## Title

Data analysis to modeling to building theory in NK cell biology and beyond: How can computational modeling contribute?



## Journal

Journal of Leukocyte Biology, 105(6)

## ISSN

0741-5400

## Authors

Das, Jayajit
Lanier, Lewis L

## Publication Date

2019-05-27

## DOI

10.1002/jlb.6mr1218-505r

Peer reviewed

DOI: 10.1002/JLB.6MR1218-505R

JLB JOURNAL OF LEUKOCYTE BIOLOGY

REVIEW

# Data analysis to modeling to building theory in NK cell biology and beyond: How can computational modeling contribute?

**Jayajit Das**[1,2,3,4] | **Lewis L. Lanier**[5]

[1]Battelle Center for Mathematical Medicine, Research Institute at the Nationwide Children's Hospital, Columbus, Ohio, USA

[2]Department of Pediatrics, The Ohio State University, Columbus, Ohio, USA

[3]Department of Physics, The Ohio State University, Columbus, Ohio, USA

[4]Biophysics Program, The Ohio State University, Columbus, Ohio, USA

[5]Department of Microbiology and Immunology and the Parker Institute for Cancer Immunotherapy, University of California, San Francisco, California, USA

**Correspondence**
Jayajit Das, Battelle Center for Mathematical Medicine, Research Institute at the Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH 43205.
Email: jayajit@gmail.com
Lewis L. Lanier, Department of Microbiology and Immunology and the Parker Institute for Cancer Immunotherapy, University of California, San Francisco, CA 94143.
Email: Lewis.Lanier@ucsf.edu

## Abstract

The use of mathematical and computational tools in investigating Natural Killer (NK) cell biology and in general the immune system has increased steadily in the last few decades. However, unlike the physical sciences, there is a persistent ambivalence, which however is increasingly diminishing, in the biology community toward appreciating the utility of quantitative tools in addressing questions of biological importance. We survey some of the recent developments in the application of quantitative approaches for investigating different problems in NK cell biology and evaluate opportunities and challenges of using quantitative methods in providing biological insights in NK cell biology.

**KEYWORDS**

CyTOF, data-driven modeling, mathematical modeling, single cell measurements, single cell trajectory, systems immunology

## 1 | INTRODUCTION

Scientific progress is marked by formation of paradigms.[1] Thomas Kuhn defines paradigms as, "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners."[1] Experimental investigations, mechanistic explanation of the acquired data, and development of frameworks to explain a broad range of observations are essential steps to generate paradigms.[1,2] This pattern has been observed historically in physical as well as in biological sciences. The application of mathematical or computational approaches in generating paradigms in physical sciences has a history of over 2000 years[1]; however, the role of quantitative approaches in producing tangible progress leading to formation of paradigms in many areas of biology including immunology is still debated. It is not uncommon for a modeler to provide convincing answers to experimental collaborators, journal referees,[3] or funding agencies to questions such as: Can models tell us something new and useful about the system that we cannot intuit from our knowledge and experience? Can models help replace costly experiments with *in silico* simulations or help us design experiments? Even mathematical modelers and bench scientists who apply quantitative tools wonder about these questions as they assess the impact of their contributions in an area that has been historically spearheaded by experimentation. These questions might not lead to conclusive answers all the time, which perhaps is one of the reasons that these questions still persist in the field. We attempt to make a case for the relevance of mathematical and computational modeling in understanding the biology of leukocytes, in particular, NK cells in the light of the above questions by reviewing a few recent investigations. We hope this will help the reader to assess these questions and lead to the answer in the context of the system she or he is interested in.

Mathematical and computational tools have been an integral part of progress in the physical sciences. This is perhaps most evident in the time and resources invested in testing theoretical predictions in

Abbreviations: CA, correspondence analysis; CyTOF, cytometry by the time of flight; KIRs, killer-cell immunoglobulin-like receptors; NKRs, NK cell receptors; PCA, principal component analysis; SFKs, Src family kinases; sc-RNA-seq, single-cell RNA sequencing; t-SNE, t-distributed stochastic nonlinear embedding.

Received: 26 December 2018 | Revised: 25 March 2019 | Accepted: 3 April 2019

  

the experiments carried out in the Large-Scale Hadron Collider and the Laser Interferometer Gravitational-Wave Observatory projects. Therefore, it would be useful to take a look at NK cell biology with the scope of a physical scientist and determine the similarities and differences between a typical physical system and NK cells, and more generally leukocytes. Analysis of a system in physical sciences begins by identifying the length and the time scales involved in the system.[4] Consider the scales involved in single NK cells to NK cell populations. Nanometer-sized NK cell receptors (NKRs) on single NK cells interact with cognate ligands of similar sizes on a target cell to mount responses composed of lysis of target cells, secretion of cytokines, and in some cases $10^3$–$10^4$-fold clonal expansion that spread to different organs (length ≥ cm).[5,6] The NKR–ligand interactions occurring in time scales of a few seconds generate responses lasting for days (e.g., generation of "memory" NK cells lasting for longer than a month). These processes describe a $10^7$-fold change in length and a $10^5$-fold change in time scale. The above change of scales is similar to processes relating the formation of snowflakes from a collection of jiggling molecules in water vapor.[7] There is a multitude of processes that work together to relay changes from the smallest to the largest scales in these systems. Mathematical models, often providing a reduced or a coarse-grained description of these physical processes, become essential to describe such multiscale systems.[8–11] Reduced models have been successful in describing several aspects of NK cell and leukocyte biology, in particular for signaling and development in NK cells, and evolution of NKR repertoires. These models are discussed in Section 1.

There is a unique aspect of NK cell biology and of the immune system in general with no counterpart in nonliving physical systems. It is the enormous diversity of the building blocks (e.g., single cells) of the system.[12,13] The size of the proteome of human NK cells is over 3000.[14] The number of phenotypically distinct NK cells in an individual can range from 6000 to 30,000.[15] The NK cell responses perhaps impact most of the human proteome[16] whose size is over 20,000 proteins.[17] Thus, application of reduced models in addressing questions where the diversity of NK cells is relevant can become challenging. As recent advances in sequencing and single-cell technologies probe single cells with more details, the diversity of the NKRs, phenotypes, and lineages keeps increasing and many appear to be important for determining NK cell responses. Therefore, a key question in the community is about how to characterize the large-scale data and then use the analysis to glean underlying mechanisms and develop therapies. A substantial research activity in recent years has been directed toward development of data analysis tools and statistical models to characterize, visualize, and, interpret functional implications of these large datasets. These advances are reviewed in Section 2. However, these data-driven techniques lack the ability of the reduced models to determine underlying mechanisms. Thus, it is an open challenge to be able to build appropriate mechanistic models in the context of the high-dimensional data describing the system. This issue is reviewed in Section 3.

We discuss the role of quantitative tools in generating theories for NK cell biology in Section 4. The definition of theory in biology can vary. A report from the National Academy of Sciences defines theory in biology as a "collection of models,"[18] whereas Shou et al.[19] describe scientific theory as a "unifying framework that can explain a large class of empirical data". The later point of view is more consistent with the concept of theory, also a hallmark of a paradigm, in physical sciences. For example, Kepler's planetary laws explained the detailed data acquired by Tyco Brahe pertaining to planetary motion, and Newton's gravitational theory described not only the planetary motion but also the motion of any object possessing a mass.[20] We will lean toward the above definition of theory in discussing efforts to develop theories for NK cell biology. The major theories in physical sciences, such as the theory of gravitation or the theory of electricity and magnetism, have been described by mathematical relations, whereas the dominant theories in biology such Darwin's evolutionary theory[21] or the clonal selection theory by Burnet[22] were proposed originally in qualitative terms. In the later years the works of Fisher, Wright, Haldane, and others provided a strong mathematical foundation for Darwin's theory.[23] Burnet's clonal selection theory and related theories were formulated mathematically by Jerne[24] and later by Perelson and Oster.[25] However, apart from these singular examples, theories or major hypotheses in many areas of biology have been pursued in qualitative terms. The lack of the use of mathematical formulations to push forward theories in biology arises in part due to the unique complexity of the biological systems. Furthermore, the technical and conceptual challenges in developing broad principles for describing many body physical systems[8] specifically that are not in equilibrium (e.g., dissipates energy) also apply to living systems.[26] The research covered in Sections 2–4 mark the progress made in the execution of the successive steps required to create a paradigm. We have highlighted the top 5 advancements in the area of computational modeling of NK cell biology in Box 1 to help the reader with the brief introduction to the field. We conclude by discussing few future directions in Section 5.

## 2 | REDUCED MODELS

Reduced models provide an intuitive and approximate description of biological systems. These models have been widely employed to investigate signaling kinetics, development, and evolution of receptor repertoire in leukocytes such as T cells, B cells, and NK cells. We will restrict our discussion to NK cells here. Excellent overviews on this topic pertaining to T and B cells are available in prior reviews.[27–32]

### 2.1 | Signaling and activation of NK cells

Coarse-grained models studying NK cell signaling kinetics describe signaling events as biochemical reactions between signaling species (usually proteins). The molecular details of protein structures in regulating protein–protein interactions are approximated into mass-action reaction rates.[33] The signaling proteins and the reactions in these models are usually chosen intuitively based on the literature and provide a reduced description of the actual detailed reactions,[34] that is, multiple phosphorylation states are described by fewer activation states or molecular details of enzymatic regulations are reduced to few Michaelis–Menten parameters. The biochemical reaction kinetics are modeled by deterministic mass action kinetics or by stochastic

**Box 1. Major advancements in application of quantitative approaches to NK cell biology**

1. *Mathematical modeling of development of Ly49 repertoire*: First proposed by Raulet and Vance in 1998,[35] and then further studied by Johansson et al.[36] and extended to KIRs by Andersson et al.[37] The model by Raulet and Vance is one of the first examples of mathematical modeling in NK cell biology.
2. *Modeling of signal integration in NKR signaling*: Das[38] and Mesecke et al.[39] developed mechanistic NKR signaling models for activating CD16 and inhibitory Ly49A receptors in mouse NK cells, and, activating NKG2D and inhibitory NKG2A receptors in human NK cells, respectively. These models quantified mechanisms that underlie signal integration in NK cells.
3. *Quantification of NK cell diversity*: Horowitz et al.[15] used mass cytometry (CyTOF) measurements and analysis of high-dimensional datasets (SPADE) to characterize the enormous diversity of NK cells (~6000–30,000 phenotypes) in a human subject and elucidated the roles of genetics and the environment in regulating the NKR diversity.
4. *Modeling of the evolutionary arms race between viruses and NKRs*: Carrillo-Bustamante et al. developed agent-based models to describe evolution of inhibitory KIRs demonstrating that "decoys" produced by viruses diversify inhibitory KIRs.[40] This model was later extended to include activating NKRs.
5. *Data-driven model with mechanistic insights for analyzing cytokine-NKR synergy*: Mukherjee et al.[41] developed a data-driven model to analyze mass cytometry data and provide mechanistic insights that underlie the synergy between IL-2 treatment and NKG2D mediated activation of human primary NK cells.

processess (e.g., Markov processess) that account for the random fluctuations in the protein copy numbers originating from the thermal fluctuations. A wide variety of software packages[42–45] are available to simulate these biochemical reactions where the user inputs the model as a set of biochemical reactions with specified kinetics rates and initial abundances of reactant species. It can be challenging to know the values of all the reaction rates and abundances of the signaling species in a model because of the following reasons: (i) Many of these values are measured in in vitro experiments, which can change in the in vivo environment and in the context of specific cell type (primary cells vs cell lines). (ii) It can be difficult to measure some of these parameters in experiments. These issues are common for developing kinetic models for cell signaling and are dealt with by carrying out sensitivity analysis of the key results against variations of the unknown parameters[46] or by estimating parameters using measured data (e.g., kinetics of a particular protein abundance).[47] Usually the number of model parameters is much larger compared to the variables that can be measured and contribute toward large confidence intervals in parameter estimations. Parameter identifiability analyses are carried out to determine such sloppy parameters, which can suggest new measurements or reparameterization of the model to constrain the parameter values.[48,49] This remains an active area of research in systems biology.[50]

A reduced model was set up to quantitatively characterize the "missing-self" hypothesis in mouse NK cells stimulated by activating CD16 and inhibitory Ly49A receptors.[38] The missing-self hypothesis states that healthy target cells express ligands cognate to activating and inhibitory NKRs such that the opposing signals generated by these interactions are balanced out and result in tolerance in NK cells interacting with the target cells (Fig. 1). This balance is disrupted in transformed or virally infected target cells due to down or up regulation of inhibitory or activating ligands leading to NK cell activation and lysis of the target cells. In the model, the Src family kinases (SFKs) phosphorylated ITAMs and ITIMs associated with activating and inhibitory NKRs, respectively. The Syk family kinases (Syk and Zap70), recruited by phosphorylated ITAMs, phosphorylated the guanine nucleotide exchange factor Vav1, and the phosphatase, SHP-1 bound to phospho-ITIMs, and dephosphorylated p Vav1. Vav1 phosphorylation resulted in Erk phosphorylation, which was used as a marker for NK cell cytotoxicity. The modeling showed that pVav1 increases sharply as the abundance of activating ligands is increased. The sharp Vav1 activation profile arises due to the competitive nature of enzymatic activation and deactivation of Vav1 molecules mediated by phosphorylated Zap70 or Syk and receptor-bound SHP-1 molecules, respectively. This is a consequence of zeroth order ultra-sensitivity[51] in the biochemical reactions regulating Vav1 phosphorylation. In such cases, the ratio (R) of the concentrations of the enzymes mediating activation and deactivation plays a decisive role in producing activation. Thus, the model suggested integration of activating and inhibitory signals in NK cells occurs as a ratio rather than as a sum.

Another modeling study by Mesecke et al.[39] considered activation of the human NK cell line NKL by activating NKG2D and inhibitory CD94-NKG2A receptors. The authors considered 72 different computational models constructed by considering different combinations of 7 different signaling modules such as association of SFKs with phosphorylated NKG2D-DAP10 complexes, exclusion of CD45 from the immunologic synapse, and dephosphorylation of SHP-1 by SFKs. In the model, SFKs phosphorylated Vav1 and SHP-1 dephosphorylated pVav1. Comparison of model results against measurement of pVav1 at different concentrations of agonist antibodies ($\alpha$-NKG2D and $\alpha$-NKG2A) cognate to activating NKG2D and inhibitory CD94-NKG2A receptors showed that the physical association of SFKs with NKG2D-DAP10 is necessary to generate the sharp increase in Vav1 activation as the concentration of activating (or inhibitory) ligands increased (or decreased). The study also measured abundances of several proteins in the NKL cells that were used in developing the model. NK cell signaling models for other NKRs such as CD16 and 2B4, have been developed recently.[52] Spatial clustering of killer-cell immunoglobulin-like receptors (KIRs) and select signaling proteins play an important role in

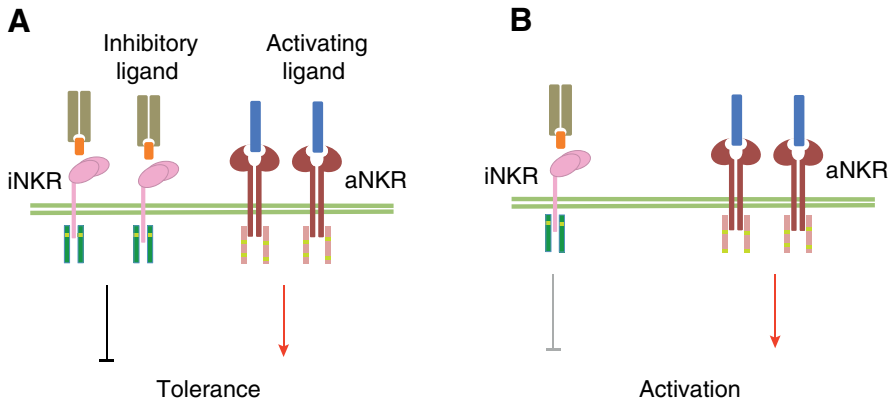


**FIGURE 1** **The missing self-hypothesis**. NK cells interact with target cells expressing ligands cognate to a diverse set of activating (aNKR) and inhibitory (iNKR) NK receptors. (**A**) Healthy target cells express ligands that lead to a "balance" in the signals generated by the dueling NKRs. (**B**) The missing self-hypothesis posited that infected or transformed target cells express lower numbers of inhibitory ligands (e.g., MHC class I molecules) that leads to a bias in the activating signals generated by the aNKRs. However, there are various other scenarios (e.g., increase in activating ligands, decrease in inhibitory ligands) that can favor activating over inhibitory signals leading to NK cell activation. What is the quantitative nature of the balance between the activating and the inhibitory signals in NK cells? For example, is the balance given by the sum or the ratio of the numbers of activating and inhibitory ligands? The results from the model by Das[38] point to the later

NK cell signaling and activation.[53,54] Spatially resolved in silico models reviewed in refs. 55,56 have been developed to analyze roles of such clustering.

## 2.2 | NK cell population model

Models describing roles of NK cell populations in immune surveillance,[57] in responding to viral infections,[58,59] and in interactions with lymphocytes of the adaptive immunity have been developed.[58,59] Excellent reviews of these models investigating immune cell population kinetics are available in the literature.[55,60–62] These models are constructed using variables that describe different cell populations (e.g., NK cells, tumor cells, T cells) or viruses, and the population kinetics are described by ordinary differential equations, partial differential equations, or agent-based models. These models include processes such as lysis of tumor cells or secretion of cytokines or differentiation of NK cells quantified by mass-action rates. Recent advances in microscopy has made it possible to image development of single precursor NK cells on stromal cells over a long time (~21 days). Khorshidi et al.[63] characterized images of tracks of single NK cells in the presence of target cells in vitro and in the mouse spleen using models of random walk to find that NK cells showed more directed movements compared to a pure Brownian walk. Lee and Mace[64] modeled these movements using random Brownian walks, directed Levy walks, and constrained random walks, and found that the proportion of the cells executing directed Levy walks increased as NK cells become more mature. These different models of random movements could provide insights[65] regarding if NK cells choose specific migration properties at different stages of development to optimize contact time with stromal cells or killing of target cells.

## 2.3 | Development of NKR repertoire

Probability and agent-based models have been constructed to describe development of NKRs that are specific or nonspecific to self-MHC class I within a host. These models describe activation of NKR genes and selection of NK cells during development by few coarse-grained processes that are executed with simple probability rules. Vance and Raulet[35] developed a binomial probability-based model to evaluate consequences of 2 hypotheses describing the NK cell education, namely, a 2-step selection and a sequential model, on determining the NKR repertoire. Johansson et al.[36] developed agent-based models that simulated an extended version of the Vance and Raulet model to evaluate NKR repertoire in 4 different MHC class I backgrounds. The simulations favored the 2-step selection hypothesis, where NKRs are acquired stochastically and then selected when the inhibitory signal crosses a threshold value. Anderson et al.[37] analyzed distributions of multiple inhibitory KIRs in human donors and found correlations in coexpressions of the KIRs. These distributions were modeled using a correlated-probability model. Their analysis of the distributions of self and non-self KIRs suggested against the sequential selection model and favored random acquisition of the KIRs. An overview of these models[66] can be found in ref. 55.

## 2.4 | Evolution of NKR repertoire

NKRs coevolve with viruses such as CMV.[67] Agent-based models describing the above evolution kinetics have been developed. In agent-based models agents representing different components of the system (e.g., healthy individuals, CMV-infected individuals) evolve with time following a set of rules. Carrillo-Bustamante et al.[40] developed agent-based models where healthy individuals carrying one MHC locus and one KIR haplotype become infected by a herpes-like virus. The healthy and infected individuals also reproduce and die. The infected individuals recover or become chronically infected, and the KIR genes in the germline and the virus can also mutate. The above processes occur with specific rates. The simulations showed that the generation of decoy molecules by the virus increases the diversity of the inhibitory KIR repertoire.[40,66] Later models by Carillo-Bustamante

et al.[68] included activating NKRs and modulation of NKR-MHC class I interactions by peptide fragments.[69]

Above reduced models produced mechanisms and hypotheses and helped design experiments that were not possible by simple logical extension of the known experimental results available at the time of their construction. For example, Vance and Raulet's mathematical formulation[35] of two competing hypotheses regarding development of NK cells predicted patterns of NKR repertoires suggested experiments that can help choose one mechanism over the other. Can these models be used to unify a broad range of features in NK cell biology? It is difficult to give a straightforward answer to this question in the face of new details about NK cell biology that are emerging from recent single cell[70] and sequencing measurements.[71,72] These multidimensional datasets may lead to revision of assumptions and approximations made in these models, and some of the new data could be incompatible with these models as many of these models' behaviors are sensitive to changes in key assumptions or parameter values.[50]

## 3 | CHARACTERIZATION, VISUALIZATION, AND INTERPRETATION OF HIGH-DIMENSIONAL DATASETS

Traditional flow cytometry methods enabled us to assay about 4–10 proteins in single cells and recent developments in mass cytometry techniques such as cytometry by the time of flight (CyTOF) increased that number almost by an order of magnitude to 40–100 proteins.[73,74] This number of measured proteins still remains small (<1% of the proteome size) compared to the large number of proteins involved in NK cell responses, nevertheless, it provides a glimpse of the details that are involved in generating single-cell responses. The large dimensions (40–100) of these datasets have necessitated the use of quantitative methods to classify, characterize, interpret, and visualize the data in lower (2 to 3) dimensions. Clustering algorithms,[75] in particular, hierarchical clustering, have been used to group single cells in high-dimensional datasets generated by CyTOF measurements. Hierarchical clustering approaches can be broadly classified into 2 types: agglomerative and divisive.[76] Agglomerative clustering starts by assigning a cluster to each single cell and then successively merging the most "similar" smaller clusters into bigger clusters. The "similarity" between a pair of smaller clusters is quantified by a metric such as the Euclidean distance (Fig. 2A). The divisive clustering begins by including all the objects (e.g., single cells) in a giant cluster and then breaking it up successively based on the similarities between the objects (Fig. 2A). A list of the clustering algorithms, visualization tools, and the biological questions they address are shown in Table 1. The large variety in the algorithms used to address different aspects of the immune response bears similarity with which different and often conflicting theories were used to describe electric and magnetic phenomena in the 1700s before the physical theory of electricity and magnetism was discovered.[1] We will briefly discuss the tools that were developed to characterize diversity and function in lymphocytes, in particular NK cells, using CyTOF measurements. A summary of these methods is provided in Table 1.

SPADE was one of the earliest computational tools developed to characterize and graphically visualize CyTOF data.[77] SPADE uses an agglomerative clustering (Fig. 2A) to cluster phenotypically similar cells. These clusters are then visualized as a tree graph (SPADE tree) where the vertices representing the clusters are connected by edges that minimize the total edge length (or a minimum spanning tree construction). Qiu et al.[77] represented the hierarchy of phenotypes in population of peripheral blood mononuclear cells (PBMCs) obtained from human donors using SPADE trees. Horowitz et al.[78] applied the SPADE analysis on CyTOF measurements for PBMCs obtained from identical twins and relative expressions of NKRs in SPADE trees demonstrated that the expression for inhibitory receptors in diverse (over 6000) NK cell populations in an individual is largely determined by the genetics of the individual. However, SPADE does not preserve the single-cell resolution in the output as the agglomerative clustering method merges smaller clusters into larger clusters. Therefore, other dimension reduction schemes that preserve the resolution at the single-cell level have been used for visualization and classification of cell phenotypes.

Several studies[79–81] employed principal component analysis (PCA)[82,83] and correspondence analysis (CA)[82] to project high-dimensional single-cell data to 2 or 3 dimensions in order to visualize the data and perform further analysis in the lower dimensions (Fig. 2). PCA-based methods preserve variances in the single-cell data, but do not preserve the local geometric structure in the lower dimensional projections. Thus, clustering algorithms that use geometric distance (e.g., Euclidean distance) between pairs of data points for phenotype classification cannot be combined with such methods. CA is a PCA-based method that is applied on high-dimensional Boolean (presence or absence of proteins) data. CA was used on a Boolean representation of receptor and signaling protein expressions in NK cells in the presence or absence of HIV-infected $CD4^+$ T cells, and the quantification of the spread of the data in 2 dimensions suggested a short-term increase in NK cell diversity in the presence of HIV infection.[80]

A more popular method of visualization of CyTOF data in lower dimensions is based on t-distributed stochastic nonlinear embedding (t-SNE)[84] (Fig. 2B). t-SNE preserves the local geometric structure in the lower dimensional projections. t-SNE assigns a probability (using a t-distribution function) for a pair of cells to be separated by a Euclidean distance in high dimensions and develops a lower dimensional representation that minimally changes the assigned probabilities. Amir et al.[85] developed a visualization tool (vi-SNE) for CyTOF data using t-SNE. Shekhar et al.[86] developed a method (ACCENSE) for generating automatic clustering of single cells in using a kernel density transformation of the lower dimensional t-SNE data. The relative comparison between SPADE, vi-SNE, and ACCENSE is reported by Anchang et al.[87] k-means clustering or kernel density-based clustering (e.g., used in ACCENSE) assume a convex geometric local structure; however, the CyTOF datasets often show nonconvex local structures (Fig. 2A). Newman and Girvin developed a network detection algorithm that does not assume convex structures in the context of determining connected communities.[76] This algorithm was adapted by Levin et al.[88] (PhenoGraph) for analyzing CyTOF datasets to determine biomarkers of AML (Fig. 2C).

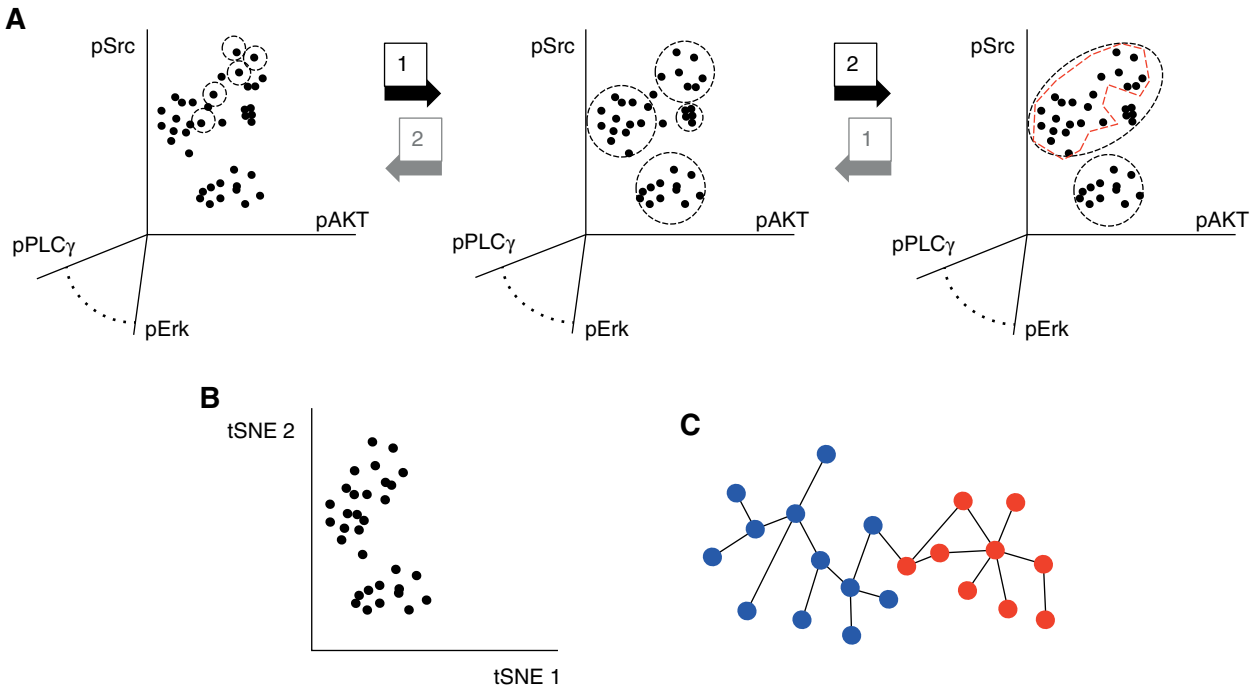


**FIGURE 2** **Clustering approaches for analyzing single cell data**. (**A**) *Agglomerative and divisive clustering*. Schematic diagram showing single cells (black dots) in the space (>3 dimensions) of measured protein abundances (e.g., pSrc, pPLC$\gamma$, pAkt, pErk) in CyTOF experiments. An agglomerative clustering method starts by assigning a unique cluster to a single cell (left most panel). Next, the most similar (e.g., clusters separated by the lowest Euclidean distance) clusters are merged into a larger cluster. This operation is carried out continuously (e.g., left to right, black arrows) until single cells separated by similarity scores lower than or equal to a user defined threshold value are binned into the same cluster. The software package SPADE follows agglomerative clustering of the CyTOF data. Divisive clustering method follows the above steps in the reverse order (right to left, gray arrows). (**B**) Data visualization methods such as vi-SNE generates a lower dimensional (e.g., 2 dimensional) representation of the high-dimensional dataset (left panel in (**A**)). (**C**) Single cells in CyTOF dataset can be clustered in shapes that are nonconvex (e.g., the crescent moon, or the cluster outlined with red dashed line in the rightmost panel in (**A**)). Well defined clustering methods such as k-means clustering[75] or kernel clustering (used in ACCENSE) assume convex shapes (e.g., an ellipse) for the clusters. k-means clustering partitions the data in k number of clusters following a specific optimization procedure (e.g., Voronoi tessellation). PhenoGraph circumvents this issue by creating a graph where single cells in (**A**) (left panel) are represented by the vertices of a graph. A pair of vertices is connected by an edge when the separation between the single cells is below a threshold value. The vertices in this graph are then clustered following a divisive clustering method proposed by Newman and Girvin[76] that uses the "betweenness" metric for the edges to perform the clustering

**TABLE 1** Summary of data analysis and visualization tools for CyTOF data

| Toolkit (Ref) | Clustering algorithm | Cell–cell variation | Biological implications |
|---|---|---|---|
| SPADE[77] | Agglomerative hierarchical clustering | No | SPADE was used in ref. 78 to quantify the diversity of NK cells in humans, to find over 6000 phenotypically different NK cell populations in an individual. The diversity of memory NK cells was also elucidated in ref. 78. |
| PCA/CA[80] | No clustering | Yes | CA was used by ref. 80 to demonstrate that virus (HIV-1, West Nile virus) infected cells help NK cells to diversify. |
| CITRUS[91] | Agglomerative hierarchical clustering | Yes | CITRUS was used by ref. 92 to characterize NK cell phenotype changes with cytokine treatment. |
| viSNE[85] | No clustering | Yes | Ref. 93 used vi-SNE to identify memory NK cells in acute myeloid leukemia patients. |
| ACCENSE[86] | Kernel density-based clustering | Yes | Ref. 86 characterized the heterogeneity in CD8+ T cell population in mice using ACCENSE. |
| PhenoGraph[88] | Community clustering | Yes | Ref. 88 quantified the heterogeneity of leukemia cells in pediatric acute myeloid leukemia patients using PhenoGraph. |

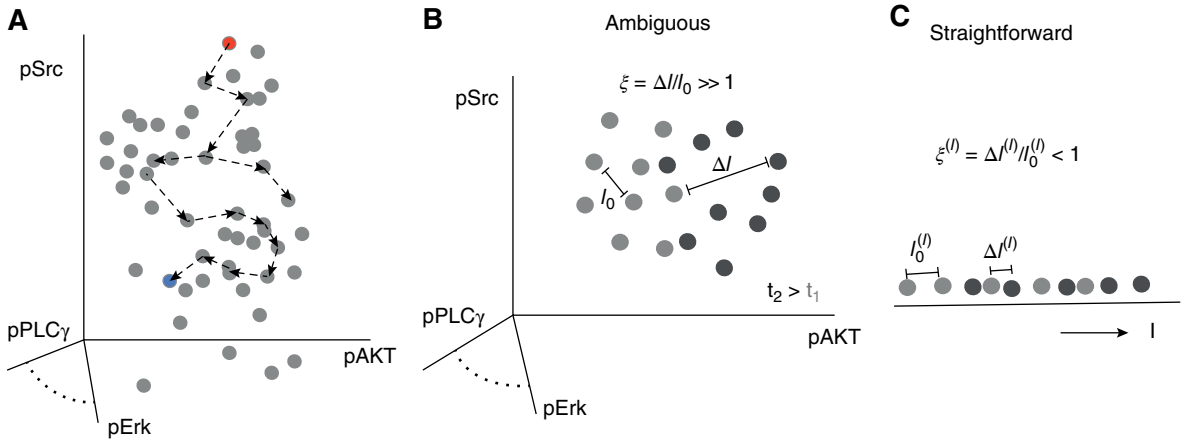


**FIGURE 3** **Reconstruction of single cell developmental or signaling trajectories.** (**A**) Shows a schematic representation of single cells in the space of protein abundances or gene expressions measured in CyTOF or single cell RNA-seq experiments. The cell population can contain single cells residing in different stages of development. The trajectory reconstruction algorithms attempt to determine potential trajectories (dashed lines) across development stages that single cells pass through. This is an example of the type 1 reconstruction as described in the main text. (**B**) Another type of reconstruction problem (type 2) can arise when trajectories of signaling kinetics in single cells are reconstructed using cytometry measurements of signaling proteins at multiple time points. In this case, the ordering of cell populations in time (e.g., cells in grey at $t_1 = 16$ min and cells in black at $t_2 = 32$ min) is known, however, single cell signaling trajectories are still unknown since the single cells are not tagged or are destroyed upon measurements. The ratio ($\xi = \Delta I/I_0$) of the length scales, $\Delta I$, the average separation between the cells at $t_1$ and $t_2$, and $I_0$, the average distance between the cells at $t_1$, determine the difficulty ($\xi < 1$ being straightforward, and $\xi > 1$ being challenging) in connecting the single cells across time. (**C**) Mukherjee et al.[99] address the challenge in (**B**) by describing the signaling kinetics using soft or invariant variables where a difficult reconstruction problem in the original variables ($\xi > 1$ in (**B**)) can turn into a straightforward reconstruction (e.g., connect cell pairs that minimize the total Euclidean distance) problem in the space of new variables I

Single-cell RNA sequencing (sc-RNA-seq) is becoming a popular method to characterize genetic diversity in single cells.[89] Sc-RNA-seq can quantify 10–30 genes in single cells and droplet-based technologies can identify 1000–3000 genes per single cell.[89] The clustering and visualization methods described above can be used to analyze these datasets as well. A challenge in sc-RNA-seq is to separate systematic measurement errors and biological noise in the cell–cell variations of the data.[90] Crinier et al.[71] performed sc-RNA-seq measurements in NK cells from different organs of mice and humans and quantified organ-specific similarities between the two species using PCA and tSNE visualizations and hierarchical clustering.

## 3.1 | Construction of single-cell trajectories

Gene regulatory processes that determine development of leukocytes depend on single-cell abundances of transcription factors and signaling proteins, as well as on intrinsic stochastic fluctuations of regulatory processes, and, signals generated in the local microenvironment. Thus, the trajectory of development initiated by a single progenitor cell can be unique, but nevertheless follows well-defined development stages. Investigation of these single-cell trajectories can provide cues regarding mechanisms that underlie differentiation and commitment to particular lineages in leukocytes. Similarly, signaling kinetics in single-cells is also affected by protein abundances in individual cells, and stochastic fluctuations intrinsic to biochemical signaling reactions. Measurement of signaling kinetics in individual cells can provide insights regarding the presence of fold change in activation, role of signal duration in mediating a specific response, and the relation between peak values of specific activation markers and downstream responses.

Single-cell mass cytometry and single-cell RNA sequencing of a cell population provide snapshot data regarding single cell expressions of proteins and RNAs, respectively, that can be categorized into 2 types (Fig. 3A and B): (1) Single cells residing at different stages of development are assayed under a particular condition. (2) A population of a specific cell type (e.g., immature CD56$^{bright}$ NK cells) is assayed at multiple time points following a specific type of stimulation. Construction of single-cell trajectories needs to address different challenges depending on the type of the measurement. For type (1) data, precursor, intermediate-stage, and end-stage cell types can be identified based on specific markers; however, the progression of the development across a range of single-cell phenotypes is not known beforehand (Fig. 3A). Thus, a computational approach is needed to order single cells according to their status in the course of development based on the available measured proteins or RNA sequences. In contrast, for the data in type (2), the temporal ordering of the single cells is known; however, the challenge is to connect single cells across successive temporal measurements. Several computational approaches have been developed in recent years to address these challenges.

Computational algorithms that assume a continuous change in phenotypes (e.g., abundance of specific proteins and transcription factors) during the course of development have been proposed to generate single-cell developmental trajectories (summarized in Table 2). These algorithms start from a precursor cell and then identify neighboring cells that lie at the immediate next stage of development. The progression of development from a precursor cell to an end-state is built by connecting single cells with appropriate neighbors. The algorithms use several methods to identify single cells in the neighborhood of a precursor cell. One of the algorithms, Monocle, first projects the

**TABLE 2** Summary of computational approaches for construction of single-cell trajectories

| Toolkit/Ref | Distance metric | Characteristics of the reconstructed trajectory | Cell–cell variation | Lower dimensional visualization |
|---|---|---|---|---|
| Wanderlust[95] | Cosine | Branching not allowed | No | No |
| Monocle[94] | Euclidean | Branching allowed | Yes | Yes |
| Wishbone[102] | Euclidean | Branching allowed | No | Yes |
| SCUBA[100] | Euclidean | Branching allowed | Yes | Yes |
| Mukherjee et al.[99] | Euclidean | No branching allowed | Yes | Yes |

high-dimensional dataset into lower dimensions using independent component analysis,[82] and then determines the trajectory by treating single cells as the vertices of a graph.[94] The Euclidean distances between the abundances of cell pairs denote the weights of the edges joining the vertices. The developmental trajectory is then calculated by finding the minimum spanning tree graph.[94] A tree graph is a graph without loops where any two vertices are connected, and the minimum spanning tree graph possesses the minimum value for the sum of the edge weights among all possible spanning tree graphs. Another algorithm, Wanderlust, creates a developmental trajectory by considering graphs in dimension of the measurement where single cells are considered as vertices of a graph where the weights (or length) of the edges connecting the vertices are given by the Cosine distance.[95] The developmental trajectory is determined by finding the shortest path that connects the precursor cell to the end-state. Kared et al.[96] used Wanderlust was to track development of immature CD56$^{bright}$ NK cells in individuals infected with HCMV using CyTOF measurements. The study found NK cells from HCMV-infected individuals showed an early loss of CD62L accompanied with an up-regulation of NKG2C, CD57, CD85j, and Tim-3.[96]

The challenge of trajectory reconstruction posed by the type (2) measurements (Fig. 3B) are dealt with in many disciplines, for example, in physics, tracking fluid particles from snapshot data obtained from microscopy experiments[97] or, in computer science, tracking individuals from snapshot data obtained from video feeds.[98] The difficulty in tracking individual objects in snapshot data is characterized by a dimensionless parameter, $\xi = \Delta l/l_0$. Here, $\Delta l$ is the average distance an object moves between two successive time recordings ($t_1$ and $t_2$), $l_0$ ($= \rho^{-1/d}$) is the average separation between the objects that are distributed with a density $\rho$ in $d$ dimensions. When $\xi \ll 1$, connecting the objects across time is straightforward, whereas when $\xi \gg 1$, matching problem becomes ambiguous. Most of the single-cell data fall in the $\xi \gg 1$ regime.[99] The reconstruction of signaling kinetics or developmental trajectories in single cells comes with several unique challenges. (1) The same single cell is present only once. This is different from particle or individual tracking where the same object can be present in multiple time frames. (2) The progression in kinetics during signaling or developments occurs in a much higher dimension (~40) compared to traditional pair-matching problems dealing with individuals or particles (~ 2–3 dimensions). (3) Sources of randomness are unique in signaling or gene regulatory processes where the randomness often cannot be described by a white gaussian noise. Marco et al.[100] developed an approach (SCUBA) to determine events of multi-lineage differentiation using gene expression and sc-RNA-seq data measured in single cells in mouse embryo at very early stages of development. The algorithm uses k-means clustering to cluster gene expressions at a particular time point and assign a lineage tree to the clustered data. Mukherjee et al.[100] proposed a fundamentally different approach by projecting the data in lower dimensions spanned by 'slow' or invariant variables that change at a much slower rate compared to that of the original variables. This transformation posed the original problem in a space of new variables (invariant or slow variables $I$) that do not change ($\Delta I^{(I)} = 0$) or change more slowly ($\Delta I^{(I)} \to 0$) with time, while still varying appreciably between the objects at a fixed time point. The matching problem cast in terms of the new variables will result in a substantial reduction in the parameter $\xi^{(I)}$ ( $= \Delta I^{(I)}/l_0{}^{(I)}$) and can even fall in the range ($\xi^{(I)} \ll 1$) where tracking objects is straightforward (Fig. 3C). The method was applied on synthetic CyTOF data generated for in silico signaling networks and produced excellent to reasonable reconstructions for a range of conditions. Fundamental constraints on reconstructing trajectories and inferring underlying models using snapshot data were studied by Weinreb et al.[101] using a physical flux balance law.

## 4 | DATA-DRIVEN MODELS WITH MECHANISTIC INSIGHTS

Reduced models to glean mechanistic insights are usually constructed intuitively. This approach works well when the processes of interest are well characterized or questions of general nature are addressed in the modeling study. However, many of the signaling reactions in NK cell signaling and activation are not well characterized. Furthermore, the existence of multiple pathways to produce activation can make it difficult to choose pathways for specific NK cell simulation. In the absence of these details it can be challenging to set up a mechanistic model because of the large number of possibilities for constructing reaction rules between the interacting components. Data-driven models that characterize the data in large dimensions obtained by single-cell technologies are usually statistical in nature and do not have the capability of the reduced models to investigate mechanistic hypothesis with molecular details. Several recent studies have attempted to bridge this gap. Krishnaswamy et al.[103] developed an information theory-based method to analyze signal processing in naïve and antigen-exposed T cells. The approach estimated a conditional probability distribution function ($P(y|x)$), which described the probability of finding a single cell with a protein Y with an abundance $y$ when another the abundance of another protein X is fixed to a value of $x$. The dependence of the conditional probability $P(y|x)$ on the values of $x$ determine the effect of

the protein species X on Y. An information metric was calculated using $P(y|x)$, which vanishes when $x$ and $y$ are independent of each other and increases as $y$ becomes dependent on $x$. Application of the method on CyTOF measurements in $CD4^+$ T cells showed that upon stimulation by crosslinking CD3, CD28, and CD4, the influence of pErk on pS6 increases with time in both naïve and memory $CD4^+$ T cells; however, the magnitude of the influence was larger in the naïve compared to that of the memory T cells. This prediction was further tested in T cells obtained from Erk-knockout mice.

Mukherjee et al.[41] developed a novel data-driven approach where the CyTOF measured signaling kinetics in a particular time interval are effectively described by a system of coupled first-order chemical reactions. In the model, the estimated reaction rates and the associated flux of molecules between pairs of proteins describe the strengths of the effective causal interactions between pairs of molecular species. The framework has several advantages: (1) It provides an effective mechanistic description of the signaling kinetics in a time interval. (2) The effective kinetics separates the contributions of basal (tonic) signaling versus IL-2 pre-treatment or priming and the receptor (e.g., CD16 or NKG2D)-induced signaling kinetics in single-cell protein abundances post-receptor stimulation. (3) The model kinetics can be solved analytically in a closed expression, thus allowing for precise estimation of the rate constants. The available CyTOF data analyses methods[103] are unable to provide the above properties, e.g., property #2. This method was applied to analyze signaling kinetics in immature $CD56^{bright}$ and mature $CD56^{dim}$ NK cells stimulated by NKG2D antibodies. The in silico analysis of fluxes between the protein pairs showed a predicted involvement of CD45 in inducing large changes in the signaling pathways in the IL-2-treated $CD56^{bright}$ NK cells. IL-2 treatment increased the abundance of CD45 in $CD56^{bright}$ NK cells by ~2-fold, which led to stronger Src kinase activation, resulting in increased Erk activation and CD107a mobilization after NKG2D stimulation. Thus, a prediction from this mechanism is that the IL-2-treated immature $CD56^{bright}$ NK cells possessing CD45 abundances closer to that of control media-treated $CD56^{bright}$ NK cells will display substantially less amounts of CD107a on the cell surface following NKG2D stimulation. This prediction was tested by comparing the IL-2-treated $CD56^{bright}$ NK cells against low and high CD45 expression at *a later time-point* ($t = 256$ min, *not used in model training*) after NKG2D stimulation and found that the NK cells with lower CD45 abundances indeed displayed less amounts of CD107a on the cell surface. Another model prediction tested was that both IL-2- and media-treated NK cells produce similar amounts of pErk if CD45-mediated Erk activation pathway is bypassed. PMA + ionomycin stimulation bypasses the need for Src kinases for Erk activation and we found that both IL-2-treated and control media-treated NK cells showed similar increases of pErk.

## 5 | TOWARD THEORY

The central framework for explaining NK cell activation or tolerance, known as the missing-self hypothesis, was proposed by Kärre.[104,105] The original form of the missing-self hypothesis was proven to be too simplistic and was generalized by Lanier and others.[106] In the recent years, several experiments, in particular, the absence of responsiveness of NK cells that lack any self-MHC inhibitory receptor during development, has contradicted the missing-self hypothesis. Based on the recent experimental results Pradeu et al.[107,108] proposed a discontinuity theory, which stated that NK cells (and immune cells in general) respond to a discontinuous change in external signal but become tolerized to signals changing continuously. The authors set up a mathematical description that expressed an output variable denoting the activation of immune cells as a sigmoidal function of an input variable representing an external signal. The output at any instance of time depended on the sum of the changes in the input variable over a time interval in the past. The mathematical relationship reproduced the basic conditions of the theory, namely, increase in the output when the input variable changed abruptly, and decay of the output for a continuous change in the input variable. As expected from a theoretical framework, their theory makes several testable predictions, for example, the chronic activation of the immune system in autoimmune disorders is generated from the change in auto-antigens during the course of the illness. The mathematical tools used in developing this theory precisely quantify the basic propositions; however, the predictions made from the theory in Pradeu et al.[107,108] do not depend on the non-trivial derivations of the mathematical formulation. In his autobiography, Darwin commented, "I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics; for men thus endowed seem to have an extra sense".[109] Obtaining this "extra sense" should be the aspiration of modelers striving to generate theories for NK cell biology and in general biological systems.

## 6 | FUTURE DIRECTIONS

The studies reviewed here demonstrated that mathematical and computational tools help analyze complex data, glean mechanisms, create mathematical models, and generate theoretical frameworks in NK cell biology. What is the future of application of quantitative approaches in NK cell biology? We think the future of this area of research will be exciting, and as we discuss below, computational and mathematical methods can generate transformative results in several areas of NK cell biology in the coming years.

Development of mechanistic computational and mathematical models for describing signal integration and activation in single NK cells stimulated by diverse ligands cognate to activating and inhibitory NKRs, and adhesion receptors will help us comprehend NK cell activation and tolerance beyond the missing-self hypothesis, which has been found to be too simplistic to describe NK cell activation in many recent experiments, for example, NK cell activation due to changes in peptide repertoire.[54] Furthermore, super resolution and confocal microscopy experiments have demonstrated nontrivial changes in spatial reorganization of NKRs,[110–113] NKR-associated signaling proteins,[112] cytokine receptors,[114] and coordinated reorganization of cytoskeletal elements[115] and transport of cytolytic granules[115,116] during NK cell signaling and activation. Quantitative models with predictive powers that account for the above diverse interactions as well as spatial changes could potentially provide valuable mechanistic insights that

underlie activation of different types NK cells such as educated,[117] uneducated, or "memory" NK cells.[118] In silico predictions from such models regarding outcomes of specific NK cells interacting with target cells expressing designed ligands can be relevant for vaccine development.[119] Importantly, a better understanding of these NK cell activation mechanisms is relevant to understanding why NK cells tolerate the existence of tumors in cancer patients that have lost MHC class I—and theoretically should be eliminated by "miss-self" recognition.

Diverse NK cell populations (e.g., educated or uneducated NK cells, "memory" NK cells) appear to be relevant for controlling viral infections such as HCMV infection or proliferation of tumor cells or graft rejection. High-dimensional datasets such as RNA-seq data describing changes in NK cell gene expressions[16,72,120] NKR gene sequences,[72,121] and protein expressions in single NK cells (e.g., CyTOF)[70,72,122] provide detailed description regarding the involvement of NK cell populations in immune responses elicited by a viral infection or by an organ transplant. Data-driven models are being increasingly applied to develop predictive computational frameworks to determine precision biomarkers in patients for optimizing NK cell[123] and T cell responses[124] in immunotherapies or organ transplants. However, a key challenge in this endeavor is to know what measured variables should be included in such computational models. In addition, these measurements still probe a small fraction of the complex NK cell response in these patients; therefore, knowing what additional variables need to be measured so that the models can be better trained is another important question. The difficulty in addressing the above challenges arise due to the lack of a principled framework for combining the diverse datasets that are acquired at different scales (e.g., single cells, organs, human subjects) and time points (e.g., different days after an organ transplant). Similar problems in modeling complex datasets are encountered in diverse areas such as protein structure prediction (https://deepmind.com/blog/alphafold/) or modeling the climate.[125] Recently developed computational tools have made substantial progress in generating precise predictions in these complex systems, and some of these tools will hopefully be useful for analyzing the above questions in NK cell biology.

NK cell response in an individual, as discussed in the Introduction, involves a wide range of scales. At present, we have mechanistic models that are developed to describe NK cell response in a particular scale (e.g., single cell or cell populations). Since the processes in these scales interact seamlessly during a host response against an infection or a tumor it will be essential to combine these models in order to gain a mechanistic understanding into the role of NK cells within the host response. Data-driven learning models combine data from these different scales,[126] but provide little or no mechanistic interpretation. Thus, a challenge for quantitative researchers is to be able take the advantages of the rich multidimensional datasets and the learning algorithms and create approaches to delineate mechanisms that underlie complex NK cell responses. Similar challenges of integrating models spanning wide range of scales exist in different disciplines such as statistical physics[127] and materials science.[128] Borrowing quantitative tools from these areas and combining those to the existing mechanistic and data-driven models in NK cell biology can generate novel solutions.

## AUTHORSHIP

J.D. and L.L.L. contributed to the scientific content and wrote the manuscript.

## ACKNOWLEDGEMENTS

This work was partially supported by the Grant R56AI108880-01 from NIAID and support from the Research Institute at the Nationwide Children's Hospital to J.D. J.D. thanks Veronica J. Vieland for a critical reading of the manuscript.

## DISCLOSURES

The authors declare no conflicts of interest.

## REFERENCES

1. Kuhn TS. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press; 1964.
2. Dyson F. The Key to Everything. *The New York Review of Books*. 2018. https://www.nybooks.com/articles/2018/05/10/the-key-to-everything/
3. Goldstein RE. Point of view: Are theoretical results 'Results'? *eLife*. 2018;7:e40018.
4. Barenblatt GI. *Scaling*. Cambridge, UK: Cambridge University Press; 2003.
5. Lanier LL. Up on the tightrope: natural killer cell activation and inhibition. *Nat Immunol*. 2008;9:495–502.
6. Sun JC, Lanier LL. NK cell development, homeostasis and function: parallels with CD8+ T cells. *Nat Rev Immunol*. 2011;11:645.
7. Barrett JW, Garcke H, Nürnberg R. Numerical computations of faceted pattern formation in snow crystal growth. *Phys Rev E*. 2012; 86:011604.
8. Anderson PW. More is different. *Science*. 1972;177:393–396.
9. Laughlin RB, Pines D. The theory of everything. *Proc Natl Acad Sci*. USA 2000;97:28–31.
10. Nelson PC, Bromberg S, Hermundstad A, Prentice J. *Physical Models of Living Systems*. New York, NY: WH Freeman; 2015.
11. Phillips R. Musings on mechanism: quest for a quark theory of proteins? *FASEB J*. 2017;31:4207–4215.
12. Nurse P, Hayles J. The cell in an era of systems biology. *Cell*. 2011;144:850–854.
13. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*. 1999;402:C47.
14. Scheiter M, Lau U, van Ham M, et al. Proteome analysis of distinct developmental stages of human natural killer cells. *Mol Cell Proteom*. 2013;M112:024596.
15. Horowitz A, Strauss-Albee DM, Leipold M, et al. Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. *Sci Transl Med*. 2013;5:208ra145.
16. Rieckmann JC, Geiger R, Hornburg D, et al. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol*. 2017;18:583.
17. Kim M-S, Pinto SM, Getnet D, et al. A draft map of the human proteome. *Nature*. 2014;509:575.

18. Council NR. *The Role of Theory in Advancing 21st-Century Biology: Catalyzing Transformative Research*. Washington, DC: National Academies Press; 2008.
19. Shou W, Bergstrom CT, Chakraborty AK, Skinner FK. Theory, models and biology. *Elife*. 2015;4:e07158.
20. Hawking S. *On the Shoulders of Giants: The Great Works of Physics and Astronomy*. Philadelphia, PA: Running Press; 2002.
21. Darwin C. *On the Origin of Species, 1859*. London: Routledge; 2004.
22. Burnet SFM. *The Clonal Selection Theory of Acquired Immunity*. Cambridge, UK: The Cambridge University Press; 1959.
23. Servedio MR, Brandvain Y, Dhole S, et al. Not just a theory—the utility of mathematical models in evolutionary biology. *PLoS Biol*. 2014;12:e1002017.
24. Jerne N. Clonal selection in a lymphocyte network. cellular selection and regulation in the immune response. *J Soc Gen Physiol*. 1974: 39–48.
25. Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol*. 1979;81:645–670.
26. Goldenfeld N, Kadanoff LP. Simple lessons from complexity. *Science*. 1999;284:87–89.
27. Chakraborty AK, Das J. Pairing computation with experimentation: a powerful coupling for understanding T cell signalling. *Nat Rev Immunol*. 2010;10:59.
28. Lever M, Maini PK, Van Der Merwe PA, Dushek O. Phenotypic models of T cell activation. *Nat Rev Immunol*. 2014;14:619.
29. Goldstein B, Faeder JR, Hlavacek WS. Mathematical and computational models of immune-receptor signalling. *Nat Rev Immunol*. 2004;4:445.
30. Das J, Jayaprakash C. *Systems Immunology: An Introduction to Modeling Methods for Scientists*. Boca Raton, FL: CRC Press; 2017.
31. Chakraborty AK. A perspective on the role of computational models in immunology. *Ann Rev Immunol*. 2017;35:403–439.
32. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser ID. Systems biology in immunology: a computational modeling perspective. *Ann Rev Immunol*. 2011;29:527–585.
33. Gunawardena J. Models in biology: Accurate descriptions of our pathetic thinking. *BMC Biol*. 2014;12:29.
34. Das J. Physical models in immune signaling. In *Systems Immunology*. Boca Raton, FL: CRC Press; 2018:227–250.
35. Vance R, Raulet D. Toward a quantitative analysis of the repertoire of class I MHC-specific inhibitory receptors on natural killer cells. In *Specificity, Function, and Development of NK Cells*. Berlin, Heidelberg: Springer 1998:135–160.
36. Johansson S, Salmon-Divon M, Johansson MH, et al. Probing natural killer cell education by Ly49 receptor expression analysis and computational modelling in single MHC class I mice. *PLoS One*. 2009;4:e6046.
37. Andersson S, Fauriat C, Malmberg J-A, Ljunggren H-G, Malmberg K-J. KIR acquisition probabilities are independent of self-HLA class I ligands and increase with cellular KIR expression. *Blood*. 2009;114: 95–104.
38. Das J. Activation or tolerance of natural killer cells is modulated by ligand quality in a nonmonotonic manner. *Biophys J*. 2010;99: 2028–2037.
39. Mesecke S, Urlaub D, Busch H, Eils R, Watzl C. Integration of activating and inhibitory receptor signaling by regulated phosphorylation of Vav1 in immune cells. *Sci Signal*. 2011;4:ra36.
40. Carrillo-Bustamante P, Keşmir C, De Boer RJ. Virus encoded MHC-like decoys diversify the inhibitory KIR repertoire. *Plos Comput Biol*. 2013;9:e1003264.
41. Mukherjee S, Jensen H, Stewart W, et al. In silico modeling identifies CD45 as a regulator of IL-2 synergy in the NKG2D-mediated activation of immature human NK cells. *Sci Signal*. 2017;10.
42. Harris LA, Hogg JS, Tapia J-J, et al. BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics*. 2016;32:3366–3368.
43. Lis M, Artyomov MN, Devadas S, Chakraborty AK. Efficient stochastic simulation of reaction-diffusion processes via direct compilation. *Bioinformatics*. 2009.
44. Angermann BR, Klauschen F, Garcia AD, et al. Computational modeling of cellular signaling processes embedded into dynamic spatial contexts. *Nat Methods*. 2012;9:283.
45. Loew LM, Schaff JC. The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol*. 2001;19:401–406.
46. Das J, Ho M, Zikherman J, et al. Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell*. 2009;136:337–351.
47. Raue A, Kreutz C, Maiwald T, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*. 2009;25:1923–1929.
48. Mitra ED, Dias R, Posner RG, Hlavacek WS. Using both qualitative and quantitative data in parameter identification for systems biology models. *Nat Commun*. 2018;9:3901.
49. Myers CR. Zen and the art of parameter estimation in systems biology. In *Systems Immunology*. Boca Raton, FL: CRC Press; 2018: 123–138.
50. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys*. 2015;143:07B201_1.
51. Goldbeter A. An amplified sensitivity arising from covalent modification in biological systems. *Proc Natl Acad Sci USA*. 1981;78: 6840–6844.
52. Makaryan SZ, Finley SD. Modeling of CD16, 2B4 and NKG2D stimulation in natural killer cell activation. *bioRxiv*. 2018:395756.
53. Treanor B, Lanigan PM, Kumar S, et al. Microclusters of inhibitory killer immunoglobulin-like receptor signaling at natural killer cell immunological synapses. *J Cell Biol*. 2006;174:153–161.
54. Das J, Khakoo SI. NK cells: tuned by peptide?. *Immunol Rev*. 2015;267:214–227.
55. Watzl C, Sternberg-Simon M, Urlaub D, Mehr R. Understanding natural killer cell regulation by mathematical approaches. *Front Immunol*. 2012;3:359.
56. Mbiribindi B, Mukherjee S, Wellington D, Das J, Khakoo SI. Spatial clustering of receptors and signaling molecules regulates NK cell response to peptide repertoire changes. *Front Immunol*. 2019;10:605.
57. Merrill SJ. A model of the role of natural killer cells in immune surveillance—I. *J Math Biol*. 1981;12:363–373.
58. Wodarz D, Sierro S, Klenerman P. Dynamics of killer T cell inflation in viral infections. *J R Soc Interface*. 2006;4:533–543.
59. Elemans M, Thiébaut R, Kaur A, Asquith B. Quantification of the relative importance of CTL, B cell, NK cell, and target cell limitation in the control of primary SIV-infection. *Plos Comput Biol*. 2011;7:e1001103.
60. Bauer AL, Beauchemin CA, Perelson AS. Agent-based modeling of host–pathogen systems: the successes and challenges. *Inf Sci*. 2009;179:1379–1389.
61. Murray JD. *Mathematical Biology*. Berlin; New York: Springer-Verlag; 1989.

62. Smith AM, Ribeiro RM, Perelson AS. Population dynamics of host and pathogens. In *Systems Immunology*. Boca Raton, FL: CRC Press; 2018:265–278.
63. Khorshidi MA, Vanherberghen B, Kowalewski JM, et al. Analysis of transient migration behavior of natural killer cells imaged in situ and in vitro. *Integr Biol*. 2011;3:770–778.
64. Lee BJ, Mace EM. Acquisition of cell migration defines NK cell differentiation from hematopoietic stem cell precursors. *Mol Biol Cell*. 2017;28:3573–3581.
65. Beltman JB, Marée AF, De Boer RJ. Analysing immune cell migration. *Nat Rev Immunol*. 2009;9:789.
66. Carrillo-Bustamante P, Keşmir C, de Boer RJ. The evolution of natural killer cell receptors. *Immunogenetics*. 2016;68:3–18.
67. Sun J, Lanier L. The natural selection of herpesviruses and virus-specific NK cell receptors. *Viruses*. 2009;1:362–382.
68. Carrillo-Bustamante P, Keşmir C, de Boer RJ. A coevolutionary arms race between hosts and viruses drives polymorphism and polygenicity of NK cell receptors. *Mol Biol Evol*. 2015;32:2149–2160.
69. Carrillo-Bustamante P, de Boer RJ, Keşmir C. Specificity of inhibitory KIRs enables NK cells to detect changes in an altered peptide environment. *Immunogenetics*. 2018;70:87–97.
70. Wilk AJ, Blish CA. Diversification of human NK cells: lessons from deep profiling. *J Leuk Biol*. 2018;103:629–641.
71. Crinier A, Milpied P, Escalière B, et al. High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice. *Immunity*. 2018;49:971–986. e5.
72. Vento-Tormo R, Efremova M, Botting RA, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*. 2018;563:347.
73. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. *Cell*. 2016;165:780–791.
74. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. *Trends Immunol*. 2012;33:323–332.
75. MacKay DJ, Mac Kay DJ. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press; 2003.
76. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E*. 2004;69:026113.
77. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. 2011;29:886.
78. Horowitz A, Strauss-Albee DM, Leipold M, et al. Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. *Sci Transl Med*. 2013;5:208ra145–208ra145.
79. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity*. 2012;36:142–152.
80. Strauss-Albee DM, Fukuyama J, Liang EC, et al. Human NK cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Sci Transl Med*. 2015;7:297ra115–297ra115.
81. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. *Trends Immunol*. 2012;33:323–332.
82. Izenman AJ. *Modern multivariate statistical techniques*. In *Regression, classification and manifold learning*. New York: Springer-Verlag; 2008.
83. Dworkin M, Mukherjee S, Jayaprakash C, Das J. Dramatic reduction of dimensionality in large biochemical networks owing to strong pair correlations. *J R Soc Interface*. 2012;9:1824–1835.
84. Maaten Lv, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2605.
85. Amir E-AD, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*. 2013;31:545.
86. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci USA*. 2014;111:202–207.
87. Anchang B, Hart TD, Bendall SC, et al. Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat Protoc*. 2016;11:1264.
88. Levine JH, Simonds EF, Bendall SC, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. 2015;162:184–197.
89. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18:35.
90. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*. 2017;19:562–578.
91. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA*. 2014;111:E2770–E2777.
92. Vendrame E, Fukuyama J, Strauss-Albee DM, Holmes S, Blish CA. Mass cytometry analytical approaches reveal cytokine-induced changes in natural killer cells. *Cytom Part B Clin Cytom*. 2017;92:57–67.
93. Romee R, Rosario M, Berrien-Elliott MM, et al. Cytokine-induced memory-like natural killer cells exhibit enhanced responses against myeloid leukemia. *Sci Transl Med*. 2016;8:357ra123–357ra123.
94. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32:381.
95. Bendall SC, Davis KL, Amir E-AD, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*. 2014;157:714–725.
96. Kared H, Martelli S, Tan SW, et al. Adaptive NKG2C+ CD57+ Natural Killer cell and Tim-3 expression during viral infections. *Front Immunol*. 2018;9.
97. Ouellette NT, Xu HT, Bodenschatz E. A quantitative study of three-dimensional Lagrangian particle tracking algorithms. *Exp Fluids*. 2006;40:301–313.
98. Lipton AJ, Fujiyoshi H, Patil RS. Moving target classification and tracking from real-time video. Fourth IEEE Workshop on Applications of Computer Vision - Wacv'98. *Proceedings*. 1998:8–14.
99. Mukherjee S, Stewart D, Stewart W, Lanier LL, Das J. Connecting the dots across time: reconstruction of single-cell signalling trajectories using time-stamped data. *R Soc Open Sci*. 2017;4:170811.
100. Marco E, Karp RL, Guo G, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci USA*. 2014;111:E5643–E5650.
101. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci USA*. 2018:201714723.
102. Setty M, Tadmor MD, Reich-Zeliger S, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol*. 2016;34:637.
103. Krishnaswamy S, Spitzer MH, Mingueneau M, et al. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. *Science*. 2014;346:1250689.
104. Lanier LL. Missing self, NK cells, and the white album. *J Immunol*. 2005;174:6565–6565.

105. Kärre K. NK cells, MHC class I molecules and the missing self. *Scand J Immunol*. 2002;55:221–228.

106. Lanier LL, Corliss B, Phillips JH. Arousal and inhibition of human NK cells. *Immunol Rev*. 1997;155:145–154.

107. Pradeu T, Jaeger S, Vivier E. The speed of change: towards a discontinuity theory of immunity?. *Nat Rev Immunol*. 2013;13:764.

108. Pradeu T, Vivier E. The discontinuity theory of immunity. *Sci Immunol*. 2016;1.

109. Darwin C, Darwin F. *Autobiography and selected letters*. Chelmsford, MA: Courier Corporation; 1958.

110. Borhis G, Ahmed PS, Mbiribindi B, et al. A peptide antagonist disrupts NK cell inhibitory synapse formation. *J Immunol*. 2013;190:2924–2930.

111. Oszmiana A, Williamson DJ, Cordoba S-P, et al. The size of activating and inhibitory killer Ig-like receptor nanoclusters is controlled by the transmembrane sequence and affects signaling. *Cell Rep*. 2016;15:1957–1972.

112. Treanor B, Lanigan PM, Kumar S, et al. Microclusters of inhibitory killer immunoglobulin–like receptor signaling at natural killer cell immunological synapses. *J Cell Biol*. 2006;174:153–161.

113. Pageon SV, Cordoba S-P, Owen DM, Rothery SM, Oszmiana A, Davis DM. Superresolution microscopy reveals nanometer-scale reorganization of inhibitory natural killer cell receptors upon activation of NKG2D. *Sci Signal*. 2013;6:ra62–ra62.

114. Bálint, Lopes FB, Davis DM. A nanoscale reorganization of the IL-15 receptor is triggered by NKG2D in a ligand-dependent manner. *Sci Signal*. 2018;11:eaal3606.

115. Mace EM, Dongre P, Hsu HT, et al. Cell biological steps and checkpoints in accessing NK cell cytotoxicity. *Immunol Cell Biol*. 2014;92:245–255.

116. Carisey AF, Mace EM, Saeed MB, Davis DM, Orange JS. Nanoscale dynamism of actin enables secretory function in cytolytic cells. *Curr Biol*. 2018;28:489–502. e9.

117. Boudreau JE, Hsu KC. Natural killer cell education and the response to infection and cancer therapy: stay tuned. *Trends Immunol*. 2018;39:222–239.

118. Cerwenka A, Lanier LL. Natural killer cell memory in infection, inflammation and cancer. *Nat Rev Immunol*. 2016;16:112–123.

119. Sun JC, Lanier LL. Is there natural killer cell memory and can it be harnessed by vaccination? NK cell memory and immunization strategies against infectious diseases and cancer. *Cold Spring Harbor Perspect Biol*. 2018;10:a029538.

120. Robinette ML, Fuchs A, Cortez VS, et al. Transcriptional programs define molecular characteristics of innate lymphoid cell classes and subsets. *Nat Immunol*. 2015;16:306.

121. Strunz B, Hengst J, Deterding K, et al. Chronic hepatitis C virus infection irreversibly impacts human natural killer cell repertoire diversity. *Nat Commun*. 2018;9:2275.

122. Patin E, Hasan M, Bergstedt J, et al. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat Immunol*. 2018;19:302.

123. Erbe AK, Wang W, Carmichael L, et al. Neuroblastoma patients' KIR and KIR-ligand genotypes influence clinical outcome for dinutuximab-based immunotherapy: a report from the Children's Oncology Group. *Clin Cancer Res*. 2018;24:189–196.

124. Łuksza M, Riaz N, Makarov V, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*. 2017;551:517.

125. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. Deep learning and process understanding for data-driven Earth system science. *Nature*. 2019;566:195.

126. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. *Cell*. 2018.

127. Fritz D, Koschke K, Harmandaris VA, van der Vegt NF, Kremer K. Multiscale modeling of soft matter: scaling of dynamics. *Phys Chem Chem Phys*. 2011;13:10412–10420.

128. Bendsøe MP, Sigmund O. *Optimization of Structural Topology, Shape, and Material*. Berlin, Heidelberg, New York: Springer; 1995.

**How to cite this article:** Das J, Lanier LL. Data analysis to modeling to building theory in NK cell biology and beyond: How can computational modeling contribute? *J Leukoc Biol*. 2019;1–13. https://doi.org/10.1002/JLB.6MR1218-505R