Trends in Biotechnology

CellPress REVIEWS

Opinion

# The Importance of Computational Modeling in Stem Cell Research

Antonio del Sol [1,2,3,4,*] and Sascha Jung[2]

**The generation of large amounts of omics data is increasingly enabling not only the processing and analysis of large data sets but also the development of computational models in the field of stem cell research. Although computational models have been proposed in recent decades, we believe that the stem cell community is not fully aware of the potentiality of computational modeling in guiding their experimental research. In this regard, we discuss how single-cell technologies provide the right framework for computational modeling at different scales of biological organization in order to address challenges in the stem cell field and to guide experimentalists in the design of new strategies for stem cell therapies and treatment of congenital disorders.**

## Stem Cell Computational Modeling in the Era of Single-Cell Big Data Generation

Stem cell research has witnessed a revolution in the last two decades after the discovery of induced pluripotent stem cells (iPSCs). This has opened new opportunities for studying human diseases; designing strategies for tissue regeneration, including **cell transplantation** (see Glossary) and developmental research. Rapid advances in single-cell approaches allow a detailed characterization of cellular phenotypes across tissues in different conditions, the discovery of new cellular subpopulations, and the reconstruction of single-cell trajectories in development and reprogramming. In particular, the increasing resolution of single-cell RNA sequencing (scRNA-seq) and the emergence of new technologies that generate other types of single-cell phenotypic omics data, such as epigenomes, proteomes, and spatial information, allow the systematic integration and analysis of these data, leading to a more comprehensive characterization of cell type classification, function, and interactions. Despite technical limitations, such as gene dropouts and low capture rates, the analysis of single-cell data attains high statistical power by considering a large number of individual samples and allows the identification of cellular subpopulations at an unprecedented resolution. Massive generation of these multiomics single-cell data enables the development of high-resolution computational models that are able to capture the collective behavior of genes at the molecular level or cells at the tissue level, thus providing an ideal framework to address key questions in the stem cell field. Indeed, computational models can generate novel predictions and provide new insights into biological mechanisms, guiding experimental research. In particular, systems biology models at different levels of complexity, including cellular, tissue, and even organ levels, can be developed to address relevant questions in stem cell research. For example, on the one hand, models at the cellular level, such as **gene regulatory network (GRN)**–based models, can improve the understanding of cellular differentiation and **cellular conversion** and can help to predict key transcription factors (TFs) and signaling molecules controlling such processes. On the other hand, models at the tissue level, including those based on **cell–cell interaction networks**, can be useful for elucidating general principles of **tissue homeostasis** and regeneration and for generating predictions of relevant cell–cell interaction events supporting the tissue regeneration capacity.

## Highlights

In the era of single-cell big data, computational modeling is a powerful tool to describe biological systems and generate predictions across different spatial and temporal scales. The steadily increasing amount of data allows the development of models that link different levels of biological organization, such as intracellular interactions, cellular behavior, and the behavior of cell populations.

The development of single-cell–based mechanistic models is necessary to better characterize biological processes and generate more accurate predictions of cellular conversion factors, cell identity transcription factors, and cell–cell interactions relevant for tissue regeneration and homeostasis.

We expect these models to accelerate the development of novel regenerative medicine strategies by guiding experimentalists in the design of stem cell transplantation and gene correction therapies.

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Esch-sur-Alzette, L-4367 Belvaux, Luxembourg
[2]CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 801 Building, 48160 Derio, Spain
[3]IKERBASQUE, Basque Foundation for Science, Bilbao 48013, Spain
[4]https://wwwen.uni.lu/lcsb/research/computational_biology

*Correspondence:
antonio.delsol@uni.lu (A. del Sol).

In summary, computational models complement statistical data analysis by providing mechanistic insights into biological processes and by generating novel predictions that can guide experimental research. In particular, we believe that a number of biological questions and challenges in the field of stem cell research and regenerative medicine can be addressed with the help of computational models. These models have been shown to be useful in guiding experimental research, particularly in the design of cellular conversion strategies. Nevertheless, the time is ripe for the development of more sophisticated models, considering the continuous improvement and appearance of new single-cell technologies that are currently able to generate large amounts of different types of data. In this regard, we are now able to develop models at different scales of biological complexity, which are more representative of biological processes. For example, a model that considers stem cell–niche interactions is more appropriate for the *in vivo* prediction of **cell fate determinants**. Therefore, these models can be particularly useful in advancing regenerative medicine strategies, such as *in vivo* reprogramming, and stem cell transplantation and rejuvenation by overcoming limitations of current experimental protocols.

## Multiscale Computational Modeling: From Intracellular Networks to Cell–Cell Interactions in the Tissue

Stem cell research comprises performing experiments and developing hypotheses that link different scales of biological organization, including intracellular interactions, cellular behavior, and the behavior of cell populations. The aim of **multiscale computational modeling** is to describe biological systems and generate predictions across these different spatial and temporal scales. The level at which the model should be constructed depends on the scientific question being addressed and the available input data.

For example, computational methods that rely on the reconstruction and analysis of intracellular GRNs have been shown to be useful in modeling cellular conversion, enabling the identification of optimal sets of **conversion factors** (Box 1). Importantly, inference of cell type–specific GRNs is an essential step for these network-based methods; however, it is not always possible to

### Box 1. Applications of Different Modeling Types to Stem Cell Research

The development of computational models can aid in addressing key questions in stem cell research. However, the choice of the modeling framework is highly dependent on the specific research question (see Table 1 in main text).

Boolean network models of GRNs work well for the identification of cell conversion factors. Although these models do not require the inference of kinetic parameters, they usually include regulatory interactions between large numbers of TFs. Moreover, this modeling framework allows the inference of cooperativity among TFs in regulation, enabling the identification of optimal combinations of TFs controlling cellular conversion.

Continuous models of GRNs based on ordinary differential equations are suited for predicting the dynamic behavior of gene expression during biological processes such as cellular differentiation. Indeed, using linear ordinary differential equations is possible to infer the regulatory relationship among several tens of genes from time-series data corresponding, for example, to cellular differentiation. Further, these equations can be employed to predict the continuous dynamic behavior of gene expression in a quantitative manner following perturbation of specific TFs or signaling pathways.

Probabilistic models, such as dynamic Bayesian networks or probabilistic Boolean networks, are adequate to simulate stochasticity in gene expression profiles and regulatory interactions in cellular systems, which is one of the determinants of cellular reprogramming efficiency. Hence, these models can be used to prioritize optimal combinations of TFs whose perturbations could induce cellular reprogramming with high efficiency.

Directed graphs representing cell–cell communication networks can be inferred from ligand–receptor pair expression without the need to estimate receptor–ligand binding affinity. This type of descriptive model is appropriate to identify relevant features of the structure of cell–cell communication networks to describe how cells interact to guarantee tissue function in homeostasis. In addition, following this approach, it is possible to detect relevant cell–cell interactions that can impair tissue regeneration in disease or aging.

### Glossary

**Cell–cell interaction networks:** maps of molecular interactions among ligands and receptors.
**Cell fate determinants:** transcription factors inducing the differentiation of stem/progenitor cells toward a specific daughter cell type.
**Cell identity TFs:** transcription factors defining a particular cell type or subtype.
**Cell transplantation:** a procedure in which cells are transferred to a tissue that is diseased or damaged.
**Cellular conversion:** a process in which one cell (sub)type is transformed into another cell (sub)type.
**Conversion efficiency:** the percentage of successfully transformed cells in cellular conversion experiments.
**Conversion factors:** the activated or inhibited genes to induce a cellular conversion.
**Gene regulatory network (GRN):** a map of molecular interactions among gene products.
**Machine learning:** algorithms used to perform a task that automatically learn from experience without specific instructions.
**Mechanistic models:** a description of a complex system with a tangible physical, chemical, or biological aspect.
**Multiscale computational modeling:** a description of a complex system at different biological levels, such as the cellular, tissue, and organ levels.
**Niche:** a microenvironment in which a cell resides that exerts various effects on it.
**Quiescent stem cell:** the reversible state of a stem cell in which it does not divide but can re-enter proliferation rapidly upon receiving signals from the niche.
**Tissue homeostasis:** an equilibrium state in which key tissue characteristics, such as cellular proliferation or cell death, are maintained within acceptable ranges.
**Transgenic:** a cell whose genome includes an additional gene that was transferred naturally or by genetic engineering.

accurately infer these GRNs, especially for newly characterized cell subtypes. In this regard, scRNA-seq is the ideal technology for capturing real interactions between genes in individual cells. Indeed, scRNA-seq captures the gene expression of thousands of individual cells in one experiment, which provides a large number of independent measurements that allow the extraction of information about gene expression heterogeneity across cells and gene–gene coexpression in individual cells. Hence, scRNA-seq allows the inference of cell type– or cell subtype–specific GRNs [1–11], which constitutes an important step in the implementation of network-based methods for cellular conversion [12,13] (Table 1). In particular, using a **machine learning** approach, it has been possible to identify different GRN configurations corresponding to different states of pluripotency, including naive and primed states [4]. This knowledge is important not only for characterizing the regulatory program of different stem cell phenotypic states but also for devising new GRN-based strategies to direct stem cell conversion. Indeed, the availability of single-cell–based GRN inference methods has led to the optimization of various cell conversion protocols in the context of differentiation, transdifferentiation, and reprogramming [14–17]. For instance, a combination of small molecules was identified to increase the efficiency of embryonic fibroblast to pluripotent stem cell conversion by inferring and analyzing the GRN governing the corresponding reprogramming trajectory [16]. Similarly, single-cell–based network inference in the context of pre–B cell to macrophage transdifferentiation revealed the TF MYC as a crucial determinant of reprogramming efficiency [15]. Moreover, a recent study demonstrated a critical role for ETS1 in cardiomyocyte differentiation based on the reconstruction of GRNs at different developmental time points [17]. Despite the recent progress in single-cell–based GRN inference, integration of scRNA-seq with other types of data, such as single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) and DNA methylation profiling, would allow more accurate inference of cell type– or cell subtype–specific GRNs and therefore would advance conversion factor predictions (Figure 1).

Complementary computational approaches overcome the intrinsic complexity of GRN inference by solely extracting relevant gene expression patterns that identify **cell identity TFs** [18,19]. Upregulation of these target cell identity TFs in the initial cell type can be used as a strategy for cellular conversion [19]. Nevertheless, these methods do not provide details about gene

Table 1. scRNA-Seq-Based Mechanistic Models

| Modeling framework | Model type | Application | Biological significance |
|---|---|---|---|
| Descriptive model | Undirected graph | Inference of gene coexpression networks from time-series data using information theory | Prioritization of functional modules |
| | Directed graph | Inference of cell–cell communication networks from ligand–receptor pair expression | Prediction of tissue cell–cell interactions |
| Logical model | Boolean network | Inference of Boolean model of GRNs based on coexpression modules and TF binding site predictions | Prediction of cell conversion factors |
| | | Inference of Boolean model of GRNs based on Granger causality analysis of time-series data | Prediction of regulatory interactions governing cellular trajectories |
| | | Inference of Boolean logic rules of GRNs | Identification of cooperative TF regulation |
| Continuous model | Regression-based | Inference of quantitative model of GRNs based on ridge regression and partial correlation | Prediction of gene expression dynamics during differentiation |
| | Linear ordinary differential equations | Inference of quantitative model of GRNs based on differentiation trajectories | |
| | | Inference of quantitative model of GRNs based on variation in cell cycle and cell state | |
| Probabilistic model | Dynamic Bayesian network | Inference of probabilistic GRN model based on prior knowledge | Prediction of reprogramming efficiency |

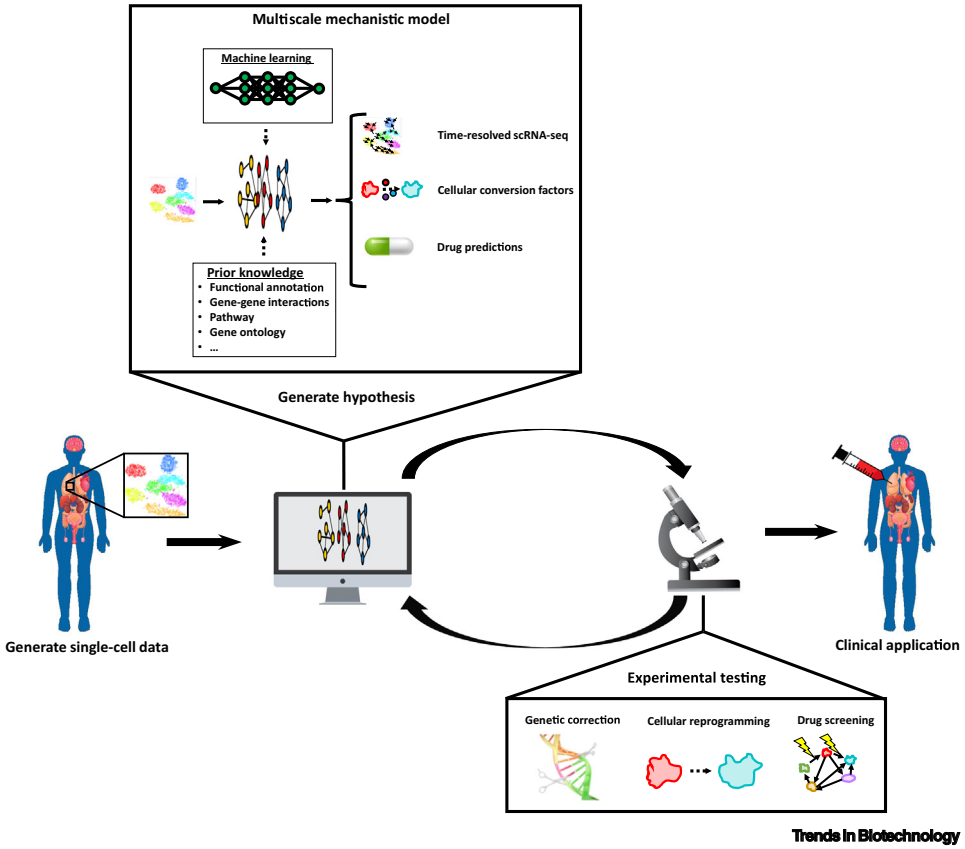


Figure 1. Accelerating the Design of Regenerative Medicine Therapies. Single-cell profiling of different scales enables the development of multiscale mechanistic models that can address key questions in stem cell research. These computational models can make use of prior biological knowledge and machine learning analyses to generate predictions that guide experimental stem cell biologists in the design of strategies for gene correction, cellular conversion, and drug screening. The translation of novel regenerative medicine therapies to the clinics will be accelerated through the feedback between computational and experimental researchers. (Image contains resources from Freepik.) Abbreviation: scRNA seq, single-cell RNA sequencing.

regulation. Combining the predictions of these computational methods with GRN inference would allow the identification of optimal sets of conversion factors.

Modeling stem cell–niche interactions would allow researchers to determine **niche** signals that maintain aberrant stem cell phenotypes in disease or aging and therefore to design strategies for counteracting the niche effect in these cells. To date, a few computational approaches have been developed that explicitly consider niche effects [20–23]. However, most of these models have been assembled manually from experimentally validated interactions and are customized to stem cells in a particular niche. By contrast, recent computational approaches that use scRNA-seq to integrate signaling and GRNs provide a general framework of stem cell–niche interactions and have been able to infer the effect of niche signals on target gene expression [24,25]. In particular, predictions of key signaling molecules that mediate niche signals to maintain cellular phenotypes have been used for cellular rejuvenation by counteracting the effect of the aging niche [26]. However, because these computational methods rely solely on scRNA-seq data for the integration of signaling and transcriptional regulatory networks, the identification of signaling molecules that mediate niche cues in stem cells to maintain their phenotypes remains a challenge. In this regard, combining scRNA-seq with bulk phosphoproteomics data from

purified cell populations would facilitate the integration of signaling and transcriptional regulatory networks. Furthermore, advances in single-cell molecular profiling technologies, such as phosphoproteomics [27,28] and perturbation studies [29], would allow investigators to capture the heterogeneity in niche-induced signaling pathways, enabling the development of higher-resolution integrative network models.

Modeling cell–cell communication mediated by receptor–ligand interactions using scRNA-seq has also been proposed to study crosstalk of different cell types in the context of development, differentiation, and disease [30–33] (Table 1 and Box 1). These models combine prior knowledge of ligand–receptor complexes with a statistical framework to predict tissue-specific cell–cell communication networks via these molecular interactions. In particular, they can generate predictions of key cell–cell interactions and motifs responsible for maintaining tissue homeostasis and supporting tissue regeneration. In this regard, the single-cell based inference of cell–cell communication networks enabled the prediction of cell–cell interactions involved in maintaining tissue homeostasis [32–34]. We expect that the comparison of these reference networks with cell–cell interactomes of pathological or injured tissues will allow the identification of dysregulated interactions and can guide the development of novel intervention strategies for restoring homeostasis and supporting tissue regeneration. In addition, computational models of cell–cell communication have provided insights into general principles underlying tissue homeostasis. For example, the analysis of cell–cell communication networks indicated the necessity of endocytosis for maintaining cell type proportions [35,36]. Although current methodologies are able to infer cell–cell communication events between different cell populations, they cannot delineate their functional differences resulting from the spatial position of cells in the tissue. Therefore, these computational approaches can be combined with imaging-based technologies for spatial transcriptome reconstruction [37–39] to enable the characterization of the complete interactome in a spatially resolved manner. For instance, single-molecule fluorescence *in situ* hybridization (smFISH) can be employed to detect genes characteristic of different tissue locations. On the basis of expression of these genes, the spatial heterogeneity within scRNA-seq data can be resolved.

Models of tissue self-organization resulting from cell–cell interactions are valuable in the study of tissue homeostasis and regeneration. In particular, the combination of computational modeling, machine learning, and mathematical optimization has been employed to predict experimentally testable perturbations that generate desired multicellular spatial patterns in human iPSC colonies [40]. Hence, this data-driven approach enables the prediction and control of spatial self-organization of heterogeneous populations of stem cells. An extension of this method to model systems composed of different cell types can potentially be used to study and control processes such as tissue homeostasis and regeneration. Additionally, they can characterize fundamental self-organized patterns for tissue regeneration, such as the existence of distinct stem cell GRN configurations governing different aspects of the cell's response to environmental cues [4], which could be further implicated in tissue regeneration [4].

## Mechanistic Models versus Machine Learning Approaches

Machine learning approaches have been widely used for pattern recognition, classification, and prediction in the study of biological systems, including in stem cell research [1,4,40]. Some examples include the construction of 3D stem cell images from fluorescence microscopy images [41], the identification of regulatory relationships between genes from single-cell gene expression datasets [1], and the prediction of experimental conditions that yield specific multicellular patterns [40].

Although machine learning models have high predictive power, they require vast amounts of data, such as imaging and omics datasets, for establishing statistical relationships between input data and predicted output. In addition, they focus exclusively on predictions and not on understanding complex processes, preventing them from providing mechanistic insights into biological processes. Machine learning algorithms can be classified into two main categories: supervised and unsupervised learning. The supervised learning algorithm learns from labeled training data to predict the outcome for unforeseen data, whereas the unsupervised learning algorithm tries to make sense of unlabeled data by extracting features and patterns on its own.

An illustrative example is the application of a supervised machine learning classifier that compares gene expression profiles of the cell states during differentiation with those from a training library of hundreds of cell type–specific gene expression patterns curated from the literature in order to investigate the identity of these cell states [42]. By contrast, **mechanistic models** rely on the generation of mechanistic hypotheses inferred from experimental observations to produce novel predictions and describe the behavior of the system (Box 2). These models are most often built on simplified conceptual and mathematical formulations of the observed phenomenon (Boxes 1 and 3). For example, a model for binary stem cell differentiation relies on the assumption that the stem/progenitor cell phenotype is characterized by the balanced gene expression pattern of groups of opposing lineage specifiers, which reside in GRN feedback loops. Following this assumption, a Boolean network–based computational model has been developed to systematically predict cell fate determinants using as input gene expression profiles of stem/progenitor and daughter cells [43]. Moreover, a single-cell–based extension of this model has been employed to elucidate mechanisms of cell fate dysregulation associated with congenital diseases [13].

### Box 2. Overview of Mechanistic Modeling Frameworks for Stem Cell Research

Mechanistic models characterize a biological system on the basis of a known or hypothesized relationship between its entities, such as genes, where the nature of the association is specified in terms of the biological process that has given rise to the experimental observations. Thus, a hallmark of mechanistic models is the explicit incorporation of physical, biological, or chemical properties of biological systems. In recent decades, different conceptual and mathematical frameworks have been developed, including logical and continuous models, to represent a biological system at different levels of detail. The choice of the modeling framework determines the faithfulness of the represented biological reality; the ability to model system dynamics; and, importantly, the ability to make predictions. In general, more faithful representations of biological systems require more data but provide a higher predictive power.

Logical models are a qualitative framework for representing the relationship between genes and are based on a set of rules governing the dynamics of the system. In their simplest form, each entity in the model can attain one of two values – 'active' or 'inactive' – representing its state, which is determined by the states of its regulators. In the past, logical models have proved useful in addressing key questions in stem cell research. For instance, a GRN model of embryonic blood development was inferred from scRNA-seq data to predict TFs directing the differentiation of mesodermal progenitors into endothelial and primitive blood cells [55]. Similarly, a GRN model of hematopoietic differentiation was compiled from the literature and was employed to predict factors inducing directed transdifferentiation and reprogramming [56]. In addition to GRNs, logical models have been employed to infer cell–cell communication by integrating gene regulation, intracellular signaling pathways, and extracellular ligand–receptor interactions to predict molecules maintaining the phenotypic state of tissue cell populations [24].

The increasing availability of single-cell data has allowed investigators to model biological systems at a higher level of detail using continuous models. By contrast to logical models, this framework can provide a quantitative characterization of biological systems and does not require the classification of entities as being 'active' or 'inactive.' Consequently, the temporal dynamics of the system are represented at a continuous time scale, allowing the direct comparison of experimental data and the model's state. These continuous models have been employed to predict key TFs of stem cell differentiation, hematopoietic lineage specification, and transdifferentiation [5,57]. However, the applicability of continuous models is not limited to the prediction of TFs. For instance, a model of stem cell–niche interaction determined mechanical properties affecting cellular differentiation [58]. Using this model, predictions of the culture duration and stiffness were employed to optimize cellular differentiation protocols.

### Box 3. Modeling Types in Stem Cell Research

The main mechanistic modeling types that have been used to address specific problems in stem cell research can be classified into five categories, namely graphs, Boolean networks, regression-based models, ordinary differential equations, and Bayesian networks (see Table 1 in main text).

Graphs consist of vertices and edges that represent relationships between them, such as coexpression between genes or interactions between cells. Depending on the type of the relationship, the graph is either directed or undirected. Importantly, graphs do not impose modeling assumptions; they merely represent the data.

Boolean networks are graph-based models in which each vertex is associated with a Boolean value representing its state. These states depend on the state of other vertices in the network, which are formalized as Boolean functions. The main assumption of Boolean networks is that the activity of vertices can be classified as being active or inactive. In the context of GRNs, this assumption is justified because the rate of transcription follows as a Hill function, which approximates a Boolean classification [59]. Here, Boolean functions represent the collective regulation of a gene by several TFs. The identification of Boolean functions typically requires time-series or perturbation data. However, the increasing amount of scRNA-seq data corresponding to different cell subpopulations can be employed for the inference of Boolean functions.

By contrast to Boolean networks, continuous models associate vertices with continuous states, such as the continuous gene expression values. In this context, regression is a commonly used approach to assess the relationship between vertices. In particular, ridge regression has been developed for assessing the dependency between vertices having near-linear relationships. Besides regression, linear ordinary differential equations constitute another common modeling type that relates a function with a linear combination of its derivatives, such as gene expression with the rate of transcription. Although often justified in the context of GRNs, the assumption of linearity prevents considering nonadditive effects, including competitive TF regulation. Inference of continuous models typically requires time-series data representing biological variation over time.

Finally, dynamic Bayesian networks (DBNs) can model temporal stochastic processes. Interactions of a DBN are probabilistic and represent the conditional dependencies between its vertices. The principal assumption of DBNs is that an event can cause another event in the future, but not in the past, a justifiable assumption in the context of gene regulation. Similarly to continuous models, the inference of DBNs requires time-series measurements.

Despite the fact that mechanistic models can provide insights into and understanding of mechanisms, their simplified assumptions can sometimes overlook the complexity of biological processes. This impediment may prevent these models from capturing the underlying principles of biological processes and therefore limit their predictive power. Hence, mechanistic and machine learning models should be seen not as competitors but rather as complementary approaches. In particular, mechanistic models can make use of inferred patterns, learned from the data, to improve their efficiency. An example of this is the successful use of machine learning techniques to optimize the parameters of a computational model for describing and controlling pluripotent stem cell dynamics [40]. Moreover, several machine learning approaches for identifying TF binding sites have been developed whose predictions can be incorporated into current single-cell based GRN inference methods to prioritize relevant regulatory interactions [44–47]. In addition, machine learning can be employed to preprocess the input data of mechanistic models, including the imputation of different types of single-cell data and the integration of multiple types of data obtained from different experiments [48,49].

## Applications of Computational Modeling to Regenerative Medicine

**Cell transplantation** is one of the main strategies in regenerative medicine to replace damaged or aged cells with healthy functioning cells. Various clinical applications of autologous iPSC-derived cell transplantation have been initiated and are currently ongoing [50,51]; however, a number of challenges must be overcome for this approach to reach its full potential. For example, transplanted cells from a different tissue or *in vitro* experiments are not always successfully integrated into the target tissue [52]. A strategy to overcome this problem requires the improvement of *in vitro* manufacturing of donor tissue cells to gain the appropriate gene expression identity of host tissue cells. Indeed, existing *in vitro* experimental protocols often suffer from low **conversion**

**efficiency**, forcing experimentalists to spend a large amount of resources in order to collect enough target cells for subsequent functional experiments or clinical use. In addition, *in vitro* cell conversion often results in the creation of immature, nonfunctional variants of target cells, failing to generate cells with desired phenotypes and functionalities. In this regard, computational modeling can help address these limitations. Advancements in scRNA-seq technologies and the development of computational models of the kind mentioned before can enable researchers to characterize functionally distinct cell subtypes, exhibiting subtle gene expression differences, and to identify conversion factors for *in vitro* generation of such cell subtypes. Furthermore, network-based models that combine the hierarchical organization of cell identity TFs with GRNs could allow the prediction of optimal combinations of conversion factors that include cell type- and cell subtype-specific TFs. Overexpression of predicted conversion factors that include cell type-specific and cell subtype-specific identity TFs can facilitate the *in vitro* production of functionally mature cell subtypes [52,53].

Combining computational approaches with existing or novel experimental techniques would open new avenues in the design of protocols for treating congenital disorders and enhancing regeneration. For example, it has recently been shown that transplantation of **transgenic** keratinocyte cultures regenerated a fully functional epidermis in a patient with junctional epidermolysis bullosa, a genetic disorder characterized by a depletion of epidermal stem cells and impaired keratinocyte differentiation [54]. However, because the genetic cause of congenital diseases affecting cell differentiation is often unknown, computational methods can guide the identification of disease-related cell fate determinants in order to generate transgenic cells for transplantation. Additionally, computational predictions of dysregulated cell fate determinants caused by congenital disorders can be employed in the design of *in vivo* gene correction strategies using techniques such as CRISPR/Cas9.

Stem cell rejuvenation is another promising strategy to revert impaired stem cell function and optimize tissue repair processes in cases of degenerative diseases and age-related disorders. One of the reasons for stem cell dysfunction is the disruption of signaling pathways of endogenous stem cells due to the detrimental effect of the diseased or aged niche. The **quiescent stem cell** phenotype is maintained by the constant activation or inhibition of key signaling pathways that target functionally relevant TFs. Computational models of stem cell–niche interactions can help to identify these niche-specific cues and the corresponding affected stem cell signaling pathways to provide insights into the mechanism of stem cell functional dysregulation in aging and disease [26]. For example, models that integrate signaling and transcriptional regulatory networks using scRNA-seq data and, if available, phosphoproteomics data can be useful in identifying signaling molecules responsible for the stable maintenance of the disease- or aging-related stem cell phenotypes. Indeed, this type of integrative model is able to capture the downstream effect of niche-specific signaling cues on the transcriptional program that defines these stem cell phenotypes. Targeted perturbation of the predicted signaling molecules can be used to counteract the niche effect for stem cell rejuvenation in case of aging or degenerative disease.

## Concluding Remarks and Future Perspectives

A number of challenges in stem cell research can be addressed with the development of multiscale computational models. Indeed, with the increasing amount of available single-cell data, especially scRNA-seq data, we are now in the position of developing computational models at different levels of complexity, including intracellular and cell–cell communication network-based models. In this regard, generation and integration of different types of single-cell omics data would offer a more complete characterization of the biological system and improve the

### Outstanding Questions

How can we optimally combine machine learning approaches with mechanistic models to obtain more insight into complex biological systems?

How can we integrate different types of single-cell data to develop accurate mechanistic models for the design of novel stem cell therapies? Moreover, how can we develop strategies for integrating data profiled in different single cells of the same subpopulation?

How can we generate functionally mature cell subtypes *in vitro* with high efficiency that can be used for transplantation?

How can we develop computational models to prioritize cell–cell interaction events relevant for tissue homeostasis? Can we predict molecules that can be used for intervention strategies to restore tissue homeostasis?

accuracy of multiscale models. For example, by integrating scRNA-seq with scATAC-seq, we would be able to better characterize novel cellular subtypes/phenotypes, including their gene regulatory landscape, which allows the implementation of more accurate models for cellular conversion to produce specific cell subtypes. Similarly, integration of scRNA-seq with single-cell phosphoproteomics would facilitate the identification of signaling pathways mediating niche cues maintaining stem cell phenotypes. In addition, we expect that the combination of scRNA-seq with imaging data will enable the inference of the spatial cell–cell communication network.

Nevertheless, due to the inherent differences and limitations in the experimental designs, integrating multiple single-cell omics data remains an ongoing challenge for many researchers (see Outstanding Questions). For example, it is often the case that different omics layers are profiled in different single cells from the same cell population, which hinders data integration. Advances in single-cell experimental and computational techniques will help to overcome these challenges.

Although machine learning approaches have been employed successfully in the context of pattern recognition and classification, they do not provide insights into biological processes. The implementation of mechanistic models is necessary to better understand causal mechanisms and generate predictions based on simplified assumptions. The combination of machine learning and mechanistic models can enable the generation of more accurate predictions while providing mechanistic understanding of complex biological systems. For example, machine learning approaches can initially assist in data processing and analysis, including the imputation of different types of single-cell data, and the integration of multiple types of data obtained from different experiments. Consequently, mechanistic models can leverage these processed data to increase their predictive power.

Advances in regenerative medicine would greatly benefit from the guidance of computational modeling. In fact, several challenges in the field, such as the increase of cellular reprogramming efficiency and the control of tissue homeostasis and regeneration, require a deeper understanding of the underlying biological processes. Multiscale modeling can generate novel hypotheses to assist experimentalists in the design of stem cell therapies for tissue regeneration and treatment of congenital disorders. For instance, single-cell-based computational models for cellular conversion enable the prediction of cell subtype-specific conversion factors and therefore can improve current protocols for *in vitro* generation of target cells for transplantation, and they can provide valuable mechanistic insights for the treatment of congenital disorders. Models of stem cell–niche interactions can assist the design of novel strategies for stem cell rejuvenation. A larger-scale model based on cell–cell communication networks can help in the identification of intercellular interactions that impair regeneration in aged tissues. This would aid in establishing novel intervention strategies for increasing the tissue regenerative capacity.

We encourage a closer collaboration between experimental and computational researchers to accelerate stem cell research, including the translation of regenerative medicine applications to the clinic. In particular, experimental and computational stem cell biologists should coordinate research efforts from the onset to specify the adequate computational model to address a particular biological question, the necessary input data for this model, and the experimental design to validate the model predictions.

**Acknowledgments**

The authors would like to thank Amy Parrish for helping with the manuscript language editing.

## References


1. Chan, T.E. *et al.* (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267.e3
2. Aibar, S. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086
3. Papili Gao, N. *et al.* (2018) SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 34, 258–266
4. Stumpf, P.S. and MacArthur, B.D. (2019) Machine learning of stem cell identities from single-cell expression data via regulatory network archetypes. *Front. Genet.* 10, 2
5. Matsumoto, H. *et al.* (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* 33, 2314–2321
6. Woodhouse, S. *et al.* (2018) SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.* 12, 59
7. Sanchez-Castillo, M. *et al.* (2018) A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* 34, 964–970
8. Aubin-Frankowski, P.-C. and Vert, J.-P. (2020) Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics* Published online June 17, 2020. https://doi.org/10.1093/bioinformatics/btaa576
9. Moerman, T. *et al.* (2019) GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 35, 2159–2161
10. Deshpande, A. *et al.* (2019) Network inference with Granger causality ensembles on single-cell transcriptomic data. *bioRxiv* Published online January 30, 2019. https://doi.org/10.1101/534834
11. Qiu, X. *et al.* (2020) Inferring causal gene regulatory networks from coupled single-cell expression dynamics using Scribe. *Cell Syst.* 10, 265–274.e11
12. Kamimoto, K. *et al.* (2020) CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv* Published online April 21, 2020. https://doi.org/10.1101/2020.02.17.947416
13. de Soysa, T.Y. *et al.* (2019) Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature* 572, 120–124
14. Liu, X. *et al.* (2019) Single-cell RNA-seq of the developing cardiac outflow tract reveals convergent development of the vascular smooth muscle cells. *Cell Rep.* 28, 1346–1361.e4
15. Finnegan, A. *et al.* (2019) Single-cell transcriptomics reveals spatial and temporal turnover of keratinocyte differentiation regulators. *Front. Genet.* 10, 775
16. Tran, K.A. *et al.* (2019) Defining reprogramming checkpoints from single-cell analyses of induced pluripotency. *Cell Rep.* 27, 1726–1741.e5
17. Ruan, H. *et al.* (2019) Single-cell reconstruction of differentiation trajectory reveals a critical role of ETS1 in human cardiac lineage commitment. *BMC Biol.* 17, 89
18. D'Alessio, A.C. *et al.* (2015) A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep.* 5, 763–775
19. Okawa, S. *et al.* (2018) Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. *Nat. Commun.* 9, 2595
20. Saçma, M. *et al.* (2019) Haematopoietic stem cells in perisinusoidal niches are protected from ageing. *Nat. Cell Biol.* 21, 1309–1320
21. Yachie-Kinoshita, A. *et al.* (2018) Modeling signaling-dependent pluripotency with Boolean logic to predict cell fate transitions. *Mol. Syst. Biol.* 14, e7952
22. Mahadik, B. *et al.* (2019) A computational model of feedback-mediated hematopoietic stem cell differentiation in vitro. *PLoS One* 14, e0212502
23. Thalheim, T. *et al.* (2018) Linking stem cell function and growth pattern of intestinal organoids. *Dev. Biol.* 433, 254–261
24. Browaeys, R. *et al.* (2020) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* 17, 159–162
25. Ravichandran, S. *et al.* (2020) SigHotSpotter: scRNA-seq-based computational tool to control cell subpopulation phenotypes for cellular rejuvenation strategies. *Bioinformatics* 36, 1963–1965
26. Kalamakis, G. *et al.* (2019) Quiescence modulates stem cell maintenance and regenerative capacity in the aging brain. *Cell* 176, 1407–1419.e14
27. Qin, X. *et al.* (2020) Cell-type-specific signaling networks in heterocellular organoids. *Nat. Methods* 17, 335–342
28. Krishnaswamy, S. *et al.* (2014) Systems biology: conditional density-based analysis of T cell signaling in single-cell data. *Science* 346, 1250689
29. Dixit, A. *et al.* (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17
30. Vento-Tormo, R. *et al.* (2018) Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353
31. Efremova, M. *et al.* (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* 15, 1484–1506
32. Camp, J.G. *et al.* (2017) Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546, 533–538
33. Skelly, D.A. *et al.* (2018) Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell Rep.* 22, 600–610
34. Raredon, M.S.B. *et al.* (2019) Single-cell connectomic analysis of adult mammalian lungs. *Sci. Adv.* 5, eaaw3851
35. Zhou, X. *et al.* (2018) Circuit design features of a stable two-cell system. *Cell* 172, 744–757.e17
36. Adler, M. *et al.* (2018) Endocytosis as a stabilizing mechanism for tissue homeostasis. *Proc. Natl. Acad. Sci. U. S. A.* 115, E1926–E1935
37. Karaiskos, N. *et al.* (2017) The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358, 194–199
38. Halpern, K.B. *et al.* (2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* 542, 352–356
39. Rodriques, S.G. *et al.* (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467
40. Libby, A.R.G. *et al.* (2019) Automated design of pluripotent stem cell self-organization. *Cell Syst.* 9, 483–495.e10
41. Chen, J. *et al.* (2018) The Allen Cell Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. *bioRxiv* Published online December 18, 2018. https://doi.org/10.1101/491035
42. Stumpf, P.S. *et al.* (2017) Stem cell differentiation as a non-Markov stochastic process. *Cell Syst.* 5, 268–282.e7
43. Okawa, S. *et al.* (2016) A generalized gene-regulatory network model of stem cell differentiation for predicting lineage specifiers. *Stem Cell Rep.* 7, 307–315
44. de Boer, C.G. *et al.* (2020) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* 38, 56–65
45. Lan, G. *et al.* (2019) Cross-cell-type prediction of TF-binding site by integrating convolutional neural network and adversarial network. *Int. J. Mol. Sci.* 20, 3425
46. Li, H. *et al.* (2019) Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res.* 29, 281–292
47. Zhou, J. *et al.* (2019) MTTFsite: cross-cell type TF binding site prediction by using multi-task learning. *Bioinformatics* 35, 5067–5077
48. Stuart, T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21
49. Lopez, R. *et al.* (2019) A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *Submitted to the ICML Workshop on Computational Biology.* Published online May 6, 2019. http://arxiv.org/abs/1905.02269
50. Mandai, M. *et al.* (2017) Autologous induced stem-cell-derived retinal cells for macular degeneration. *N. Engl. J. Med.* 376, 1038–1046
51. Kikuchi, T. *et al.* (2017) Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. *Nature* 548, 592–596

52. Xu, J. *et al.* (2015) Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell* 16, 119–134
53. Berneman-Zeitouni, D. *et al.* (2014) The temporal and hierarchical control of transcription factors-induced liver to pancreas transdifferentiation. *PLoS One* 9, e87812
54. Hirsch, T. *et al.* (2017) Regeneration of the entire human epidermis using transgenic stem cells. *Nature* 551, 327–332
55. Moignard, V. *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* 33, 269–276
56. Collombet, S. *et al.* (2017) Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc. Natl. Acad. Sci. U. S. A.* 114, 5792–5799
57. Ocone, A. *et al.* (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* 31, i89–i96
58. Peng, T. *et al.* (2017) A mathematical model of mechanotransduction reveals how mechanical memory regulates mesenchymal stem cell fate decisions. *BMC Syst. Biol.* 11, 55
59. Bhaskaran, S. *et al.* (2015) Hill equation in modeling transcriptional regulation. In *Systems and Synthetic Biology* (Singh, V. and Dhar, P., eds), pp. 77–92, Springer Netherlands