Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions

Grace Hui Ting Yeo1,2,5, Sachit D. Saksena1,2,5 & David K. Gifford1,3,4✉

Existing computational methods that use single-cell RNA-sequencing (scRNA-seq) for cell fate prediction do not model how cells evolve stochastically and in physical time, nor can they predict how differentiation trajectories are altered by proposed interventions. We introduce PRESCIENT (Potential energy underlying Single Cell gradients), a generative modeling framework that learns an underlying differentiation landscape from time-series scRNA-seq data. We validate PRESCIENT on an experimental lineage tracing dataset, where we show that PRESCIENT is able to predict the fate biases of progenitor cells in hematopoiesis when accounting for cell proliferation, improving upon the best-performing existing method. We demonstrate how PRESCIENT can simulate trajectories for perturbed cells, recovering the expected effects of known modulators of cell fate in hematopoiesis and pancreatic β cell differentiation. PRESCIENT is able to accommodate complex perturbations of multiple genes, at different time points and from different starting cell populations, and is available at https://github.com/gifford-lab/prescient.
modeling developmental landscapes is essential to improving our understanding of how cells are driven to transient and terminal states in vivo and to enable precise manipulation of cell fates in vitro. Single-cell RNA-seq analysis (scRNA-seq) has enabled the study of developmental landscapes by the observation of gene expression in single cells sampled at multiple stages of differentiation. However, these studies provide snapshots of a given differentiation process and do not directly observe lineage relationships between cells at different time points in development. Recently, experimental lineage tracing methods that couple various barcoding strategies with scRNA-seq have been described that identify lineage relationships. These methods provide ground truth for computational models of differentiation.

Existing computational approaches for modeling differentiation typically summarize observations of cell states and couplings emergent of the underlying process and have limited to no capacity for modeling differentiation as a continuous process (Fig. 1a). The predominant approach is pseudo-temporal inference, which orders cells along an arbitrary one-dimensional measurement representing differentiation time, and hence cannot model differentiation dynamics with respect to real, physical time. Other methods have also emerged for the specific task of cell fate prediction. For example, Waddington-OT predicts long-range cell–cell probabilistic couplings by reframing the task of inferring cell relationships between population snapshots as an unbalanced optimal transport problem. Another method, Fate-ID iteratively builds ensemble cell-type classifiers from labeled terminal cell states. However, these methods only summarize observations of cell states and couplings emergent of the underlying differentiation process. Recently, a small number of methods have described approaches to modeling differentiation as a process, but they have been limited either in how the model is solved, or in modeling capacity. For example, Population Balance Analysis (PBA) solves a reaction-diffusion partial differential equation describing differentiation but is forced to use a non-parametric solution due to computational constraints. Similarly, pseudodynamics models a diffusion process but only in a one-dimensional cell state.

We introduce PRESCIENT (Potential enerGy undErlying Single Cell gradIENTs), a generative modeling framework fit using longitudinal scRNA-seq datasets to model complex potential landscapes. PRESCIENT extends previous work by Hashimoto et al. that showed that a global potential function of a time-series is recoverable via a diffusion-based model fit to well-mixed, cross-sectional observations. PRESCIENT builds upon this by enabling the model to operate on large numbers of cells over many timepoints with high-dimensional features, and by incorporating cellular growth estimates. We validate PRESCIENT on a newly published lineage tracing dataset by evaluating PRESCIENT’s ability to generate held-out timepoints and to predict cell fate bias, i.e. the probability a cell enters a particular fate given its initial state. We show that when accounting for cell proliferation, PRESCIENT outperforms existing methods on predicting cell fate bias. Unlike existing methods, PRESCIENT learns a stochastic, parametric, queryable form of the differentiation landscape via a generative neural network, which enables simulations of high-dimensional trajectories with arbitrary initializations in physical time. This enables simulation of trajectories for cells unobserved during training, including cells with computationally perturbed gene expression profiles, which none of the existing summarization methods or modeling methods are able to do (Fig. 1a). This is also in contrast to other generative methods like scGen, which was proposed for predicting shifts in gene expression space in response to perturbations via autoencoder latent space arithmetic. While scGen is a promising approach for generating cell profiles under different perturbations for initializations of PRESCIENT models (Discussion), it is not time resolved, does not generate distributions of cells, and does not explicitly model cellular differentiation. We demonstrate how PRESCIENT can be used to model perturbations of multiple genes, at different time points, and from different starting states. We are able to recover expected changes in final cell fate distributions when interrogating our models using perturbations of known regulators of cell fate in hematopoiesis and pancreatic β cell differentiation. PRESCIENT enables large unbiased in silico perturbation experiments to aid the design of in vitro genetic perturbational screens.

Results

Learning a generative model of cellular differentiation from high-dimensional scRNA-seq data. PRESCIENT models cellular differentiation as a diffusion process over a gene expression landscape parameterized by a potential function that we wish to identify given only time-series population snapshots of single-cell RNA expression. In this diffusion process, evolution of a cell’s state at a given time is governed by a drift term, corresponding to the force acting on that cell given its current state, and a noise term, corresponding to stochasticity. In particular, the drift term is defined to be the negative gradient of the potential function, such that the potential induces a force that naturally drives cells toward regions of low potential (Fig. 1b; Methods). This stochastic process can then be simulated via first-order time discretization to sample trajectories for a given cell. The potential function is fit by minimizing a regularized Wasserstein loss between empirical and predicted populations at the observed time points (Fig. 1c; Methods). Previous work has shown the recoverability of diffusion dynamics from cross-sectional observations via this objective function.

To enable the modeling framework to operate on large scRNA-seq datasets, we fit models on PCA projections of the scaled gene expression data, which has been successfully used in down-stream scRNA-seq analysis methods such as clustering and cross-dataset integration. The potential function is given by a neural network, which operates as a black-box function approximator, hence enabling complex parameterizations of the landscape (Fig. 1c).

Finally, we take into account cell proliferation by weighting each cell in the source population according to its expected number of descendants in the objective. To assess the importance of incorporating cell proliferation, we study models assuming a priori knowledge of cell proliferation, which can be directly estimated from the data by computing the number of descendants for each starting cell given lineage tracing data, as well as models where cell proliferation is estimated from gene expression.

Fate outcomes generated by PRESCIENT align with experimental lineage tracing when taking into account cell proliferation. We validate our model on a recently published lineage tracing dataset by Weinreb et al. which used DNA barcodes to track clonal trajectories during mouse hematopoiesis. We evaluate our model on two tasks: recovery of a held-out time point, and cell fate prediction (Fig. 1d, e).

We first evaluated whether models were able to recover the marginal cell population at a held-out time point, day 4, when trained only on days 2 and 6, using cells that have lineage tracing data available. We evaluated the Wasserstein distance between the simulated and the empirically observed cell populations for days 4 (testing distance) and 6 (training distance) on the epoch with the lowest training distance. Simulated populations generated by our model outperform baselines, including the distance of the
simulated population at day 4 to the actual populations at day 2 and 6, as well as a linearly interpolated population as predicted by WOT8 (Fig. 2a, Supplementary Note).

We next evaluated on cell fate prediction, which we define to be the task of predicting the clonal fate bias of a given barcoded clone as described by Weinreb et al. This is defined as the number of neutrophils divided by the total number of neutrophils and monocytes for that clone (Methods). To predict, we simulate 2000 trajectories initialized with only the starting cell of each clone until the final time point. For each of these trajectories, we classify the cell at the final time point as neutrophil, monocyte or other using an approximate nearest neighbor (ANN) classifier that had

Fig. 1 A generative model of cellular differentiation. a Existing single-cell models of development can be described as operating in pseudo-time or real time (x-axis), and by the extent to which they model the underlying differentiation process (y-axis). PRESCIENT is highlighted in red. b Observations of population-level time-series data are used in a generative framework that models the underlying dynamic process in physical time. Evolution of a cell's state is governed by a drift term and a noise term. The drift, depicted by solid arrows, is defined as the negative gradient of the potential function, depicted by the color gradient in the background. Dashed lines correspond to noise. The model is fit using observations of population-level time-series data, depicted as solid circles. Simulations of cell states are depicted as dashed circles. c Cartoon depicting model fitting process. The neural network parameterizing the underlying drift function $\mu$ takes as input the PCA projections of gene expression data at observed time points (again depicted as solid lines). The stochastic process is then simulated via first-order time discretization to produce a population at the next time step, and so on. This proceeds until the next observed time point, at which the loss between the simulated and predicted population is minimized. The model was validated using two tasks. d Held-out recovery, where the model was asked to predict the marginal distribution of a held-out time point, and e–f, fate prediction, where the model was asked to predict the fate distribution outcome of a given progenitor cell. Fate prediction can be applied to cells observed in the dataset (e) or cell states in which some perturbation has been applied in silico (f). As shown, the perturbation results in a significant shift of fate distribution outcomes.

of neutrophils divided by the total number of neutrophils and monocytes for that clone (Methods). To predict, we simulate 2000 trajectories initialized with only the starting cell of each clone until the final time point. For each of these trajectories, we classify the cell at the final time point as neutrophil, monocyte or other using an approximate nearest neighbor (ANN) classifier that had
been trained using cell-type labels provided by Weinreb et al. (Methods). We then evaluate the clonal fate bias on the test set, by ensembling predictions over the last 5 evaluated epochs (Figs. 2c, d, S2a). We measure performance as the Pearson correlation with respect to the actual clonal fate bias given by the lineage tracing data, as well as AUROC of classifying a given cell as having a clonal fate bias of >0.5, since the metric is strongly bimodal (Fig. S2b).

We first fit a PRESCIENT model on the subset of data for which lineage tracing data is available (Fig. 2e). Figure 2d depicts the learned potential and drift functions for models trained with and without cell proliferation. Qualitatively, we observe that incorporating cell proliferation changes the potential landscape near the earliest time point. When the model does not take into account cell proliferation, its performance is similar to existing fate prediction methods like Waddington-OT with provided empirically derived cell proliferation ($r = 0.150$, AUROC = 0.599) and FateID ($r = 0.225$, AUROC = 0.602), achieving $r = 0.196 \pm 0.020$, AUROC = 0.601 ± 0.006 over 5 seeds. Incorporating empirically derived cell proliferation rates into the PRESCIENT modeling framework greatly improves performance with mean $r = 0.347 \pm 0.029$, AUROC = 0.692 ± 0.012, more closely approaching the upper-bound performance estimated using held-out clonal data of $r = 0.487$, AUROC = 0.771 (Figs. 2f, g; S2).

Accounting for cell proliferation improves the performance of the fate prediction task. However, these models were fit with empirically derived cell proliferation rates calculated from the lineage tracing data, which is usually unavailable (Figs. 2e, S2c). We next looked at whether cell proliferation rates derived from gene expression could achieve similar performance. To this end,
we modified an approach described by Schiebinger et al.8 to use KEGG annotations of cell cycle and apoptosis genes which were also highly variable in the dataset to estimate the number of descendants a cell is expected to have. These estimates correlated weakly but significantly (r = 0.207, p = 1e−45) with the empirical rates calculated from the lineage tracing data (Fig. 2c). We compared models fit to the same set of cells using either our gene-expression-derived or empirical proliferation estimates. We found that models incorporating gene-expression-derived estimates achieved r = 0.399 ± 0.025 and AUROC = 0.725 ± 0.008, a slight improvement over empirical proliferation rate based models (Fig. 2f, g).

Generative models can simulate trajectories for cells not observed during training. We next hypothesized that PRESCIENT should be able to predict the fate of cells not observed during training. We expect that the model has learned a good approximation of the underlying potential function from the training data and hence should be able to generalize to unseen data points. To test our hypothesis, we used our proliferation estimates to fit models to all cells with and without lineage tracing data.

We found that model performance was similar when the cells in the test set were included in the training dataset (r = 0.391 ± 0.035, AUROC = 0.723 ± 0.013), and when they were not (r = 0.407 ± 0.019, AUROC = 0.727 ± 0.014) (Fig. 2f, g). Furthermore, although performance as measured by correlation and AUROC is similar to when models were fit only on cells with lineage tracing data, the fraction of test set cells for which the model predicted at least one cell entering a neutrophil or monocyte fate increased slightly from 0.74 ± 0.01 to 0.80 ± 0.03 (all cells) and 0.83 ± 0.02 (only cells without lineage tracing data) (Fig. 2h), suggesting that the models did benefit from observing more data.

We tested if different magnitudes of the perturbation had different effects (Methods). We found that increasing the magnitude of the perturbation resulted in larger changes in the relative cell fractions (Fig. 3d, e). Furthermore, while individual perturbations resulted in a mixture of significant and non-significant changes in the final neutrophil populations, ensemble perturbations consistently resulted in significant changes (p < 0.05; Fig. 3f, g). We also tested multiple sets of randomly selected non-TFs to ensure the changes in final cell fate were not simply a result of perturbations causing random model changes. These randomly selected non-TFs do not result in an observed shift in neutrophil and monocyte fates in the final time point (Fig. 3h), suggesting that our model is robust to random effects.

PRESCIENT predicts expected changes in cell fate when perturbing transcription factors involved in the regulation of hematopoiesis. We hypothesized that a PRESCIENT model trained on the Weinreb et al. dataset should be able to recapitulate the effects on cell fate when perturbing transcription factors (TFs) known to be involved in regulation of neutrophil or monocyte differentiation. (Figs. 3a, b, S3a, b). In particular, we focus on a set of TFs previously identified to be potentially antagonistically correlated with either monocyte and neutrophil fate in progenitor cells by MetaCell analyses and CRISPR-seq experiments of a haematopoietic stem cell dataset17, many of which are supported by existing literature18–21.

We first introduced perturbations to Lmo4, Cebpe, Mxl1, and Dach1. TFs previously identified to be involved in granulopoiesis, the production of mature neutrophils. As expected, we observed that down-regulation of these TFs led to a relative decrease in the fraction of neutrophils while up-regulation of these TFs led to a relative increase in the fraction of neutrophils with respect to the unperturbed population. We next perturbed TFs involved in monocyte development, including Ifit1, Ifit5, Klf4, and Nr4a1, and observed similar results (Fig. 3c).

We next tested if different magnitudes of the perturbation had different effects (Methods). We found that increasing the magnitude of the perturbation resulted in larger changes in the relative cell fractions (Fig. 3d, e). Furthermore, while individual perturbations resulted in a mixture of significant and non-significant changes in the final neutrophil populations, ensemble perturbations consistently resulted in significant changes (p < 0.05; Fig. 3f, g). We also tested multiple sets of randomly selected non-TFs to ensure the changes in final cell fate were not simply a result of perturbations causing random model changes. These randomly selected non-TFs do not result in an observed shift in neutrophil and monocyte fates in the final time point (Fig. 3h), suggesting that our model is robust to random effects.

PRESCIENT enables in silico simulations of perturbed cell profiles. The ability to simulate trajectories for unobserved cells allows the model to make predictions of the fate distribution outcome of cells with perturbed gene expression profiles. We demonstrate this ability on two model systems drawn from published studies with time-series scRNA-seq measurements of differentiation. For these experiments, we first simulate trajectories for an unperturbed initialization (Fig. S2d, e). We then introduce perturbations to the same initial cell population by introducing different levels of overexpression or knock-down of genes that have been reported in the literature to modulate cell fate outcome (Methods). These perturbations can involve multiple genes, or be introduced at different time points, or in different starting populations. The resulting gene expression profile is then transformed into PCA space to initialize simulations of perturbed cell trajectories.

PRESCIENT predicts expected outcomes of transcription factor perturbations in endocrine induction introduced at different timepoints and developmental stages. We next applied PRESCIENT to a 7 time-point scRNA-seq time course of another...
well-characterized differentiation system, the production of pancreatic islet cell types in vitro. This dataset did not include lineage tracing measurements.

We first hypothesized that PRESCIENT should be able to recapitulate the effects on cell fate when perturbing TFs known to be involved in the regulation of endocrine induction and specification in the starting population. Previous work has shown that NEUROG3 and NKX6 activation is associated with the endocrine lineage, while PTF1A and HES1 is associated with the exocrine lineage. When introducing ensembled in silico perturbations of NEUROG3 and NKX6.

Figure 3: In silico perturbations of hematopoiesis results in expected shifts in fate distribution. a The distribution of cells at the final time-point generated by the model initialized with unperturbed cells (left) and cells with perturbations of Lmo4, Cebpe, Mxd1, and Dach1 upregulated (z = 10) during neutrophil differentiation (right). b Proportions of generated cell types from day 2 to 6 initialized with unperturbed cells (left) and cells with perturbations of transcription factors upregulated during neutrophil differentiation (right). c Fraction of neutrophil and monocyte cells at final time point with single-gene perturbations. Individual genetic perturbations made to transcription factors involved in monocyte development are indicated in green, while transcription factors involved in neutrophil development are indicated in orange. Control genes (in gray) indicate experiments when perturbing genes from a random set of non-TFs as in (h) d e. f–g Ensemble perturbations of transcription factors involved in neutrophil (orange) and monocyte (green) differentiation have a stronger effect. h Ensemble random perturbations of non-transcription factors without proliferative signatures (gray). In c–h, boxplots are of randomly initialized unperturbed vs. perturbed simulations (n = 10) with 200 cells for each initialization, and red asterisks indicate Welch’s independent two-sided t-test at p < 0.05. Boxplots indicate median (middle line), first and third quartiles (box), and the upper whisker extends from the edges to the largest value no further than 1.5 × IQR (interquartile range) from the quartiles and the lower whisker extends from the edge to the smallest value at most 1.5 × IQR of the edge, while data beyond the end of the whiskers are outlying points that are plotted individually as diamonds.

Figure 4: Time courses of transcription factors involved in hematopoiesis. a Neutrophil-associated TFs and b Monocyte-associated TFs. c–d Individual genetic perturbations made to transcription factors involved in neutrophil and monocyte differentiation have an increased effect at higher dosages. e–f Ensemble perturbations of transcription factors involved in neutrophil (orange) and monocyte (green) differentiation have a stronger effect. g Ensemble random perturbations of non-transcription factors without proliferative signatures (gray). In c–h, boxplots are of randomly initialized unperturbed vs. perturbed simulations (n = 10) with 200 cells for each initialization, and red asterisks indicate Welch’s independent two-sided t-test at p < 0.05. Boxplots indicate median (middle line), first and third quartiles (box), and the upper whisker extends from the edges to the largest value no further than 1.5 × IQR (interquartile range) from the quartiles and the lower whisker extends from the edge to the smallest value at most 1.5 × IQR of the edge, while data beyond the end of the whiskers are outlying points that are plotted individually as diamonds.

Figure 5: In silico perturbations of hematopoiesis results in expected shifts in fate distribution. a The distribution of cells at the final time-point generated by the model initialized with unperturbed cells (left) and cells with perturbations of Lmo4, Cebpe, Mxd1, and Dach1 upregulated (z = 10) during neutrophil differentiation (right). b Proportions of generated cell types from day 2 to 6 initialized with unperturbed cells (left) and cells with perturbations of transcription factors upregulated during neutrophil differentiation (right). c Fraction of neutrophil and monocyte cells at final time point with single-gene perturbations. Individual genetic perturbations made to transcription factors involved in monocyte development are indicated in green, while transcription factors involved in neutrophil development are indicated in orange. Control genes (in gray) indicate experiments when perturbing genes from a random set of non-TFs as in (h) d e. Individual genetic perturbations made to transcription factors involved in monocyte development are indicated in green, while transcription factors involved in neutrophil development are indicated in orange. Control genes (in gray) indicate experiments when perturbing genes from a random set of non-TFs as in (h) d e. f–g Ensemble perturbations of transcription factors involved in neutrophil (orange) and monocyte (green) differentiation have a stronger effect. h Ensemble random perturbations of non-transcription factors without proliferative signatures (gray). In c–h, boxplots are of randomly initialized unperturbed vs. perturbed simulations (n = 10) with 200 cells for each initialization, and red asterisks indicate Welch’s independent two-sided t-test at p < 0.05. Boxplots indicate median (middle line), first and third quartiles (box), and the upper whisker extends from the edges to the largest value no further than 1.5 × IQR (interquartile range) from the quartiles and the lower whisker extends from the edge to the smallest value at most 1.5 × IQR of the edge, while data beyond the end of the whiskers are outlying points that are plotted individually as diamonds.

Figure 6: Time courses of transcription factors involved in hematopoiesis. a Neutrophil-associated TFs and b Monocyte-associated TFs. c–d Individual genetic perturbations made to transcription factors involved in neutrophil and monocyte differentiation have an increased effect at higher dosages. e–f Ensemble perturbations of transcription factors involved in neutrophil (orange) and monocyte (green) differentiation have a stronger effect. g Ensemble random perturbations of non-transcription factors without proliferative signatures (gray). In c–h, boxplots are of randomly initialized unperturbed vs. perturbed simulations (n = 10) with 200 cells for each initialization, and red asterisks indicate Welch’s independent two-sided t-test at p < 0.05. Boxplots indicate median (middle line), first and third quartiles (box), and the upper whisker extends from the edges to the largest value no further than 1.5 × IQR (interquartile range) from the quartiles and the lower whisker extends from the edge to the smallest value at most 1.5 × IQR of the edge, while data beyond the end of the whiskers are outlying points that are plotted individually as diamonds.

Perturbed simulation (Lmo4, Cebpe, Mxd1, Dach1; day 6) Unperturbed simulation (day 6)

Time steps from day 2 to day 6 (dt = 0.1)

Neutrophil-associated TFs Monocyte-associated TFs

Magnitude of perturbation

Fraction of neutrophils

Fraction of monocytes

Random set of non-TFs

Overexpression (z=5)

Underexpression (z=5)

Fraction of neutrophils

Fraction of monocytes

Non-TF ensemble

Non-TF ensemble

Magnitude of perturbation
screens (Fig. 5a). Mean expression of TFs on day 0 was not correlated with predicted log2(FC) of cell fraction, showing that we could predict the effects of TFs even if their relative expression was low on day 0 (Fig. S4d).

We next asked if PRESCIENT could recapitulate the known timing of TFs during endocrine induction by introducing perturbations to the same endocrine/exocrine axis TFs described above at multiple timepoints (Methods). Endocrine/exocrine induction is known to occur early in the time-course, and the results corroborate this as early perturbations to NKX6.1 and NEUROG3 (z = 10) result in an increase in endocrine cells, but this effect diminishes with perturbations induced at later time points (Fig. 4d).

PRESCIENT also enables perturbations of cells sampled from different starting cell populations along a given differentiation trajectory. To demonstrate this, we introduced perturbations of selected TFs from the above screen to cells sampled from different stages of the endocrine induction pathway as labeled by Veres et al. (Fig. 5b, NKX2.2, NKX6.1, PAX4, and ARX are found early in endocrine specification and are often the first signal of the production of specific terminal endocrine cell fates23. We found that perturbations of these TFs in pancreatic progenitors result in a significant increase in final β-cell proportion (p < 0.05) and this effect is minimal in cells further along the endocrine induction pathway (Figs. 5b, S4e, f). In contrast, PDX1 has been shown to continue to promote β-cell neogenesis late into the endocrine induction pathway40. We show that perturbations to PDX1 increase the fraction of β-cells in both progenitor cells and cells with late NEUROG3 expression (neurog3late), recapitulating the multiple contexts in which perturbations to PDX1 can modulate endocrine fate. Similarly, MAFB and MAFA are known to continue to modulate fate late into endocrine induction37,41, and we show that perturbations to both MAFA and MAFB result in increases in β-cell proportion when introduced to both early pancreatic progenitors and cells in late endocrine induction. While ISL1 has been previously reported to stimulate islet cell production, the timing of ISL1 activation is less well characterized. Our model suggests that ISL1 has a role in β-cell specification in both early pancreatic progenitor phase and late in the endocrine induction pathway.

Finally, we observed that modulation of α-cell fate largely occurs via perturbations induced in the pancreatic progenitor
Discussion

PRESCIENT is a generative modeling framework for learning potential landscapes from population-level time series scRNA-seq data. PRESCIENT marks a departure from the predominant methods for analyzing scRNA-seq studies of cellular differentiation. Computational methods for lineage inference have been dominated by pseudo-time approaches that do not attempt to model the stochastic or dynamic nature of cell fate determination. More recent fate prediction methods either summarize observations of the emergent process or suffer from modeling limitations (see Introduction). However, we have demonstrated an important predictive advantage to fully generative models that seek to describe the underlying differentiation landscape. After the model has learned the landscape, it can generate trajectories for unseen data points.

We can hence interrogate PRESCIENT to propose hypotheses for possible perturbations. We show that inducing perturbations in well-studied regulatory genes of hematopoiesis and β-cell differentiation result in expected changes in fate outcome. We also show that we can model the differentiation outcomes of complex perturbations consisting of multiple genes, or at different time points or in different starting populations. This enables large, combinatorial in silico experiments that can help limit the number of in vitro experiments needed to achieve a desired cell fate. PRESCIENT can be used to identify targets for genetic- and small molecule-based screens and aid the design and fine-tuning of new directed differentiation and reprogramming protocols. We show an example of this type of unbiased, large-scale screen for in vitro β-cell differentiation in which we perturbed 200+ TFs and identified target genes that could cause significant shifts in β-cell, α-cell, and EC-cell fates. While this work was limited to TFs to show the utility of the method, non-TF targets, such as signaling pathway effectors, can also be tested using PRESCIENT. However, the model is subject to constraints and assumptions, for example requiring that the final time point of the dataset be at steady-state. The model also improves when incorporating growth rates, and would likely benefit from better estimates of proliferation. There also remain challenges to confidently suggest gene sets for experimental perturbation. One problem is that information is lost about individual genes when transforming data into PCA space, and lowly-expressed genes important to cell fate decisions may be dropped altogether in the scRNA-seq data. This can be addressed by methods that have proposed ways to directly generate gene expression counts from latent cell states or by approaches that model perturbation profiles in the original gene expression space. These profiles can then be used to initialize PRESCIENT simulations or even extend PRESCIENT to include end-to-end encoding. Another problem is that the association of certain genes with specific cell fates does not necessarily imply causality.

Future extensions of PRESCIENT would accommodate other data or modeling approaches. For example, PRESCIENT’s objective can be modified to maximize the likelihood of observing individual trajectories given lineage tracing data. Further, first-principle approaches to modeling dynamics such as RNA velocity are complementary approaches that are non-generative but can be used to constrain PRESCIENT models, e.g. RNA velocity has been proposed for constraining flows across timepoints to local velocities within timepoints. We expect that integrating additional sources of information or increasing sampling density of timepoints should improve the quality of the underlying landscape inferred.

Methods

Identifying the latent dynamics of cellular differentiation. Following Hashimoto et al. we model cellular differentiation as a diffusion process $X(t)$ given by the stochastic differential equation

$$dX(t) = f(X(t))dt + \sqrt{2\sigma^2}dW(t)$$

where $X(t)$ represents the k-dimensional state of a cell at time $t$, $f(X(t))$ is a drift term representing the force acting on a cell given its state, and $W(t)$ corresponds to unit Brownian motion. In particular, the drift function is defined to be the negative of the gradient of a potential function $\phi(x)$, such that intuitively, the potential function $\Phi(x)$ can be thought of as inducing a gradient field driving cells from regions of high potential to low potential. Within the conceptual framework of Waddington’s epigenetic landscape, this potential function corresponds to the height of the landscape. This process can be simulated via first-order time discretisation

$$X(t + \Delta t) = X(t) + f(X(t))\Delta t + \sqrt{2\sigma^2}\Delta W(t)$$

where $\Delta t$ are i.i.d. standard Gaussians. This converges to the diffusion process as $\Delta t \to 0$.

We define the marginal distribution at time $t$ to be $p(x(t)) = P(X(t) = x)$. The inference task identifies the potential function $\Phi(x)$, and hence the underlying drift function $f(x)$, given only samples from the marginal distribution $(x(t), \sim p(x(t))) \in \{1...m\}$, where $m$ is the number of sampled cells at time $t$ and $n$ is the number of time points where data was observed. In practice, this data corresponds to gene expression profiles of cells sampled over the course of a time-series experiment. Inference proceeds by finding the potential function $\Phi$ in a family of functions $K$ that minimizes the objective

$$\min_{\Phi \in K} \left( \sum_{i=1}^{m} \Phi(x_i) \right) + \lambda \left( \sum_{i=1}^{n} \Phi(x_i) \right)$$

Incorporating cell proliferation. Using notation from Feydy et al. computing the Wasserstein distance involves solving the optimization problem

$$\min_{\phi \in K} \sum_{i=1}^{m} \pi_i \psi_i(\phi)$$

where $\pi_i$, $\psi_i$, and $\phi$ are parameters controlling the strength of the entropic regularizer. To motivate this loss metric, it is helpful to compare to a case in which we can actually observe ground-truth trajectories of a diffusion process $X(t)$, in which case prediction error can be directly measured as the Euclidean distance between observed plots along the trajectory and samples from the predicted distribution of $X(t)$ under the model. Wasserstein distance is the direct analog of Euclidean distance when instead considering cross-sectional observations of indistinguishable particles along trajectories of a given diffusion process. In our case, each time point in longitudinal scRNA-seq is a cross-section of cell populations along multiple differentiation trajectories.
The drift function is visualized as unit arrows, where the point of origin is given by the same grid and the vector is given by the drift evaluated at the point of origin.

Preprocessing of existing scRNA-seq datasets. Preprocessed data for the Weinreb et al. experiments was downloaded from https://github.com/AllonKlein/Lab/paper-data/blob/master/Lineage_tracing_on_transcriptional_landscapes/links_state_to_fate_during_differentiation/README.md (commit: d8f0969). The set of highly variable genes was determined as by Weinreb et al. by first filtering for highly variable genes, and then excluding genes correlated with cell cycle (SPRING, correlation coefficient > 0.7). Normalized gene expression for variable genes was selected and projected to 50 dimensions via PCA, which was then used as input to the modeling framework. For experiments evaluating the model on the held-out time point, preprocessing was fit to only the training set consisting of days 2 and 6, and then used to transform all data including day 4. For all other experiments, preprocessing was fit to all data across time points. All visualizations using umap were fit with 30 neighbors.

Data for the Veres et al. experiments was downloaded from GEO (GSE114412)22. Raw counts were first pre-processed using the standard Seurat pipeline (v3.1.5)2 to obtain normalized counts. For feature selection, genes were further filtered for those observed in at least 10 cells. Then, the ‘FindVariableFeatures’ function was used to identify the top 2500 most variable genes. Scaled gene expression was then computed as was for the Weinreb et al. dataset. For projection into PCA space (30 PCs), the variable gene set was filtered again to remove genes correlated with TOPO2A (r > 0.15), as described by Veres et al. This was used as input to the modeling framework, and for visualization via UMAP. For estimation of proliferation rates, the full variable gene set was used (Fig. S3a, b, c, d).

Experiments on recovery of a held-out time point. For comparison to Waddington-OT (WOT)8, which uses held-out recovery (interpolation) as a benchmark, we fit models via pre-training on day 6 and then trained on days 2 and 6 for evaluation on day 4. Models were fit using only the subset of data for which lineage tracing data was available to enable comparison of models incorporating empirically derived cell proliferation rates. All models were trained for 2500 epochs.

To evaluate the models at a given time point, 10,000 cells were sampled at day 2 with replacement according to the expected cell proliferation rate. Then, the model was used to sample a single trajectory for each of the sampled cells until the time point under evaluation. The Wasserstein distance was then computed between the simulated cell population and the empirically observed cell population. Models were evaluated at day 4 (held-out, testing) and day 6 (training) every 100 training epochs. The testing distance is reported for the epoch with the lowest training error (Fig. S1c).

To compute the linear interpolation baseline, we used Waddington-OT (WOT), which uses a similar optimal transport formulation but lacks an explicit parametric form. WOT enables recovery of a held-out time point via linear interpolation using transport maps built between sets of cells in early and late time points. To run WOT, we used python code available on GitHub (https://github.com/broadinstitute/wot). The input to WOT is a set of time-point labeled gene expression matrices and growth rates and the output is an optimal transport map. The optimal transport map was built with the full set of cells with lineage barcodes from day 2 (n = 4638 cells) to day 6 (n = 29,679 cells). The empirical proliferation rates derived from clonal expansion of this set of cells from day 2 to 6 were provided to WOT and three growth iterations were permitted. The parameters for best fit of the optimal transport map were as follows: k = 1, λ = 10,000. With the transport map built between day 2 and day 6, 10,000 cells at day 4 were interpolated using the interpolate_with_ot() function from WOT. This maps a point at the midpoint of each of the pairs in the optimal transport map. The testing distance of these interpolated points from the observed day 4 cells was computed as reported above.

Predicting clonal fate bias. To predict clonal fate bias, models were first trained on data from all three time points. Models were fit on three sets of data: (a) the subset of cells for which lineage tracing data is available, (b) the subset of cells for which no lineage tracing data is available, and (c) all cells. Models were trained for 2500 epochs and evaluated every 500 epochs. Then, cells were simulated until the final time point via the first-order discretization as parameterized by the trained model.

We evaluated the clonal fate bias metric described by Weinreb et al. on the test set as defined in their paper. The ANN classifier that we used to classify cells as Neutrophil, Monocyte or other at the final time point was fit with 10 trees, 20 neighbors and using Euclidean distance in PCA space (50 pcs). The model was first fit on a random 80% split of the data. When evaluated on the held-out 20% test split, the model achieved a macro-average f1-score of 0.98. Splits were stratified by cell type. The model was then re-fit to the full dataset. After classification, the clonal fate bias was then computed as the number of neutrophils divided by the total number of monocytes and neutrophils. Since the model did not always predict any cell within the 2000 sampled trajectories to be a monocyte or neutrophil, we also added a pseudocount of 1. In those cases, clonal fate bias would hence be 0.5.
We observed variation in predictions made by models at each epoch (Fig. S2a). Since no validation set is available to formulate a stopping criterion, we chose to ensemble predictions over the last 5 epochs evaluated (i.e., epoch 2100, 2200, 2300, 2400, 2500) by taking the mean across the estimated clonal fate bias across those epochs. In most cases, performance metrics for WOT, PBA, and FateID were calculated using the predictions already pre-computed and made available by Weinreb et al.

Introducing and evaluating in silico perturbations. Perturbation experiments were performed similarly to the clonal fate bias experiments, except using perturbed cells as input to the first-order discretization. Generally, perturbations were introduced in silico by setting the scaled normalized expression of target genes to zero—values less than zero for knockdowns and greater than 0 for overexpression. The resulting perturbed gene expression profile was then transformed via PCA into lower-dimensional space for input to forward simulation of the trained model. For the in vitro hematopoiesis dataset, a single PRESCIENT model was used for predicting the effect or silico perturbations to this dataset was seed 1 and epoch 2500 of a model trained with a neural network architecture of 2 layers of 400 units. For all experiments, 200 undifferentiated cells (annotations from Weinreb et al.) were randomly sampled from day 2 weighted by KEGG-derived growth rate estimates, resulting in biased sampling for actively proliferative cells. These sampled cells were simulated forward 40 steps with a df of 0.1 to the final time point (day 6). This process was repeated with random initializations of both unperturbed and perturbed cells. Cells at the final time point were then classified using the same ANN classifier used for the clonal fate bias experiments. Relevant TFs for the target cell fate were identified by searching the literature for studies that had experimentally verified sets of TFs involved in early cell fate decisions by progenitor populations and the highly variable feature set was filtered for these TFs. Perturbations were focused on monocytes and neutrophils due to the availability of experimentally correlated or confirmed perturbations for these cell types and the focus of neutrophil/monocyte fate in the Weinreb et al. lineage tracing. The Neutrophil-associated transcription factors were Lyp, Lyp, and Nrk1a1. First, the effect of perturbations to individual genes were tested by perturbing each target gene with a z-score of −2.5 for knockdown and 5 for overexpression. To test if there was a significant shift in neutrophil/monocyte cell fractions at the final time point, Welch’s t-test with two-sided two-tailed p-values were performed between unperturbed simulations and perturbed simulations for each target TF. Next, the effect of perturbational magnitude was evaluated by introducing perturbations of −2.5, −1, −0.5, 2, 5, and 10 to each target gene individually. The same statistical test was performed in comparison to unperturbed simulations. With 30 PCs trained with 2 layers of 400 units across all perturbations and each outcome by perturbing sets of TFs with z-scores of −2.5, −1, −0.5, 2, 5, and 10. As a control, the combined perturbation was repeated with following randomly selected non-TF control genes: Geh, Gfr, Dhr2, Trafi, Lekrk, Lgmn, Il13, and Sgk1. The same statistical test was performed in comparison to unperturbed simulations. For the in vitro beta-cell differentiation dataset, a single PRESCIENT model was trained with 2 layers of 400 units. The 400-unit model was chosen as it achieved a macro-average f1 score of 0.939 when discriminating between β-cells, β-secretory, β-secretory, and EC cells. The ANN classifier was then re-fit on the full dataset. For time point sampling experiments, 200 cells were sampled at each time point weighted by KEGG-derived growth rates, based on metadata from Veres et al. Cells were iteratively sampled from days 0–6 and simulated forward with a dt (step size parameter of 0.1) to the final time point (day 7) under both unperturbed and perturbed conditions. Perturbations (z = −2.5, −1, −0.5, 2, 5, 10) were introduced to sampled cells from each timepoint. To test if there was a significant shift in neutrophil/monocyte cell fractions at the final time point, Welch’s t-test was performed on Welch’s independent two-sided t-tests were performed between unperturbed simulations and perturbed simulations for each target TF. For cell-type subpopulation experiments, 200 cells were randomly sampled from the S0X2 + progenitor, NKX6.1 + progenitor, and NFURG3 early/ mid/late, weighted by KEGG-derived growth rates. For the screen of 200 cells, 200 cells were first simulated without perturbations and the same cells were simulated with perturbations (z = 0) of each TF. This was repeated with 10 random initializations. For each tf, to test for a significant change in sign between all cell-type fractions, two-sided paired t-tests were conducted between the unperturbed and perturbed simulations. To show that TF expression at static timepoints is not necessarily indicative of fate bias, a control analysis was completed on the Veres et al. 2019 dataset by computing the log-fold change of each TF between cell types of interest at day 7 and plotted against the log-fold change in cell fraction predicted by PRESCIENT perturbation analysis (Fig. S4d).

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

Data availability
Data for the Weinreb et al. experiments was downloaded from https://github.com/AllonKleinLab/paper-data/blob/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation/README.md (commit: 8b0969f). Data for the Veres et al. experiments was downloaded from GEO (GSE114121)2. Trained models from this study are available at https://zemodo.org/record/4687634, YHExNBK19, or https://www.github.com/gifford-lab/prescient-analysis.

Code availability
An open-source implementation documentation, and tutorial vignettes of PRESCIENT is available at https://ccs.csail.mit.edu/prescient/. Source code can be found at https://github.com/gifford-lab/prescient. Notebooks to reproduce figures and analyses are available at https://github.com/gifford-lab/prescient-analysis.

Received: 2 October 2020; Accepted: 22 April 2021; Published online: 28 May 2021

References
1. Ferrell, J. E. Jr Bistability, bifurcations, and Waddington’s epigenetic landscape. Curr. Biol. 22, R458–R466 (2012).
2. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. Nature. Rev. Genet. (2020) https://doi.org/10.1038/s41576-020-0223-2.
3. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science (2020) https://doi.org/10.1126/science.aaw3381.
4. Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. Nature 564, 219–224 (2018).
5. Saenz, W., Connood, R., Todorov, H. & Saeyes, Y. A comparison of single-cell trajectory inference methods. Nature Biotechnol. 37, 547–554 (2019).
6. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578 (2012).
7. Tanay, A. & Regev, A. Scaling single-cell genomics from phenology to mechanism. Nature 541, 331–338 (2017).
8. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 176, 928–943.e22 (2019).
9. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nat. Methods 15, 379–386 (2018).
10. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. USA 115, E2467–E2476 (2018).
11. Fischer, D. S. et al. Inferring population dynamics from single-cell RNA-seq sequencing time series data. Nat. Biotechnol. 37, 461–468 (2019).
12. Hashimoto, T., Gifford, D. & Jaakkola, T. Learning population-level diffusions with generative RNNs. Proc. 33rd Int. Conf. Mach. Learn. 48, 2417–2426 (2016).
13. Lottolfah., M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. Nat. Methods 16, 928–943.e22 (2019).
14. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888–1902.e21 (2019).
15. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296 (2019).
16. Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nat. Cell Biol. 20, 836–846 (2018).
17. Feinberg, M. W. et al. The Kruppel-like factor KLF4 is a critical regulator of hematopoietic lineage and identity in direct reprogramming. EMBO J. 26, 4138–4148 (2007).
18. Hannan, K. N. et al. The transcription factor NrfA1 (Nue17) controls bone marrow differentiation and the survival of Ly6C- monocytes. Nat. Immunol. 12, 778–785 (2011).
19. Yamanaka, R. et al. Impaired granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-deficient mice. Proc. Natl. Acad. Sci. USA 94, 13167–13172 (1997).

Nature Communications | https://doi.org/10.1038/s41467-021-23518-w | www.nature.com/naturecommunications

ARTICLE

11
24. Dassaye, R., Naidoo, S. & Cerf, M. E. Transcription factor regulation of pancreatic organogenesis, differentiation, and maturation. *Islets* 8, 13–34 (2016).

25. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl. Acad. Sci. USA* 97, 1607–1611 (2000).

26. Johansson, K. A. et al. Temporal control of neurogenin3 activity in pancreas progenitors reveals competence windows for the generation of different endocrine cell types. *Dev. Cell* 12, 457–465 (2007).

27. Jensen, J. et al. Control of endodermal endocrine development by Hes-1. *Nat. Genet.* 24, 39–44 (2000).

28. Schaefer, K. N. & Peifer, M. Wnt/Beta-catenin signaling regulation and a role for biomolecular condensates. *Dev. Cell* 48, 429–444 (2019).

29. Collombat, P. et al. The simultaneous loss of Arx and Pax4 genes promotes a somatostatin-producing cell fate specification at the expense of the alpha- and beta-cell lineages in the mouse endocrine pancreas. *Development* 132, 2969–2980 (2005).

30. Collombat, P. et al. Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes Dev.* 17, 2591–2603 (2003).

31. Xu, Y. et al. Generation and phenotype identification of PAX4 gene knockout rabbit by CRISPR/Cas9 system. *G3* 8, 2833–2840 (2018).

32. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* 3, 346–360.e4 (2016).

33. Papizan, J. B. et al. Nkx6.2 repressor complex regulates islet β-cell specification and prevents β-to-α-cell reprogramming. *Genes Dev.* 25, 2291–2305 (2011).

34. Schaffer, A. E. et al. Nkx6.1 controls a gene regulatory network required for establishing and maintaining pancreatic Beta cell identity. *PLoS Genet.* 9, e1003274 (2013).

35. Schaffer, A. E., Freude, K. K., Nelson, S. B. & Sander, M. Nkx6 transcription factors and Ptf1a function as antagonistic lineage determinants in multipotent pancreatic progenitors. *Dev. Cell* 18, 1022–1029 (2010).

36. Gannon, M. et al. pdx-1 function is specifically required in embryonic beta cells to generate appropriate numbers of endocrine cell types and maintain glucose homeostasis. *Dev. Biol.* 314, 406–417 (2008).

37. Artner, I. et al. Mafa and MafB regulate genes critical to beta-cells in a unique temporal manner. *Diabetes* 59, 2530–2539 (2010).

38. St-Onge, L., Sosa-Pineda, B., Chowdhury, K., Mansouri, A. & Gruss, P. Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. *Nature* 387, 406–409 (1997).

39. Sander, M. et al. Genetic analysis reveals that Pax6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes Dev.* 11, 1662–1673 (1997).

40. Yang, Y.-P., Thorel, F., Boyer, D. F., Herrera, P. L. & Wright, C. V. E. Context-specific α- to β-cell reprogramming by forced Pdx1 expression. *Genes Dev.* 25, 1680–1685 (2011).

41. Hang, Y. & Stein, R. Mafa and Mafb activity in pancreatic β cells. *Trends Endocrinol. Metab.* 22, 364–373 (2011).

42. Yeo, G. H. T. et al. A multiplexed barcodelet single-cell RNA-seq approach elucidates combinatorial signaling pathways that drive ESC differentiation. *Cell Stem Cell* 26, 938–950.e6 (2020).

43. Liu, Y. et al. CRISPR activation screens systematically identify factors that drive neuronal fate and reprogramming. *Cell Stem Cell* 23, 758–771.e8 (2018).

44. Luginbühl, J. et al. Decoding neuronal diversification by multiplexed single-cell RNA-Seq. *Stem Cell Reports* (2021) https://doi.org/10.1016/j.stemcr.2021.02.006.

45. Shiffman, M. et al. Reconstructing probabilistic trees of cellular differentiation from single-cell RNA-seq data. *arXiv [q-bio.QM]* (2018).

46. Yuan, B. et al. CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Syst.* 12, 128–140.e4 (2018).

47. La Manno, G. et al. RNA velocity of single cells. *Nature* 560, 494–498 (2018).

48. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420 (2018).

49. Saksena, S. D. & Yeo, G. *gifford-lab/prescient*: Software release of PRESCIENT. (2021). https://doi.org/10.5281/zenodo.4683408. 10/02/2020; 04/22/2021.

**Acknowledgements**

We thank Jennifer Hammelman and Ernest Fraenkel for helpful discussion and comments. We gratefully acknowledge funding from NIH grants 5 R01 NS109217, 5 R01 HG008754, and 5 R01 HG008363 (D.K.G.) and the National Science Scholarship (PhD) from the Agency for Science, Technology and Research Graduate Academy (G.Y.).

**Author contributions**

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing, visualization: G.Y. S.S.; conceptualization, writing, formal analysis, supervision, funding acquisition: D.K.G.

**Competing interests**

D.K.G. is a founder of Think Therapeutics. All other authors declare no competing interests.

**Additional information**

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23518-w.

Correspondence and requests for materials should be addressed to D.K.G.

Peer review information *Nature Communications* thanks Mahmoud Ibrahim and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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