Mechanistic Deep Learning Framework on Cell Traits Derived from Single-Cell Mass Cytometry Data

Bo Wang[1], Sandra S. Zinkel[2,3], and Eric R. Gamazon[1,4,5,6]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA

[3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[4]Vanderbit Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[5]Data Science Institute, Vanderbilt University Medical Center, Nashville, TN

[6]Clare Hall, University of Cambridge, Cambridge, United Kingdom

Send correspondence to:

Eric R. Gamazon <eric.gamazon@vumc.org> or Sandra Zinkel <sandra.zinkel@vumc.org>

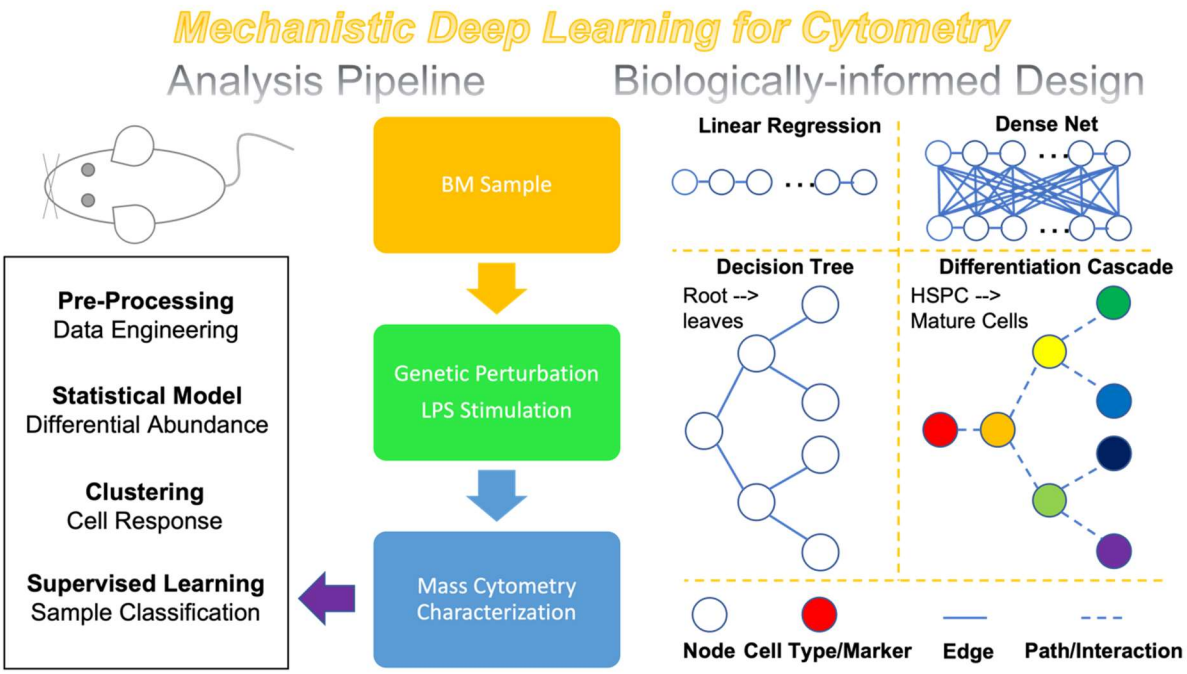
Abstract

Germline genetic variations can alter cellular differentiation, potentially impacting the response of immune cells to inflammatory challenges. Current variant- and gene- based methods in mouse and human models have established associations with disease phenotypes; however, the underlying mechanisms at the cellular level are less well-understood. Immunophenotyping by multi-parameter flow cytometry, and more recently mass cytometry, has allowed high-resolution identification and characterization of hematopoietic cells. The obtained characterization yields increased dimensionality; however, conventional analysis workflows have been inefficient, incomplete, or unreliable. In this work, we develop a comprehensive machine learning framework – MDL4Cyto – that is tailored to the analysis of mass cytometry data, incorporating statistical, unsupervised learning, and supervised learning models. The statistical modeling can be used to illuminate cell fate decision and cell-type dynamics. The unsupervised learning models along with complementary marker enrichment analyses highlight genetic perturbations that are significantly associated with alterations in cell populations in the hematopoietic system. Furthermore, our supervised learning models, including deep learning and tree-based algorithms, address the bottleneck to data pre-processing that characterizes conventional workflows and generate inferences (e.g., on marker/cell-type interactions) from raw experimental characterization. Notably, we reveal a close relationship among network design, prediction performance, and the underlying biological context. We show that the network architecture extracted from the differentiation cascade of the investigated biological system yields enhanced prediction performance. The presented methodology will enable new insights into hematopoietic differentiation at baseline and following perturbation.

Highlights

- Analysis pipeline on mass cytometry data with high-performance implementation of statistical, unsupervised learning, and supervised learning models
- Concordance of machine learning results with biological contexts
- Biologically-informed neural network designs enhance prediction performance

Graphical Abstract



Keywords

Machine Learning, Mass Cytometry, CyTOF, Hematopoiesis, Cellular Differentiation, Regulatory Cascade, Clonal Hematopoiesis of Indeterminate Potential

Introduction

Hematopoiesis is a hierarchical and dynamic process. Proper blood cell development requires highly coordinated and tightly regulated cytokine signals, affecting cell and stage-appropriate gene expression. Genetically modified mouse models are an invaluable model system for studying hematopoiesis, due to their defined genetic background, wealth of reagents to precisely define hematopoietic cell populations, ability to undergo precise genetic modification, and the wealth of data and assays arising from decades of study. Cytometry tools have proved to be essential in simultaneously analyzing labeled molecules at the single-cell level and, consequently, have been widely implemented in basic and clinical research[1]. While traditional fluorescence cytometers can detect up to 35 markers, the technique is limited by spectral overlap of fluorophores. Single-cell mass cytometry (Cytometry by Time-Of-Flight [CyTOF]) enables multiple-parameter tracking by the use of metal-isotope-tagged antibodies, overcoming the inherent limitations of spectral overlap. Therefore, a more comprehensive panel of cellular functions (achieved from significantly expanded feature capacity[2]) along with intracellular signaling states can be explored to identify disease-associated changes or functionally-relevant states; differential expression under external perturbation or intracellular response is more easily elucidated[3].

Insights from cytometry characterization are traditionally extracted from manual gating. Automated workflows such as FlowCore[4], PhenoGraph[5] or the recently reported Tracking Responders EXpanding (T-REX)[6] have been introduced with significantly enhanced processing efficiency. However, these pipelines are still limited by expert intervention into (manual) data pre-processing, where key thresholds for transformation, gating or cell type clustering are tailored to the investigated systems. Recently, machine learning algorithms have attracted significant attention for their extensive application to the analysis of cytometry data[6–8]. Specifically, beyond functioning as embedded algorithms in pipelines, machine learning tools provide “end-to-end” solutions, generating inferences or predictions

78 directly from raw experimental characterization, with potential to uncover novel associations that are
79 obscured by conventional gating strategies.

80 Hematopoiesis has served as a paradigm of stem cell biology and cellular differentiation[9].
81 Genetic perturbations within cells (in particular, Hematopoietic Stem and Progenitor Cells [HSPCs])
82 induce challenges to cellular functions, thus leading to various diseases or conditions, e.g., Clonal
83 Hematopoiesis of Indeterminate Potential (CHIP), Myelodysplastic Syndromes (MDS), and Acute Myeloid
84 Leukemia (AML). Tet2 inactivating mutations are among the early mutations that initiate clonal
85 expansion and progression to MDS or AML. As the predominant Tet gene in the hematopoietic tissue
86 (especially in HSCs), Tet2 functions as a tumor suppressor whose haplo-insufficiency initiates myeloid
87 and lymphoid transformations[8]. Tet2 has been shown to regulate resolution of inflammatory responses.
88 Tet2 KO mice and cell lines display prolonged and enhanced inflammatory cytokine signaling, which
89 impacts HSPC differentiation.

90 Hematopoietic homeostasis is further controlled through tight regulation of programmed cell
91 death. The two major types of cell death, apoptosis (programmed cell death) and necrosis (regulated
92 necrotic cell death), maintain homeostasis by downregulating detrimental inflammation and prevent
93 transformation by allowing efficient removal of damaged cells. It has been widely recognized that the
94 genes in the Bcl-2 family are critical in regulating apoptosis[10,11], comprised of both pro-survival and pro-
95 death members. The functions of BID (pro-apoptotic BH3-only member) were investigated in our recent
96 human phenome-wide association study (PheWAS)[12,13]. In follow-up studies, we evaluated the
97 multidomain effectors Bax, Bak, and Bid in murine models[14]. We found that deletion of Bax/Bak inhibits
98 both bone marrow apoptosis and bone marrow necroptosis. Further deletion of Bid (TKO) leads to
99 robust activation of necroptosis, which results in a phenotype characterized by dysplasia and cytopenia,
100 resembling the bone marrow failure disorder MDS[14].

101 To address the limitations of conventional cytometry data analysis pipelines[5–8,15], we developed
102 a comprehensive *Mechanistic Deep Learning for Cytometry* (MDL4Cyto) framework, implementing
103 statistical, unsupervised learning, and supervised learning models. The framework was evaluated against
104 bone marrow samples of murine models that have been genetically perturbed in the hematopoietic
105 system, including Wild Type (WT), *VavCreBaxBakBid* triple-knockout (TKO), and *VavCreTet2* knockout
106 (Tet2 KO) mice, and validated with biological observations. Notably, building on our previous work[16,17],
107 we designed the architectures of the neural networks using the intrinsic hierarchy of the hematopoietic
108 system as a guide, establishing a *biological feature – network architecture – phenotype* relationship. We
109 show that biologically-informed network designs could function as probes on the regulatory cascade
110 with prediction performance as the evaluation metric. Our framework may inform future machine
111 learning implementations to address fundamental questions in cell biology.

112

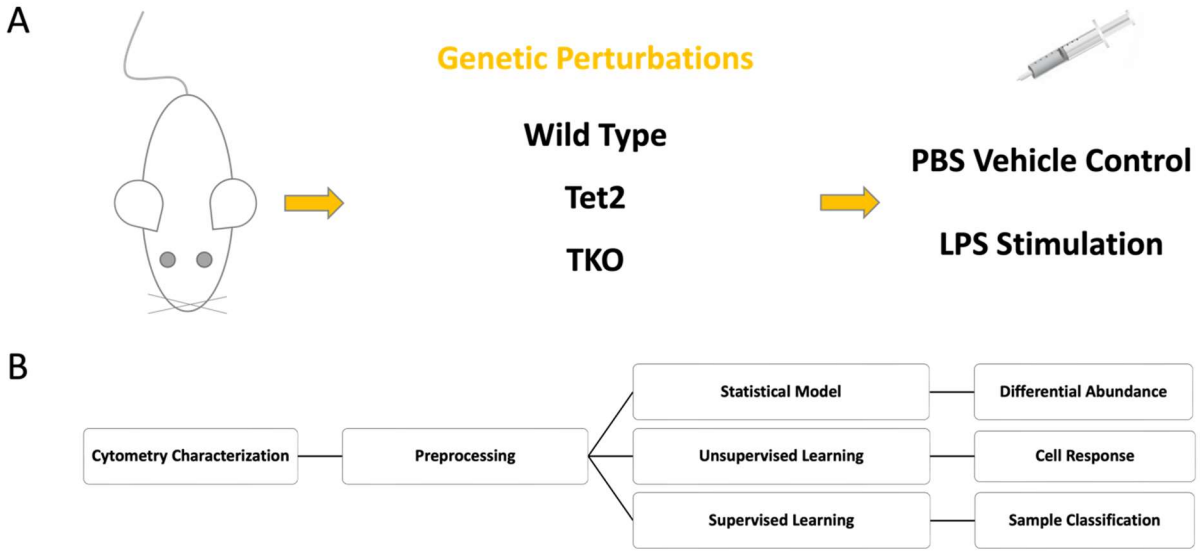
113 ## Results

114 **Overview of mass cytometry data and MDL4Cyto framework**

115 Mice from three genetic murine models – WT, Tet2 and TKO – were treated with Phosphate-
116 buffered saline (PBS) or lipopolysaccharide (LPS). LPS is a major surface membrane polysaccharide found
117 on Gram-negative bacteria and is a typical pathogen-associated molecular pattern (PAMP).
118 Inflammation triggered by LPS stimulation activates a robust innate immune response and can induce
119 the proliferation of hematopoietic progenitor cells[18] and the production of cytokines such as TNFα,
120 interleukin and IL-2. Bone marrow from these mice was subsequently harvested and processed for
121 cytometry experiments. Besides genetic perturbations, all the samples challenged by LPS stimulation
122 were compared to the PBS vehicle control (Figure 1A). Across different types of samples (WT-PBS, WT-
123 LPS, Tet2-PBS, Tet2-LPS, TKO-PBS, and TKO-LPS), lineage marker groups (stem cells, pan leukocytes, T

124 lymphocytes, mature myeloid cells, B lymphocytes, and erythrocytes) and functional marker groups
125 (cytokines, cell death, and proliferation), were tagged and monitored (Table S1). As sample data profiles
126 obtained from the mass cytometry measurements varied in length, we imputed data points at each of
127 the six sample data profiles via binning methods with 30 seconds.

128 We developed MDL4Cyto, an analysis pipeline for mass cytometry data (Figure 1B and
129 **Methods**), implementing statistical, unsupervised, and supervised learning models. Statistical tests on
130 the cell count of key lineage and functional markers were performed, aiming for a glimpse of the
131 underlying biology. As both genetic perturbations and LPS stimulation may induce inflammatory changes
132 in hematopoiesis, we applied Generalized Mixed-Effects Model (GMM) estimation to determine the
133 extent to which cell count was statistically altered by these modifications. Two established workflows,
134 namely PhenoGraph[5] and T-REX[6], were benchmarked as established pipelines capable of annotating cell
135 populations or capturing cell responses. In addition, supervised learning models were explicitly designed
136 to probe marker/cell-type interactions within the regulatory cascade, as well as to predict sample type
137 directly from the raw mass cytometry characterization. Nine machine learning models were evaluated,
138 including Linear Regression (LR), Dense Neural Network (DNN), Boosted Tree (BT), Random Forest (RF),
139 stacked Dense Neural Network - Random Forest (DNN-RF), Convolutional Neural Network in One
140 Dimension (CNN-1D), Convolutional Neural Network in Two Dimensions (CNN-2D), Recurrent Neural
141 Network with Long Short-Term Memory cell (RNN-LSTM), and Recurrent Neural Network with Gated
142 Recurrent Units cell (RNN-GRU).



143
144 **Figure 1. MDL4Cyto pipeline for analysis of single-cell mass cytometry data.** Bone marrow samples (**A**)
145 were harvested from genetically perturbed murine models (Wild Type [WT], *VavCreBaxBakBid* triple-
146 knockout [TKO], and *VavCreTet2* knockout [Tet2 KO]), which had been treated with LPS stimulation or
147 PBS vehicle control. Statistical, unsupervised, and supervised learning models were implemented in our
148 *Mechanistic Deep Learning for Cytometry* (MDL4Cyto) framework. The analysis pipeline **(B)** was
149 benchmarked, using the reference dataset obtained from mass cytometry characterization, for the tasks
150 of differential abundance analysis, cell response detection, and sample type classification.

151

152

153 **Statistical modeling of genetically-determined or condition-dependent effects**

154 The cell count summary statistic, which was calculated on the reference dataset, was found to
155 be altered by genetic perturbations and/or LPS stimulation (Figure S1). Clear differences were observed
156 for all markers. As reported in our previous work[14] and other studies[19], the TKO sample leads to necrosis
157 with increasing production of cytokines. Enhanced inflammatory signaling was also observed in Tet2-
158 deficient samples, as has also been previously described[19]. We observed a statistically significant
159 difference in cell count for Sca1 (LSK cells), CD45 (leukocytes), TNFα (cytokines), and Caspase 3 (cell

160 death indicator) with respect to genetic perturbation (Figure 2A). Markers of cytokines were found to be

161 highly correlated with other surface markers, especially under LPS stimulation (Figure 2B).

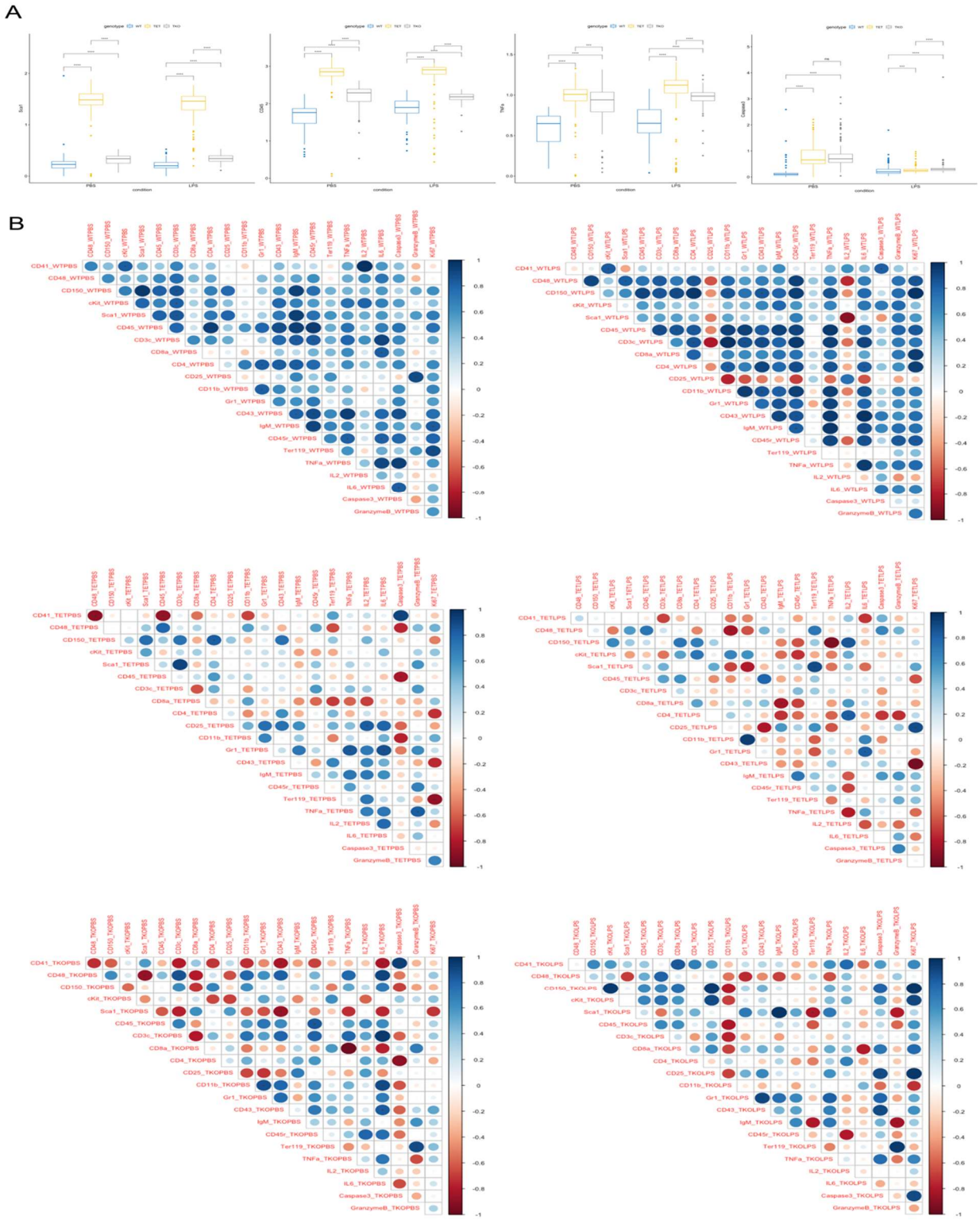
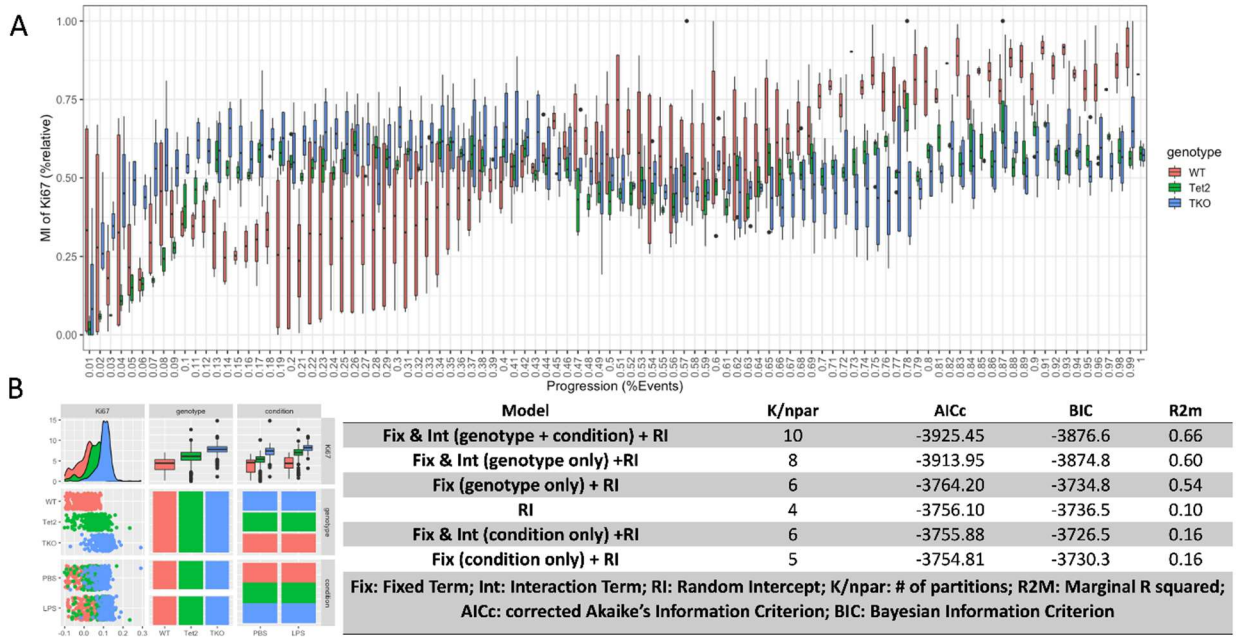


Figure 2. Impact of genetic perturbations on cell populations and correlations among the markers. A statistically significant change in cell count was observed for Sca1 (LSK cells), CD45 (leukocytes), TNFα

165 (cytokines), and Caspase 3 (cell death indicator) in both the PBS vehicle control and LPS stimulation
166 condition, demonstrating genetic perturbations altering cell populations **(A)**. Cytokines from genetic
167 knockout samples were highly correlated with other surface markers, in particular under LPS stimulation
168 **(B)**, which is consistent with the biological context. (Wilcoxon test was performed for the pairwise
169 comparisons; Pearson correlation was calculated for the markers along with 95% confidence interval.
170 *p<.05, **p<0.01, ***p<0.001, ****p<0.0001).

171

172 The data profiles obtained from cytometry characterizations are dynamical panels, i.e., have a
173 time component. To estimate the overall effects of genotype and condition, we analyzed the
174 progression of marker Intensity for Ki67 from the aggregated data panel with 1 hour binning window
175 (Figure 3 and **Methods**). Ki67 is a proliferation protein that is increased in the S phase of the cell cycle
176 and indicates proliferating cells. We observed progression divergence according to genotype (Figure 3A)
177 and differentiated marker intensity between conditions (PBS vs LPS, Figure 3B). This is consistent with
178 our understanding that genetic perturbations and LPS stimulation impose pressures on differentiation
179 and cell cycle as dynamical processes. To achieve a better quantitative estimate of the contributions
180 arising from genotype and condition, we implemented GMM estimation. Besides the fixed and/or
181 interaction terms (**Methods**), the time term was considered in the model (Figure 3C). We applied
182 corrected Akaike's Information Criterion (AICc) and Bayesian Information Criterion (BIC) for model
183 evaluation and Marginal R squared (R2m) for variance coverage. Models built from genotype
184 demonstrated the best performance, which corroborates the fact that genetic perturbations have
185 significant impact on cellular differentiation and function. We note that the statistical modeling
186 implemented in MDL4Cyto (Figure 1B and **Methods**) is broadly applicable, which can facilitate the study
187 of cell fate decision and cell-type dynamics.



Model	K/npar	AICc	BIC	R2m
Fix & Int (genotype + condition) + RI	10	-3925.45	-3876.6	0.66
Fix & Int (genotype only) +RI	8	-3913.95	-3874.8	0.60
Fix (genotype only) + RI	6	-3764.20	-3734.8	0.54
RI	4	-3756.10	-3736.5	0.10
Fix & Int (condition only) +RI	6	-3755.88	-3726.5	0.16
Fix (condition only) + RI	5	-3754.81	-3730.3	0.16

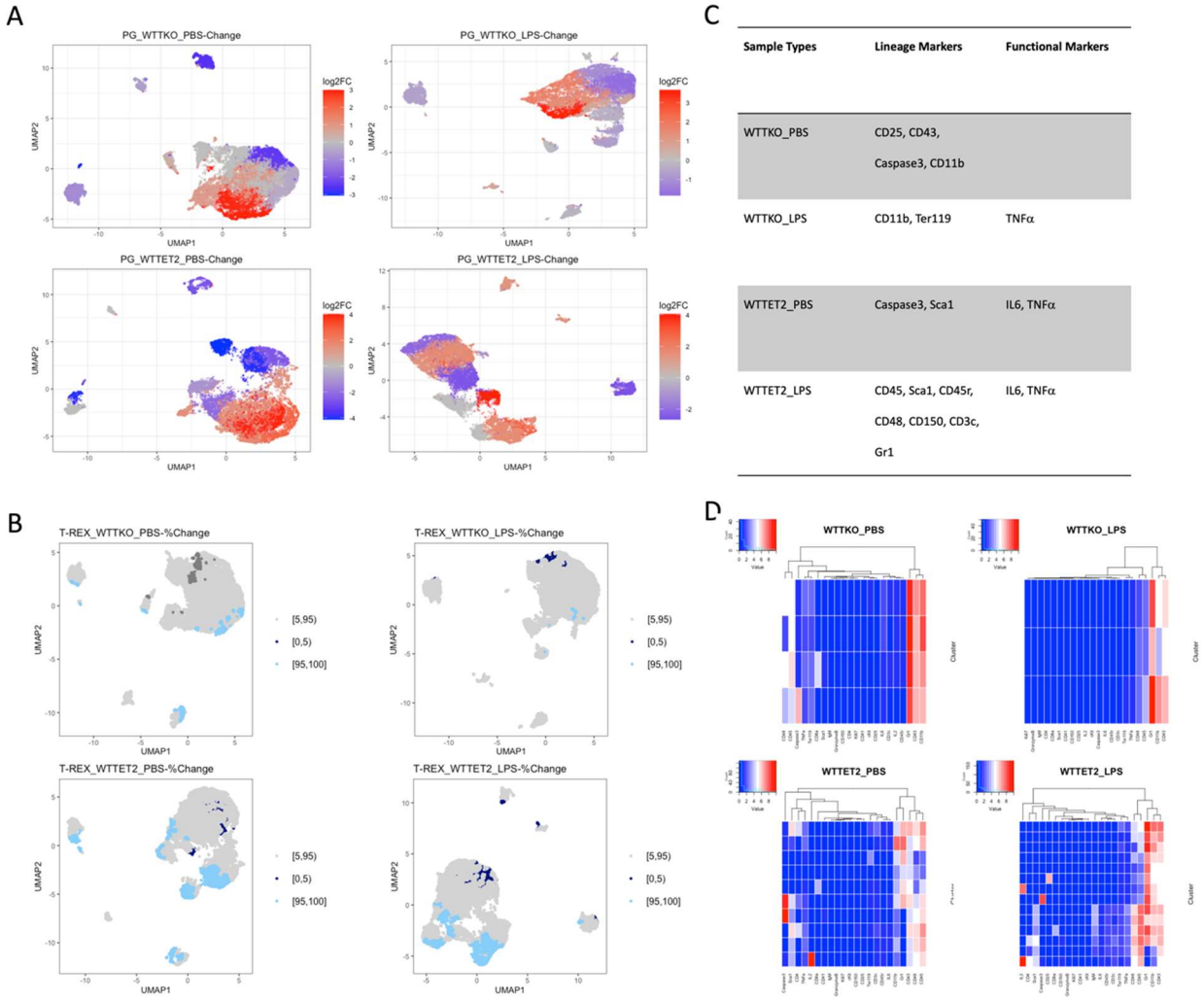
Fix: Fixed Term; Int: Interaction Term; RI: Random Intercept; K/npar: # of partitions; R2M: Marginal R squared; AICc: corrected Akaike's Information Criterion; BIC: Bayesian Information Criterion

Figure 3. Analysis of dynamic data panel. Progression (%Events over the course of time) **(A)** and histogram **(B)** of cell count of proliferation marker Ki67 diverged with genetic background and shifted with condition. We implemented Generalized Mixed-Effects Model (GMM) estimation that considered the time term in addition to a fixed term and interaction term. While using genotype and condition as main parameters, several models were constructed, as represented by the equations in **(C)**. The performance of each model was further evaluated using the corrected Akaike's Information Criterion (AICc), Bayesian Information Criterion (BIC), and Marginal R squared (R2m). Genotype was identified as an important feature.

Quantifying cell response and population alteration via unsupervised learning

Identifying cell response from cell populations has been a key interest in mass cytometry analysis. Hematopoietic responses, induced either by genetic perturbations or by pathogen-like stimulation (LPS), were analyzed via pairwise comparisons, including WT-PBS vs TKO-PBS, WT-LPS vs TKO-LPS, WT-PBS vs Tet2-PBS, WT-LPS vs Tet2-LPS. Two established unsupervised learning algorithms, T-REX and PhenoGraph, were benchmarked in clustering our reference dataset (see Table S3 for the hyperparameters). Both algorithms apply k-nearest neighbors (kNN) for clustering and Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and visualization. As illustrated in the UMAP plot, genetic perturbations induced alterations in cell populations (Figure 4A&4B). Notably,

207 differential cell count of markers under a genetic perturbation, including Sca1, CD45, TNFα, and Caspase
208 3, could be seen in the mapping of cell count to the clusters (Figure S2, T-REX algorithm). These results
209 are consistent with the comparisons from our statistical tests (Figure 2A). Compared to conventional
210 methods, unsupervised learning showed a unique advantage in efficiently capturing changes in cell
211 populations. In our framework, cell subpopulations within the clusters were further investigated in two
212 ways. While reference datasets trained here contained only a single sample for each type, we created
213 two pseudo-replicates from a single sample (**Methods**) for the PhenoGraph algorithm, thus enabling
214 differential abundance analysis for the markers within each cluster (Figure 4C). In addition, cell
215 populations from the T-REX algorithm were screened and enriched through Density-Based Spatial
216 Clustering and Application with Noise (DBSCAN) and Marker Enrichment Modeling (MEM)[20] (Figure 4D).
217 Strikingly, in contrast to the standard T-REX method, our framework uncovered a strong signal from the
218 cytokines, i.e., IL6 and TNFα, from the differential abundance analysis (Figure 4C). This observation could
219 well be attributed to the known LPS-driven inflammation, as demonstrated by the enrichment analysis
220 from MEM (Figure 4D). Collectively, the TNFα and IL6 differential abundance results are consistent with
221 myeloid-driven or innate inflammation.



Sample Types	Lineage Markers	Functional Markers
WTTKO_PBS	CD25, CD43, Caspase3, CD11b	
WTTKO_LPS	CD11b, Ter119	TNFα
WTTET2_PBS	Caspase3, Sca1	IL6, TNFα
WTTET2_LPS	CD45, Sca1, CD45r, CD48, CD150, CD3c, Gr1	IL6, TNFα

Figure 4. Cell response identified by the unsupervised learning. Clustering from Phenograph **(A)** and T-REX **(B)** algorithms was performed on the paired genetically perturbed samples. PBS vehicle control and LPS stimulation were separately explored and compared. Using both algorithms, we found that the genetic perturbations resulted in altered cell populations. In our framework, clusters of significant fold changes identified from PhenoGraph were further explored for differential abundance of markers. Markers with significant responses that were captured in more than two clusters were identified **(C).** In addition, cell populations obtained from T-REX were screened for those with significant changes and enriched through DBSCAN and MEM **(D)**.

Probing the regulatory cascade with biologically-informed neural networks

The hematopoietic system has classically been represented as a rigid hierarchical tree, illustrating HSC progression in stages, from multipotent to differentiated cell types. Evaluating a

235 “dynamic” model with complex interactions, we interrogated the marker/cell-type interactions within
236 the regulatory cascade from supervised learning models, considering both the network architecture and
237 prediction performance as probes. LPS stimulation impacts all stages of HSC differentiation, thus having
238 a global impact on all surface markers[18]. Therefore, we assigned PBS vehicle control and LPS stimulation
239 as predicted labels to the samples (**Methods**). Twenty-two lineage/functional makers from nine
240 different cell types (Table S1) were encoded as essential numerical features, while genotype was loaded
241 as an external categorical/global feature.

242 Considering each node as a marker/cell type, their interactions could be represented as network
243 connections. Specifically, a linear regression model reflects the additive effect from markers/cell types; a
244 dense neural network models the complex interactions among markers/cell types; and a tree model
245 considers the binary relations between levels within the hierarchy (Figure 5A). LPS stimulation on the
246 cells (the prediction label) could be treated as a perturbation on the neural network. In this model,
247 HSPCs will be differentiated to the common lymphoid progenitors, unipotent precursor cells as well as
248 megakaryocyte-erythroid progenitor cells, eventually producing mature cells. Informed by the
249 differentiation hierarchy, we implemented Boosted Tree (BT) and Random Forest (RF) models along with
250 two reference models, Linear Regression (LR) and Deep Neural Network (DNN). The four models varied
251 in performance: Linear Regression (LR) achieved a prediction accuracy of 64.8%; the DNN model with
252 three-hidden layers enhanced the accuracy to 68.1%; the Boosted Tree (BT) model further increased the
253 prediction performance to 72.7%; and the Random Forest (RF) models topped the accuracy at 77.13%
254 (Figure 5B). This overall pattern of prediction accuracy (LR < DNN < BT < RF) suggests a non-linear,
255 hierarchical relationship among the features, consistent with the classical hematopoietic hierarchical
256 tree. Interestingly, when we stacked a Deep Neural Network on top of the Random Forest model, where
257 DNN functioned as a denoising head[21], the new stacked model showed degraded performance in
258 comparison to the RF model alone.

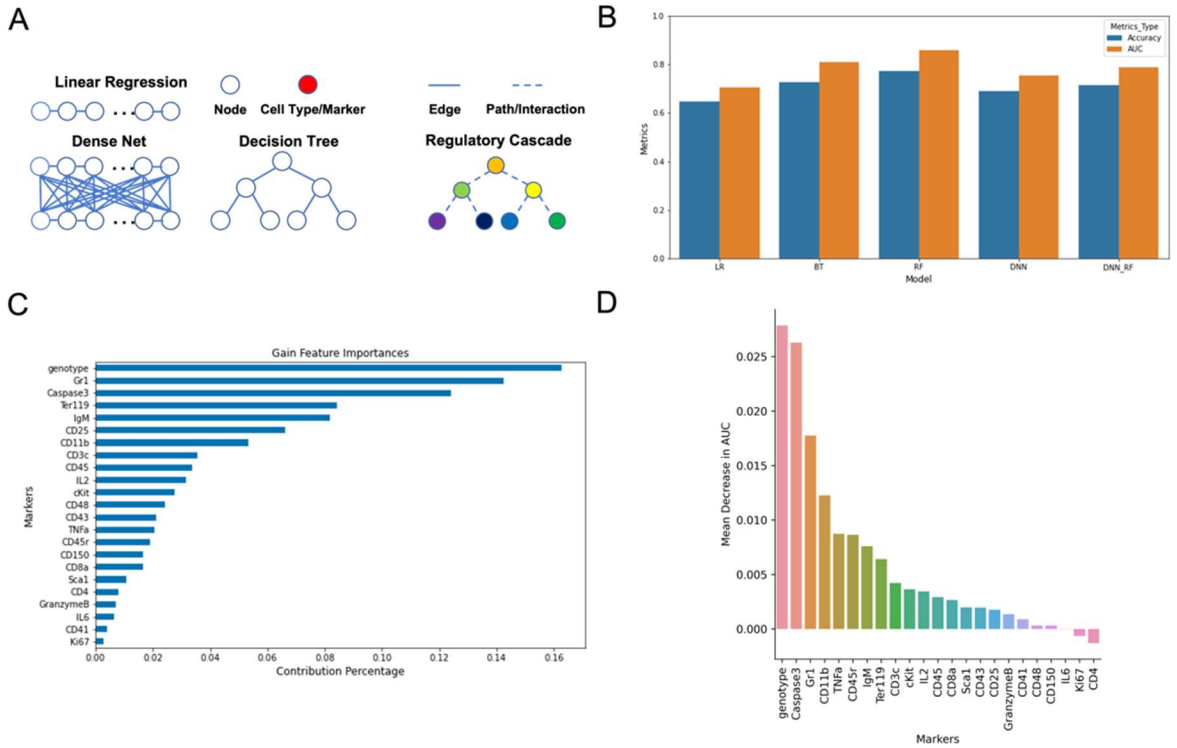


Figure 5. Mechanistic insights extracted from biologically-informed neural networks. Our neural network designs were guided by the biological context, in which we considered each node as a marker. The marker/cell-type interactions could therefore be represented as a network **(A)**, e.g., a decision tree or neural network could be linked to the regulatory cascade. With PBS vehicle control and LPS stimulation as labels, tree-based models (Boosted Tree [BT] and Random Forest [RF]) showed improved performance compared to Linear Regression (LR) and Deep Neural Network (DNN) models as well as DNN-RF stacked models **(B),** consistent with the hematopoietic tree hierarchy (i.e., representing a sequential binary branching decision process) among the markers. Feature importance of individual markers in the BT model, which implements gain-based feature importance **(C),** and RF model, which implements mean decrease in AUC **(D),** were examined. The group of markers with significant contributions would imply an overall cell response under inflammatory (LPS), e.g., pathogen-induced, challenges.

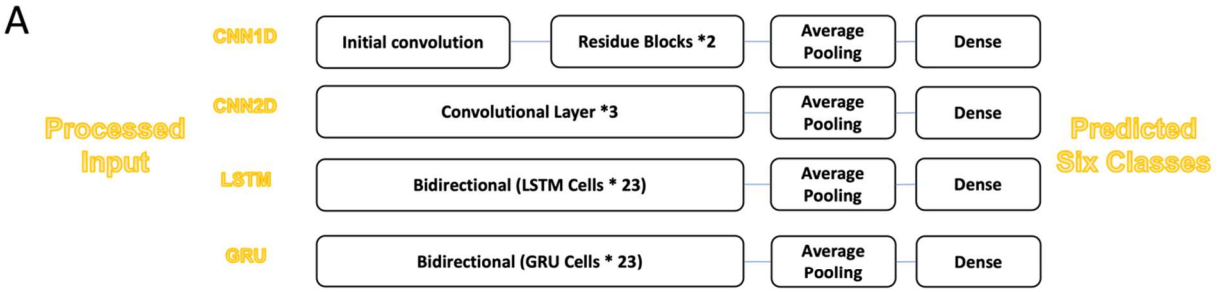
We further considered the feature contributions of individual markers using gain-based feature importance (Figure 5C) and mean decrease in AUC (Figure 5D). Gain-based feature importance reflects the relative contribution of a given feature to the model. Mean decrease in AUC indicates the decline in performance once the corresponding feature has been removed from the training dataset. We hypothesized that markers/features with the top feature contributions are biologically meaningful.

278 Indeed, the single most important contribution came from the strongest perturbation – genotype.
279 Individual markers, including Caspase3, Gr1, IgM, CD11b, and CD25, that contribute substantially to the
280 prediction of LPS stimulation, are highly biologically relevant in the full differentiation cycle of
281 hematopoiesis. Caspase 3 is an important player in the regulated cell death. Gr1 and CD11b reflect a
282 response from myeloid cells. And Ter119 indicates a strong immune response from the erythrocytes
283 under inflammatory challenges.

284

285 **Applying machine learning in sample classification**

286 While conventional methods/pipelines are limited by domain knowledge or manual
287 intervention, a predictive model that bypasses cell clustering or manual gating may provide an end-to-
288 end approach to the analysis of mass cytometry data. Here, we benchmarked a group of supervised
289 learning models, including CNN-1D, CNN-2D, RNN-LSTM, and RNN-GRU, that directly handle the raw
290 cytometry profiles for sample classification, i.e., prediction of clinical or biological outcome (Figure 6A).
291 We established multi-class output labels corresponding to six different sample types, including WT-PBS,
292 WT-LPS, Tet2-PBS, Tet2-LPS, TKO-PBS, and TKO-LPS. We used the curated ensemble of (1 hour)
293 "segments" from the initial bulk samples in evaluating the model performance. We found that CNN and
294 RNN-derived models demonstrated good prediction performance with at least 95% accuracy (Figure 6B).
295 We argue that convolutional layers and recurrent cells are efficient in learning the relevant features.
296 Algorithms such as CNN and LSTM as applied here, with cell count as the trait, captured with greater
297 fidelity the landscape of marker/cell-type interactions than the tree-based discrete models evaluated
298 above.



B

Model	Train Acc	Val Acc	Test Acc
CNN-1D	0.99783	0.97400	0.95887
CNN-2D	0.97580	0.95233	0.96537
RNN-LSTM	0.99819	0.96425	0.97186
RNN-GRU	1.000	0.98375	0.98052

Figure 6. Supervised learning identified sample type with high prediction performance. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) derived models (CNN-1D, CNN-2D, RNN-LSTM, RNN-GRU) were benchmarked for the task of sample type classification. The specifics of the architectures are shown in **(A)**. All tested models achieved high performance with over 95% in accuracy **(B),** demonstrating the importance of learning the complex interactions and high-order features.

Discussion

The current single-cell genomic revolution has enabled entire systems, such as the hematopoietic system, to be interrogated comprehensively at an unprecedented resolution, revealing a previously unappreciated degree of functional and phenotypic variation within cell compartments. In this work, bone marrow samples were harvested from mice and further characterized in mass cytometry – the baseline murine model (WT) was compared to genetic perturbations, including *Tet2*-knockout (*VavCreTet2*) and *VavCreBaxBakBid* triple-knockout (TKO). LPS served as an external stimulus or an innate immune response, simulating pathogen infection. Statistical modeling and unsupervised learning revealed the direct effect of genetic perturbations. Markers of interest that are typically processed individually could be combined in unsupervised learning models. Importantly, our study highlights a close connection among the intrinsic hierarchy of HSPC differentiation, neural network architecture, and prediction performance. Specifically, Boosted Tree (BT) and Random Forest (RF), which closely resemble

318 the stepwise binary branching decisions of the classical hierarchical model of hematopoiesis, showed
319 superior performance in comparison with Linear Regression (LR), Deep Neural Network (DNN) or even
320 the DNN-RF stacked model. The prediction performance was substantially enhanced relative to neural
321 network models, as the model reflected the cell differentiation cascade. However, as recent studies with
322 functional assays, non-invasive lineage tracing, and omics studies at single-cell resolution have shown[22],
323 hematopoiesis is a continuous process in which obvious boundaries are lacking between cell populations
324 located at different levels of the hierarchy. Consequently, given their architecture, tree-based models
325 may have an inherent limitation in capturing transition states, even with extensive optimization,
326 constraining their ability to further improve prediction performance. From this perspective, the complex
327 relationships among the cell populations (including feed-back loops in pathways) may be better
328 elucidated by neural networks.

329 Nevertheless, results of the analyses from the machine learning models were consistent with
330 biological insights such as myeloid-driven or innate inflammation (responses from TNFα and IL6). In
331 summary, the presented framework provides a unifying methodological framework for the analysis of
332 cytometry data, combining data-driven and model-driven approaches. We constructed interpretable
333 models that are anchored in biological context, with implications for ongoing and future studies in
334 single-cell biology.

335

336 Limitations of the study

337 Machine learning models, including graphical representation learning and tree-based
338 algorithms, were trained in datasets obtained from mass cytometry (CyTOF) characterization. While we
339 designed the neural network architectures with biological information as guide, assumptions of cell
340 types as nodes and edges as their interactions within a graph-based framework are based on the classic

341 discrete cell differentiation model. Cell differentiation is a continuous process, with intermediate states
342 among various levels leading to much more complex networks.

343

344 Methods

345 Mass cytometry and murine samples

346 Three generic murine models were included in this study, namely, wild type (WT), Tet2 knockout
347 (Tet2)[23], and *VavCreBaxBakBid* triple-knockout (TKO)[14]. Prior to bone marrow isolation, the mice had
348 been treated intraperitoneally with 1.5 mg/kg LPS (L4391; Millipore-Sigma) or PBS for 8 hours. The
349 obtained samples for the six genotype-condition types were subsequently characterized using mass
350 cytometry, according to previously established protocols[24]. Mass cytometry combines inductively
351 coupled plasma time-of-flight mass spectrometry and cytometry. Specifically, mass cytometry staining
352 was performed on whole bone marrow with antibodies conjugated to heavy metals. Cell responses from
353 9 different cell type and functional categories (stem cells, pan leukocytes, T lymphocytes, mature
354 myeloid cells, B lymphocytes, erythrocytes, cytokines, cell death, and proliferation) were recapitulated
355 by 16 surface/lineage markers and 6 function/intracellular markers. Markers from pre-gating channels
356 (DNA-1 and DNA-2) and instrument channels (Time/Event) were also considered (Table S1). The CyTOF
357 data generated from the six samples consisted of a maximum of 293,778 cells and a minimum of
358 121,124 cells.

359 Data pre-processing

360 Initial cell count data profiles obtained from the mass cytometry characterization were filtered
361 out for debris, doublets, and dead cells, thus reducing noise. They are represented as dynamical panels
362 on which the *arcsinh* transformation with co-factor = 5 and a binning window of 30 seconds was applied.
363 We sought to analyze the data in a way that maximizes robustness. The processed profiles were sliced

364 into ensembles of "segments" to achieve an optimized tradeoff between computational efficiency and
365 prediction accuracy. We evaluated an array of sizes ([2, 10, 30, 60, 120, 240, 480]), corresponding to
366 2*30S=1M, 10 * 30S = 5M, 30 * 30S = 15M, 120 * 30S = 1H, 240 * 30S = 2H, & 480 * 30S = 4H sampling
367 intervals (M: Minute, H: Hour), respectively. The obtained pile of "segments" was completely shuffled
368 and re-assembled into new ensembles. We implemented an RNN-LSTM model for this special
369 "hyperparameter search" (see prediction accuracy in Table S2) to determine the optimal segment size,
370 which was 1H. Signals of select markers, including Sca1, CD45, TNFα, Caspase 3, and Ki67, were
371 aggregated and averaged into 1 hour binning window for the statistical test and GMM estimation.

372 **Mechanistic Deep Learning Framework**

373 *Statistical test and model*

374 Wilcoxon test was performed on pairwise comparisons of marker intensity. Pearson correlation
375 coefficient was calculated for inter-marker association. We implemented a GMM on the marker Ki67.
376 We considered fixed effects and interactions among time, genotype, and condition (PBS control or LPS
377 stimulation). After testing the random slope of a sample, $(1 + \text{Time} \mid \text{Sample})$, we considered only the
378 random intercept from a sample, i.e., $(1 \mid \text{Sample})$. The models were evaluated using corrected Akaike's
379 Information Criterion (AICc), Bayesian Information Criterion (BIC), and Marginal R squared (R2M). The
380 optimal model was found to be given by the following expression:

381 $$\text{Marker Intensity} = \text{Time} + \text{Genotype} + \text{Condition} + \text{Genotype:Time} + \text{Condition:Time} + (1 \mid \text{Sample}) + \varepsilon$$

382 *Unsupervised learning models*

383 We analyzed two established unsupervised learning pipelines, namely Phenograph[5] and T-REX[6],
384 which were applied to the preprocessed CyTOF dataset with binning window 30S. For both pipelines,
385 Uniform Manifold Approximation and Projection (UMAP)[25] was used for dimensionality reduction; k-
386 nearest neighbors (kNN)[26] for clustering; Marker Enrichment Modeling (MEM) for generating the

387 enriched heatmap. The applied hyper-parameters are presented in Table S3. For comparison, we
388 screened and enriched the cell populations derived from the T-REX algorithm through Density-Based
389 Spatial Clustering and Application with Noise (DBSCAN) and MEM.

390 *Supervised learning models*

391 All the machine learning models were trained using the *TensorFlow (TF)* ecosystem[27], which we
392 applied to the preprocessed CyTOF dataset (with binning window size of 30S) to perform model training
393 and evaluation. Linear Regression (LR), Deep Neural Network (DNN), Boosted Tree (BT), and Random
394 Forest (RF) models were implemented on the dataset with train and test split of 8:2. The DNN was
395 implemented as a multilayer perceptron. A stacked model of Dense Neural Network and Random Forest
396 (DNN-RF) was constructed following the same splitting. Models implementing a Convolutional Neural
397 Network (CNN) or a Recurrent Neural Network (RNN) in the framework include CNN-1D, CNN-2D, RNN-
398 LSTM, and RNN-GRU. For these four architectures, pre-processed datasets were divided into train, valid,
399 and test data according to 6:2:2 ratio.

400 We trained the neural networks using mean squared error (as the loss function), *Adam* (an
401 efficient stochastic optimization method), and Leaky ReLU (as the activation function), building on our
402 previous work[16,17]. To reduce the possibility of overfitting, we implemented *dropout* assuming a dropout
403 rate of 20%.

404

405 **Acknowledgments**

406 This research is supported by National Institutes of Health (NIH) grants NHGRI R35HG010718,
407 NHGRI R01HG011138, NIA AG068026, and NHLBI R01HL133559.

408

409 **Author contributions**

410 B.W., S.S.Z., and E.R.G. conceived the project and designed the study. B.W. performed the
411 analyses and drafted the article. B.W., S.S.Z., and E.R.G. revised the article. S.S.Z. and E.R.G. supervised
412 the project and acquired funding for the study.

413

414 **References**

415 1. Davis, M. M., Tato, C. M. & Furman, D. Systems immunology: just getting started. *Nat. Immunol.* **18**,
416 725–732 (2017).
417 2. Gadalla, R. *et al.* Validation of CyTOF Against Flow Cytometry for Immunological Studies and
418 Monitoring of Human Cancer Clinical Trials. *Front. Oncol.* **9**, (2019).
419 3. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a
420 Human Hematopoietic Continuum. *Science* **332**, 687–696 (2011).
421 4. Hahne, F. *et al.* flowCore: a Bioconductor package for high throughput flow cytometry. *BMC*
422 *Bioinformatics* **10**, 106 (2009).
423 5. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that
424 Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
425 6. Barone, S. M. *et al.* Unsupervised machine learning reveals key immune cell subsets in COVID-19,
426 rhinovirus infection, and cancer therapy. *eLife* **10**, e64653 (2021).
427 7. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep
428 learning model for cytometry data. *Proc. Natl. Acad. Sci.* **117**, 21373–21380 (2020).
429 8. Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via
430 representation learning. *Nat. Commun.* **8**, 14825 (2017).

431 9. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* **182**,
432 1384–1400 (2020).
433 10. Domen, J., Cheshier, S. H. & Weissman, I. L. The role of apoptosis in the regulation of
434 hematopoietic stem cells: Overexpression of Bcl-2 increases both their number and repopulation
435 potential. *J. Exp. Med.* **191**, 253–264 (2000).
436 11. Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins: changing partners in the dance
437 towards death. *Cell Death Differ.* **25**, 65–80 (2018).
438 12. Salisbury-Ruf, C. T. *et al.* Bid maintains mitochondrial cristae structure and function and protects
439 against cardiac disease in an integrative genomics study. *eLife* **7**, e40907 (2018).
440 13. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference
441 transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
442 14. Wagner, P. N. *et al.* Increased Ripk1-mediated bone marrow necroptosis leads to myelodysplasia
443 and bone marrow failure in mice. *Blood* **133**, 107–120 (2019).
444 15. Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry
445 data. *Nat. Methods* **14**, 707–709 (2017).
446 16. Wang, B. & Gamazon, E. R. *Towards mechanistic models of mutational effects: Deep Learning on*
447 *Alzheimer's Aβ peptide*. 2021.12.19.473403
448 https://www.biorxiv.org/content/10.1101/2021.12.19.473403v2 (2021)
449 doi:10.1101/2021.12.19.473403.
450 17. Wang, B. & Gamazon, E. R. Modeling mutational effects on biochemical phenotypes using
451 convolutional neural networks: application to SARS-CoV-2. *iScience* **7**, 104500 (2022).
452 18. King, K. Y. & Goodell, M. A. Inflammatory modulation of HSCs: viewing the HSC as a foundation
453 for the immune response. *Nat. Rev. Immunol.* **11**, 685–692 (2011).

454 19. Cai, Z. *et al.* Inhibition of Inflammatory Signaling in Tet2 Mutant Preleukemic Cells Mitigates
455 Stress-Induced Abnormalities and Clonal Hematopoiesis. *Cell Stem Cell* **23**, 833-849.e5 (2018).
456 20. Diggins, K. E., Greenplate, A. R., Leelatian, N., Wogsland, C. E. & Irish, J. M. Characterizing cell
457 subsets using marker enrichment modeling. *Nat. Methods* **14**, 275–278 (2017).
458 21. Xie, J., Xu, L. & Chen, E. Image Denoising and Inpainting with Deep Neural Networks. in *Advances*
459 *in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
460 22. Zhang, Y., Gao, S., Xia, J. & Liu, F. Hematopoietic Hierarchy – An Updated Roadmap. *Trends Cell*
461 *Biol.* **28**, 976–986 (2018).
462 23. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and
463 myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
464 24. Zinkel, S. S. *et al.* Proapoptotic BID is required for myeloid homeostasis and tumor suppression.
465 *Genes Dev.* **17**, 229–239 (2003).
466 25. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for
467 Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).
468 26. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M. k-Nearest Neighbor Classification. in *Data*
469 *Mining in Agriculture* (eds. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M.) 83–106 (Springer, 2009).
470 doi:10.1007/978-0-387-88615-2_4.
471 27. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed
472 Systems. 19.
473
474
475
476

Supplementary Information

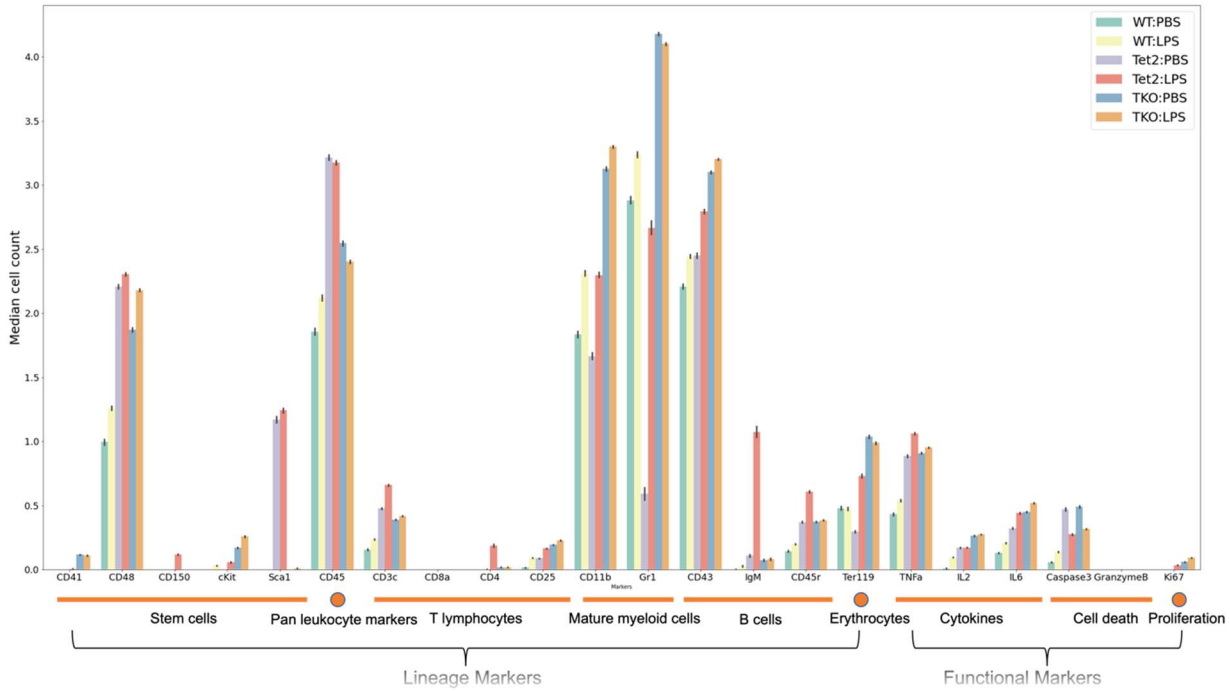
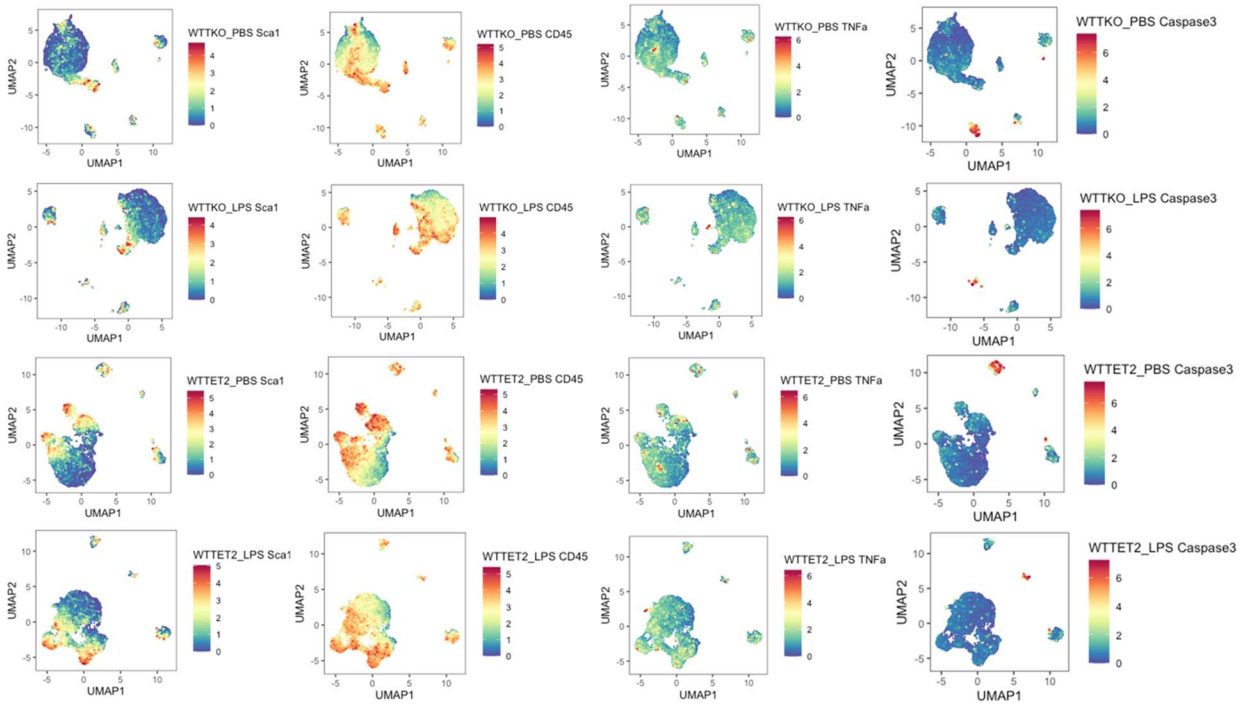


Figure S1. Cell count summary statistics. The sample data profiles (WT-PBS, WT-LPS, Tet2-PBS, Tet2-LPS, TKO-PBS, and TKO-LPS) were obtained from mass cytometry characterization, from which cell count summary statistics were computed. Lineage and functional marker groupings are defined by the cell types.



483

484 **Figure S2. Differential cell count illustrated from the unsupervised learning.** Markers of interest under

485 genetic perturbation, including Sca1 (LSK cells), CD45 (leukocytes), TNFα (cytokines), and Caspase 3 (cell

486 death indicator), were mapped to the clustering results from the T-REX algorithm.

487

488 **Table S1.** Cell markers investigated in this study

Marker Type	**Cell Type**	**Markers**
Lineage	Stem Cells	CD41, CD48, CD150, cKit, Sca1
Lineage	Pan Leukocytes	CD45
Lineage	T Lymphocytes	CD3c, CD8a, CD4, CD25
Lineage	Mature Myeloid Cells	CD11b, Gr1
Lineage	B Lymphocytes	CD43, IgM, CD45R
Lineage	Erythrocytes	Ter119

Functional	Cytokines	TNFα, IL2, IL6
Functional	Cell Death	Caspase3, Granzyme B
Functional	Proliferation	Ki67
Pre-gating		DNA1, DNA2
Instrument		Cell Length, Time

Table S2. Searching for the optimal size "hyperparameter" via RNN-LSTM model architecture

Micro Sample Size	Train Acc	Val Acc	Test Acc
2	0.73653	0.71342	0.71033
10	0.93403	0.88667	0.87375
30	0.99791	0.95369	0.95125
60	**0.99499**	**0.9325**	**0.9575**
120	1.0	0.9598	0.975
240	1.0	0.93	0.93
480	1.0	0.77551	0.88

Table S3. Hyperparameters applied in PhenoGraph and T-REX

Key components in pipeline	**Phenograph**	**T-REX**
UMAP	ret_nn = TRUE	
kNN	k = 30 for single sample, k = 60 for paired sample	