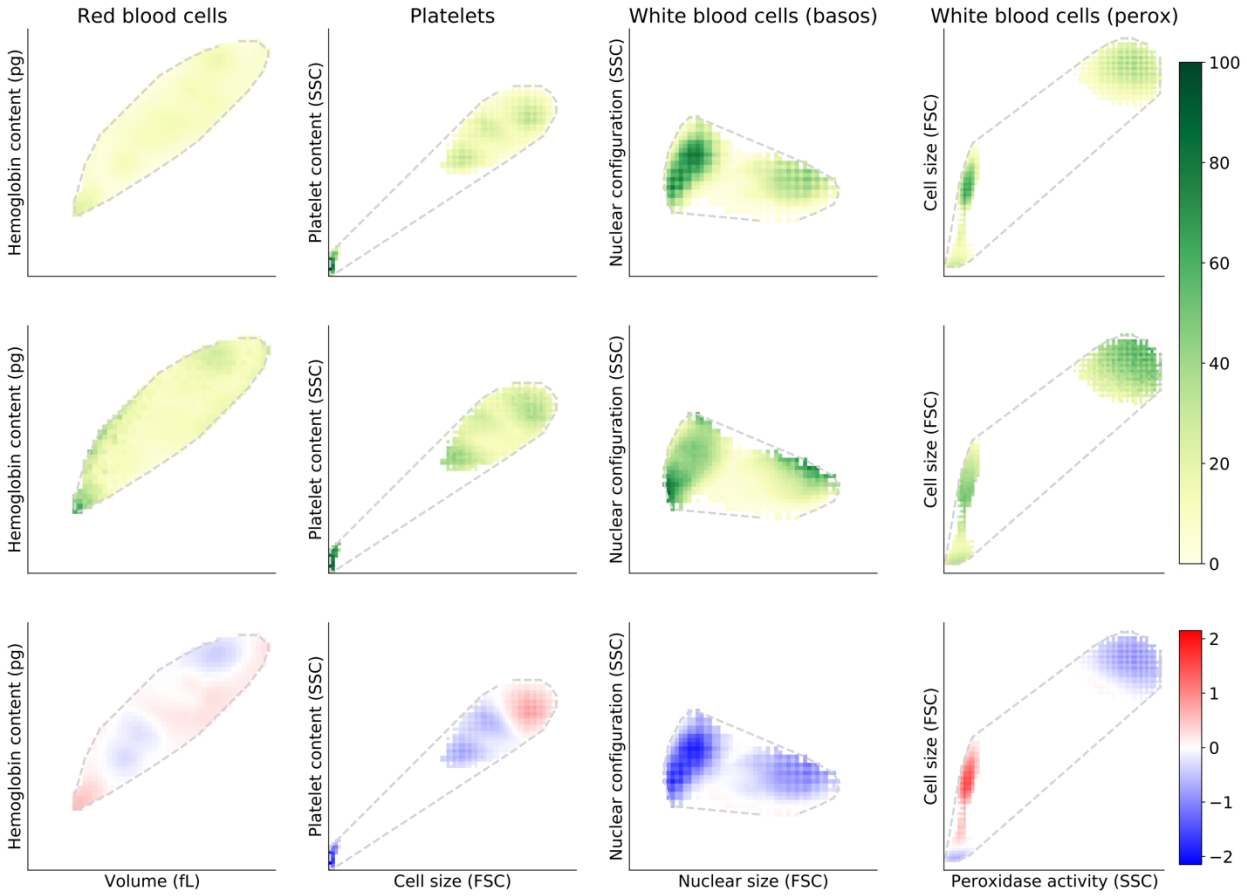
Figure 4. Down Shift (DS) complements Left Shift (LS) in the associations with diseases. Panel A shows the percentage of individuals that have Down shift (red bars) or not (blue bars) when they also have Left shift or not. It also shows the distributions of hematocrit and white cell count in the respective subgroups, showing the variation between Down Shift or not has a higher change in both hematocrit and white cell count compared to left shift. The star (*) denotes that the means are statistically significantly different using a two-sided independent t-test. Panel B shows the odds ratio of having an elevated C-reactive protein (CRP) test or an Erythrocyte Sedimentation Rate (ESR) test in presence of Left Shift (blue bars), down shift (red bars) or both simultaneously (mustard bars). Panel C shows the 10 highest odds ratios of having a diagnosis from the day of the CBC is measured up to a month after.



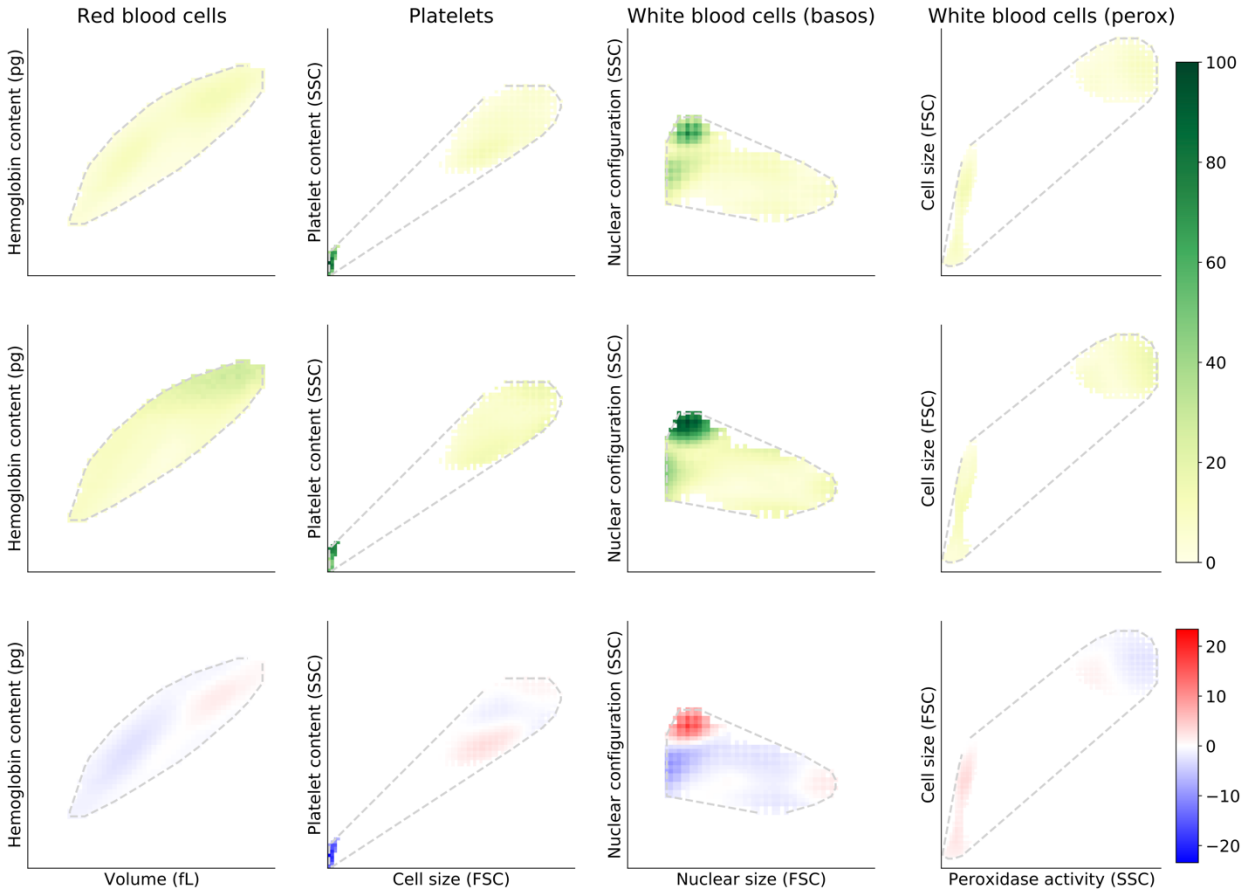
Extended Figure 1. Training pipeline for Single-cell FastShap. First, we train a surrogate MIST model able to take in input any subset of the input distributions. Second, we train the Single-cell FastShap by comparing for a random subsample of cells the sum of the Shapley values output by the model versus the output of the surrogate on the same subset. Then, we choose a value function and compute the loss, which we use to backpropagate through the single-cell FastShap model.



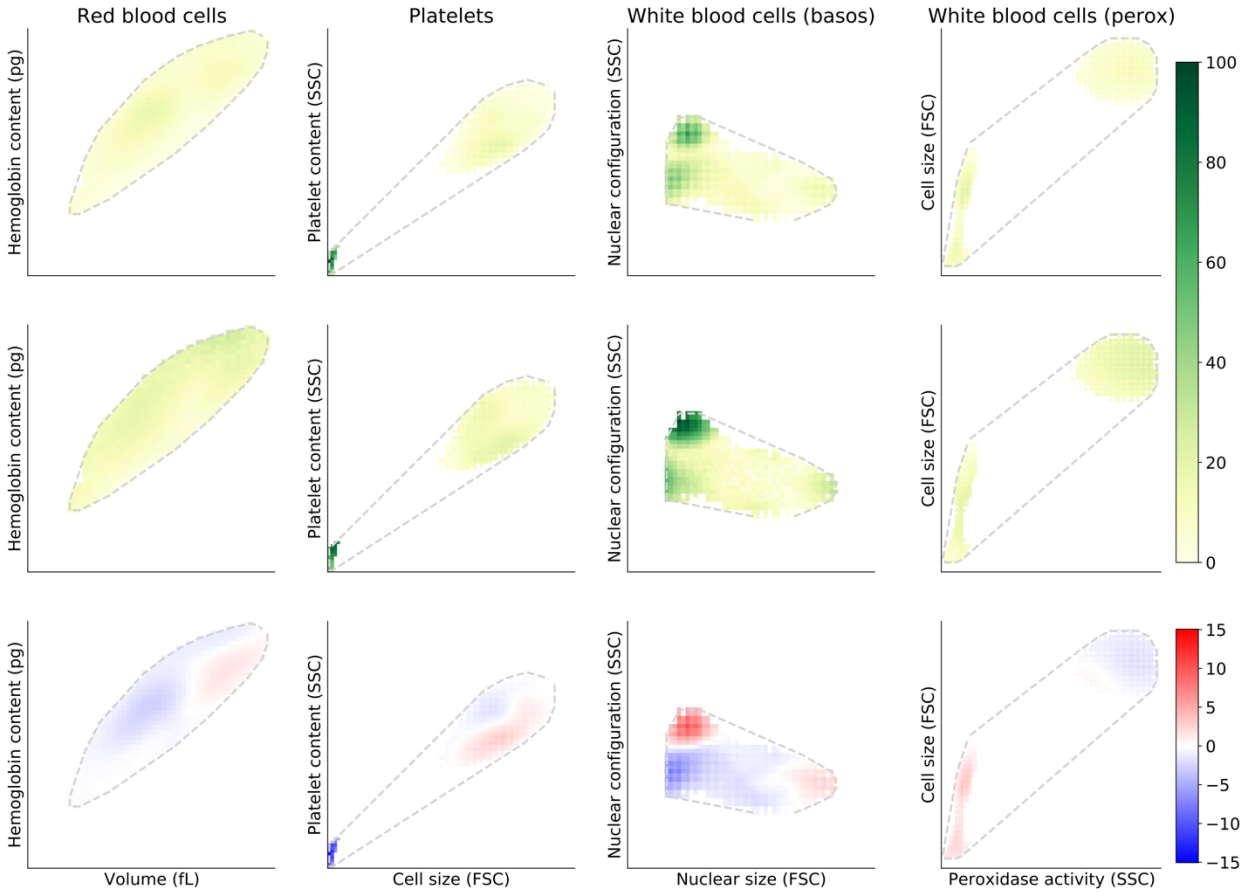
Extended Figure 2. **Interpretability maps for the prediction of platelet count.** On the first row we have the average absolute Shapley value across all distributions; in second row the standard deviation of the Shapley values; and in the last row the signed Shapley values. The platelet distribution (bottom-left corner of the PLT+RBC distribution) is the one with the highest Shapley values and higher variance, and it is positively correlated with platelet count (positive Shapley values).



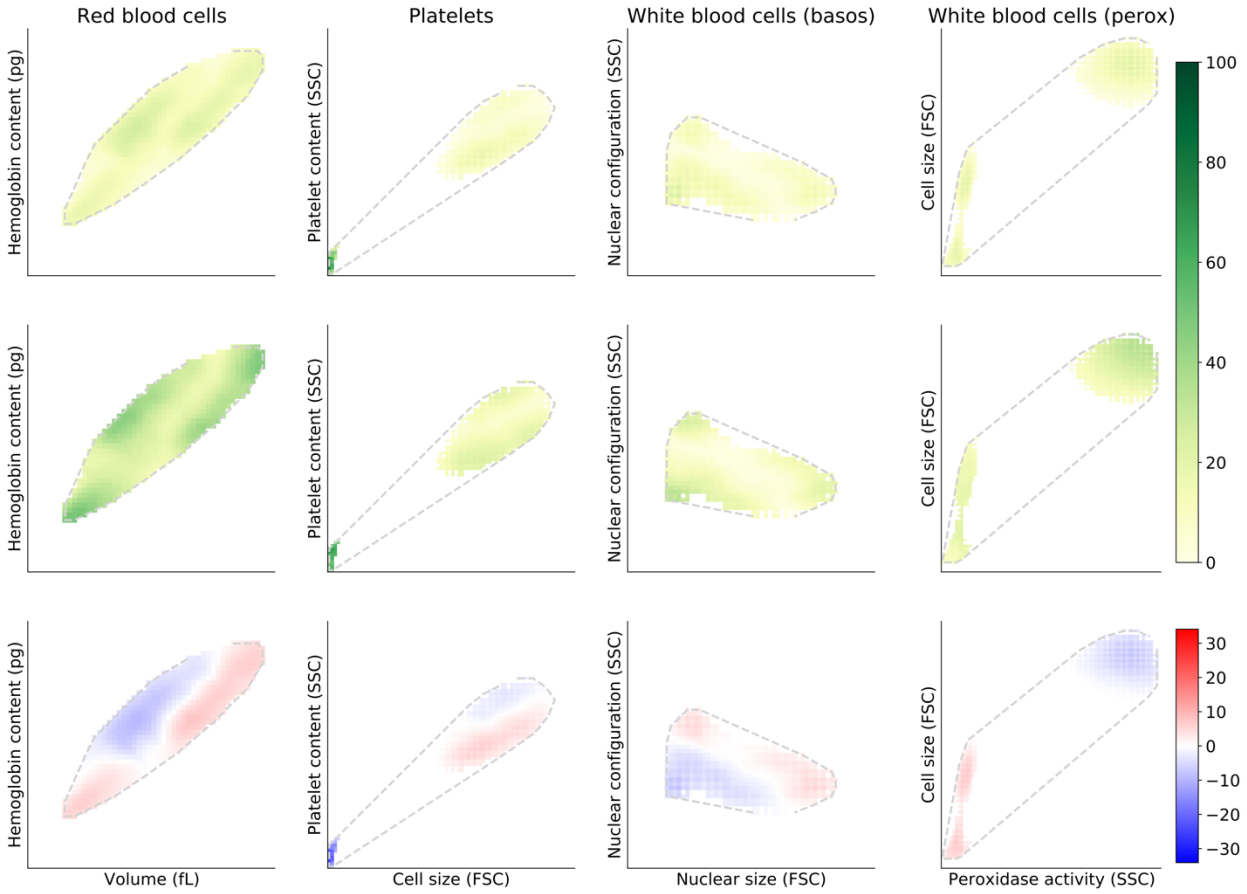
Extended Figure 3. **Interpretability maps for the prediction of white cell count.** On the first row we have the average absolute Shapley value across all distributions; in second row the standard deviation of the Shapley values; and in the last row the signed Shapley values. The prediction of white cell count is inversely correlated with the platelet distribution (bottom-left corner of the PLT+RBC distribution) and heavily relies on the BASO distribution as well as the PEROX distribution.



Extended Figure 4. **Interpretability maps for the prediction of hematocrit.** On the first row we have the average absolute Shapley value across all distributions; in second row the standard deviation of the Shapley values; and in the last row the signed Shapley values.



Extended Figure 5. Interpretability maps for the prediction of hemoglobin. On the first row we have the average absolute Shapley value across all distributions; in second row the standard deviation of the Shapley values; and in the last row the signed Shapley values.



Extended Figure 6. Interpretability maps for the prediction of red cell count. On the first row we have the average absolute Shapley value across all distributions; in second row the standard deviation of the Shapley values; and in the last row the signed Shapley values.