nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00679-5

# Extrapolating heterogeneous time-series gene expression data using Sagittarius

Received: 8 November 2022

Accepted: 2 May 2023

Published online: 22 June 2023

Check for updates

Addie Woicik[1], Mingxin Zhang[1,2], Janelle Chan[1], Jianzhu Ma[3,4] & Sheng Wang[1]

Understanding the temporal dynamics of gene expression is crucial for developmental biology, tumour biology and biogerontology. However, some timepoints remain challenging to measure in the laboratory, particularly during very early or very late stages of a biological process. Here we propose Sagittarius, a transformer-based model that can accurately simulate gene expression profiles at timepoints outside the range of times measured in the laboratory. The key idea behind Sagittarius is to learn a shared reference space for time-series measurements, thereby explicitly modelling unaligned timepoints and conditional batch effects between time series, and making the model widely applicable to diverse biological settings. We show Sagittarius's promising performance when extrapolating mammalian developmental gene expression, simulating drug-induced expression at unmeasured dose and treatment times, and augmenting datasets to accurately predict drug sensitivity. We also used Sagittarius to extrapolate mutation profiles for early-stage cancer patients, which enabled us to discover a gene set connected to the Hedgehog signalling pathway that may be related to tumorigenesis in sarcoma patients, including *PTCH1*, *ARID2* and *MYCBP2*. By augmenting experimental temporal datasets with crucial but difficult-to-measure extrapolated datapoints, Sagittarius enables deeper insights into the temporal dynamics of heterogeneous transcriptomic processes and can be broadly applied to biological time-series extrapolation.

The temporal dynamics of the transcriptome are key to the study of developmental biology[1,2], tumour biology[3,4], immunobiology[5,6] and pharmacogenomics[7,8]. As bulk and single-cell RNA-sequencing technologies have become cheaper[4,9–11], more transcriptomic datasets include gene expression measurements at multiple timepoints[12–19]. Still, it often remains challenging to measure transcriptomic profiles at very early or late stages of a biological process. For instance, senescent and extremely diseased tissues can be challenging to measure, but are of extreme interest for ageing and therapeutics.

The underlying problem here is temporal extrapolation, where timepoints of interest are outside the time range of experimental measurements. Accurate extrapolation on a single time series is challenging due to non-stationary features and temporal out-of-domain adaptation[20]. One possible solution for the extrapolation problem is to combine sparse time-series measurements from heterogeneous sequences. For example, mouse[12] and roundworm[21] transcriptomic time-series measurements, combined with developmental human measurements, can help simulate early-stage embryonic transcriptomic profiles for humans[22]. There are two major challenges in effectively utilizing other time series: unaligned measured timepoints and batch effects between experimental conditions. Existing methods are unable to simultaneously consider the full sequence of measured

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan. [3]Department of Electrical Engineering, Tsinghua University, Beijing, China. [4]Institute for AI Industry Research, Tsinghua University, Beijing, China. e-mail: swang@cs.washington.edu

timepoints[23,24] or take into account the temporal batch effects between time series[25–28].

To address these limitations we propose Sagittarius, a model that maps heterogeneous gene expression time series to a shared reference space on the basis of inferred biological age rather than the observed age, enabling multiple sparsely measured time series to jointly inform extrapolation. Sagittarius leverages a transformer-based architecture with multihead attention[29] to map the heterogeneous measurements from the irregular, unaligned and sparse time series to a latent reference space shared by all time series, using high-frequency sinusoidal embeddings of the timestamp[27,30] and experimental condition labels of each time series to define the mapping. After alignment in the shared reference space, Sagittarius can accurately predict new genomic profiles at extrapolated timepoints, as well as predicting measurements for unmeasured combinations of experimental conditions.

We evaluated Sagittarius in three diverse settings in developmental biology, pharmacogenomics and cancer genomics. On the Evo-devo development dataset[12], Sagittarius accurately extrapolated gene expression profiles with a 0.983 Pearson correlation, enabling an in-depth analysis of mouse organ differentiation. To evaluate Sagittarius's robustness to extremely sparse measurements, we next applied it to the Library of Integrated Network-Based Cellular Signatures (LINCS) pharmacogenomics dataset[15] and found that Sagittarius was able to predict drug-repurposing opportunities across drugs and cell lines. Finally, we applied Sagittarius to The Cancer Genome Atlas (TCGA) dataset[31], where Sagittarius was able to accurately extrapolate mutation profiles for patients with a long survival time. Our findings implicated a gene set related to the Hedgehog (HH) signalling pathway and *GLI* oncogene that can potentially drive tumorigenesis in sarcoma patients.

## Results

### Overview of Sagittarius

Given a heterogeneous, unaligned, sparse and irregular genomic time-series dataset, Sagittarius is able to extrapolate gene expression profiles for unmeasured timepoints (Fig. 1). Fundamentally, we hypothesize that the input time series follow a common latent trajectory, such as a general developmental trajectory shared by humans and model organisms. The key idea behind Sagittarius is to learn a shared reference space that models this trajectory. We use a transformer-based architecture to map each input measurement to and from the reference space, where the learnable mapping is informed by the experimental conditions associated with the time series. This parameterization addresses both temporal extrapolation and batch effects between experimental conditions (Methods). During inference, Sagittarius can extrapolate gene expression profiles for a timepoint and experimental condition of interest.

### Extrapolating gene expression to unmeasured timepoints

To assess the merit of our approach, we evaluated whether Sagittarius can extrapolate profiles for gene expression time series from multiple experimental conditions. We used the Evo-devo time-series data, which contain bulk RNA-Seq data from seven species and seven organs, where each time series ranges between 9 and 23 distinct measured timepoints that are not biologically aligned across species. Importantly, the developmental ranges measured by each species differ: primates include senescence measurements, while rhesus macaque and chicken do not contain early embryonic data. Therefore, the Evo-devo dataset can be used to assess whether Sagittarius can handle unaligned timepoints and differing biological ages measured across species.

To initially validate our model, we hid the last four measured timepoints from each species' organ time series to use as a test set. After training on the remaining Evo-devo data, we predicted the gene expression vectors for each species and organ combination at the four hidden timepoints and compared them with the held-out expression vectors. To benchmark Sagittarius's performance, we also evaluated the deep learning methods Conditional Variational Autoencoder[23] (cVAE), Compositional Perturbation Autoencoder[24], Multi-Time Attention Network[27], PRESCIENT[32], neural ordinary differential equation[26,28] and recurrent neural network[25], as well as classical mean and linear methods. Overall, Sagittarius achieved the best average performance between the extrapolated and measured gene expression profiles in terms of Pearson correlation comparing genes ($\rho = 0.983$), Pearson correlation comparing timepoints ($\rho = 0.458$) and root mean squared error (RMSE = 0.087), compared with 0.926, 0.142 and 0.163 respectively for the best-performing comparison approach (Methods and Supplementary Figs. 4 and 5), and this improvement was robust to many hyperparameter settings (Supplementary Note 1 and Supplementary Figs. 6–8). Sagittarius was also able to accurately extrapolate expression for many genes, with at least a 0.3 Pearson correlation comparing timepoints for more than 55% of the modelled genes (Methods).

We further stratified our extrapolation results by species and organ (Fig. 2). We found that our model achieved the best performance on all species and organs, with best absolute performance on the mouse testis time series. This demonstrates the benefit of the shared reference space, as the mouse's final training timepoint is postnatal day 0 (P0) but the model is able to learn from later development in other species to inform extrapolation for the mouse. In contrast, the two worst-performing species were the human and chicken, which we believe to reflect larger distributional shifts in the data. All methods struggled on human test data, which are at much later developmental stages than the training dataset. We therefore conducted an analogous Evo-devo experiment, this time extrapolating to the earliest four timepoints as test data (Methods). We found that Sagittarius was still the best-performing method, and had stronger performance for extrapolation to early-stage human development (Supplementary Figs. 9 and 10), supporting this hypothesis. We believe that the relatively poor chicken performance also stems from a distributional shift, as chicken is the only non-mammal in the dataset. Therefore, the chicken time series are less evolutionarily related to the other time series in the dataset, and may be harder to represent as a transformation of a shared, otherwise mammalian, developmental trajectory[12]. After better understanding Sagittarius's strengths and weaknesses, we then studied whether Sagittarius could simulate samples for unmeasured timepoints to gain new insights into tissue differentiation and ageing.

### Transcriptomic dynamics reveal organ-specific ageing genes

To further examine the Sagittarius extrapolated expression profiles, we next predicted developmental trajectories for each mouse organ, beginning at embryonic day 5.5 (E5.5) and continuing to P63. By extrapolating to early timepoints, we expect to observe a hypothetical trajectory that includes organogenesis, which takes place between E6.5 and E8.5 in mouse development[33,34]. Specifically, we expect that the earliest extrapolated timepoints will result in very similar expression profiles across the different queried organs, which would not have differentiated at this stage. On subsequent days, we would then expect the organs to diverge according to germ layer, before finally separating by organ[12,13,33,35,36]. We visualized the uniform manifold approximation and projection[37] (UMAP) embedding of the simulated organ time series (Fig. 3a,b), as well as the top principal components[38] (Supplementary Fig. 12). Our findings largely aligned with the understanding of mouse organogenesis. Namely, the developmental stage dominated the gene extrapolation measurements at the earliest timepoints, with multiple organs grouped in the same location of the UMAP space. At later timepoints, we found that the predicted expression values for brain and cerebellum were more closely grouped together, as were expression measurements for the heart, ovary and testis, consistent with the ectoderm, mesoderm and endoderm tissue germ layer classifications[12].

Given the increasing tissue-specific signal in Sagittarius's simulated gene expression vectors at later timepoints, we then investigated which genes most contributed to the differentiation of organ

https://doi.org/10.1038/s42256-023-00679-5

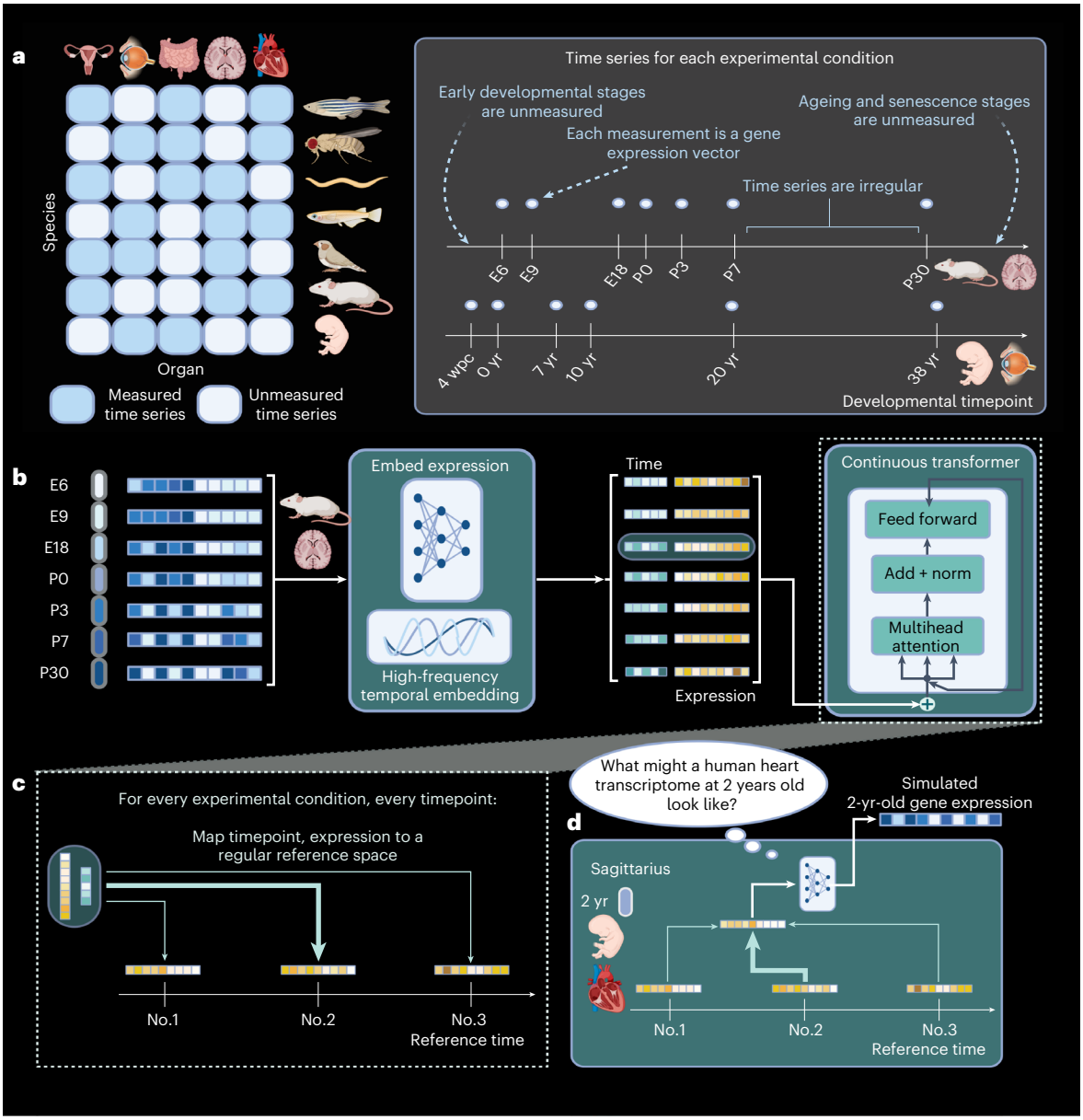


**Fig. 1 | Sagittarius model overview. a**, Sagittarius is useful in settings with many diverse time-series measurements, such as developmental gene expression data across species and organs, many combinations of which are unmeasured. The measurements in each time series are also sparse and unaligned, such as irregularly sampled mouse brain measurements from embryonic day 6 (E6) to postpartum day 30 (P30) and human eye measurements from 4 weeks-post-conception (4wpc) to 38 years (38 yr). **b**, For each time series, Sagittarius computes a conditional high-frequency sinusoidal embedding of the measured timepoints and a conditional embedding of the gene expression measurements at each timepoint on the basis of the species and organ. It then uses a continuous, multihead attention transformer to map the embedded timepoints and expression vectors to the reference space. **c**, The continuous transformer takes each pair of species- and organ-conditioned time and expression embeddings and learns a mapping to the regular reference space, translating from measured age to a shared biological age. **d**, Users can request extrapolated expression vectors from Sagittarius, such as the expression profile of a human 2-year-old heart that has not been measured in the original dataset (**a**). Sagittarius maps the request from the regular reference space back to the data space to predict the unmeasured profile.

trajectories during development. Excluding the heart and cerebellum, which we found to be the most developmentally distinct for many genes, we found that mouse *Xrn2* expression levels were comparable across organs at early extrapolated timepoints but differed significantly at later timepoints (analysis of variance (ANOVA) $P > 0.05$ and $P < 1 \times 10^{-98}$ respectively), with lower expression levels in the liver than other organs at late timepoints (Fig. 3c). Existing work has found that mouse *Xrn2* and its roundworm orthologue *xrn-2* play important biological roles during development[39–46], and *XRN2* overexpression in humans has also been tied to poor liver cancer prognosis[47].

We next sought to further examine Sagittarius's organ-specific extrapolation potential in mouse. We predicted a transcriptomic profile trajectory beginning at P14 (Methods). The latest mouse measurement in the Evo-devo dataset is taken at P63, so we used the Tabula Muris Senis single-cell RNA-Seq dataset[16], which spans from a 1-month-old mouse to a 30-month-old mouse, to validate our results. We compared the Pearson correlation of the gene expression over time between the extrapolated profiles and the Tabula Muris Senis data for each tissue, and, for mouse genes including *Egflam*, *Smoc1*, *Slc6a2* and especially *Rpl38*, which previous work has suggested

https://doi.org/10.1038/s42256-023-00679-5

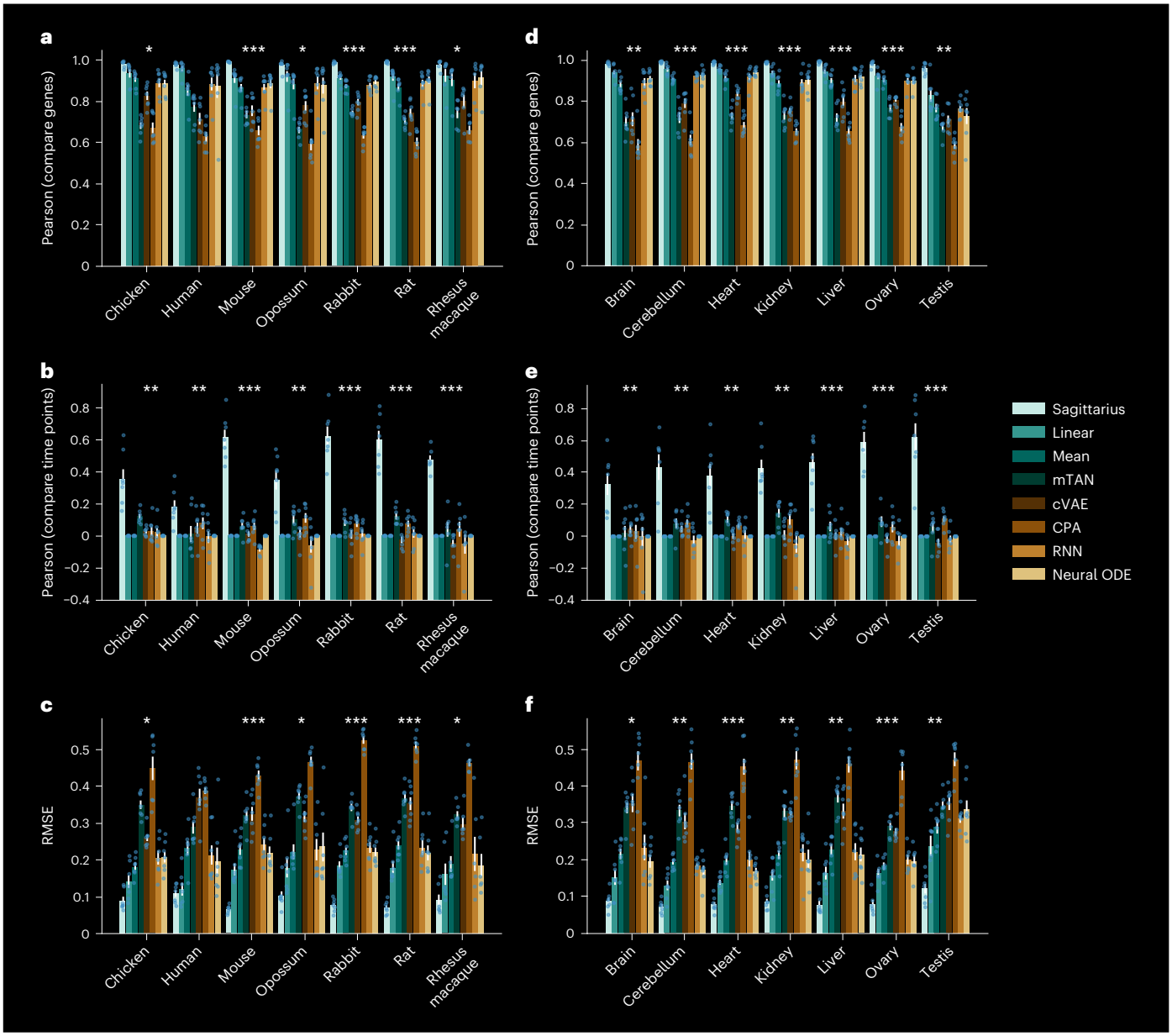


**Fig. 2 | Gene expression prediction for extrapolated timepoints in later-stage development. a–f**, Bar plots comparing the performance of Sagittarius and existing approaches when extrapolating to the four latest timepoints in the Evo-devo dataset. mTAN, Multi-Time Attention Network; CPA, Compositional Perturbation Autoencoder; RNN, recurrent neural network; Neural ODE, neural ordinary differential equation. Test sequences are subdivided by species (**a**–**c**) and by organ (**d**–**f**). For Pearson correlation, comparing genes (**a**,**d**) or comparing timepoints (**b**,**e**), higher correlations indicate better performance; for RMSE (**c**,**f**), lower error indicates better performance. Data are presented as mean values ± s.e.m. When stratified by species (**a**–**c**), $n = 6$ organ time series for Rhesus macaque and $n = 7$ organ time series for all other species. When stratified by organ (**d**–**f**), $n = 6$ species time series for ovary and $n = 7$ species time series for all other organs. The asterisks indicate that Sagittarius outperforms the next-best-performing model in the metric, with significance levels of $^{*}P < 5 \times 10^{-2}$, $^{**}P < 5 \times 10^{-3}$ and $^{***}P < 5 \times 10^{-4}$. We use a one-sided Fisher $z$-transformed test for Pearson correlation comparing genes and comparing timepoints, and a one-sided $t$-test for RMSE.

could regulate developmental processes in a tissue-specific way[48], found that Sagittarius's extrapolated ageing trajectory aligned with the Tabula Muris Senis tissue measurements better than the younger mouse trajectory taken directly from Evo-devo (Methods and Fig. 3d). We attribute this to the shared reference space, which can identify ageing and senescence patterns from other species such as human and rhesus macaque to inform transcriptomic extrapolation for mouse ageing. After applying Sagittarius to the Evo-devo dataset, we next considered whether the model could successfully extrapolate unmeasured experimental combinations in settings with multiple continuous variables.

## Sagittarius simulates unmeasured drug perturbations

We next evaluated Sagittarius on extremely sparse multivariate data with multiple continuous temporal variables, thereby exponentially increasing the space of possible experimental settings. We applied Sagittarius to the larger, high-dimensional LINCS L1000 pharmacogenomics dataset[15]. In the LINCS dataset, compounds are experimentally applied to cell lines at specific doses and for a given treatment time before measuring the drug-induced expression profiles, although only 1.77% of possible drug and cell line combinations are screened (Fig. 4a). Sagittarius models each treatment experiment in two continuous dimensions: dose and treatment time.

https://doi.org/10.1038/s42256-023-00679-5

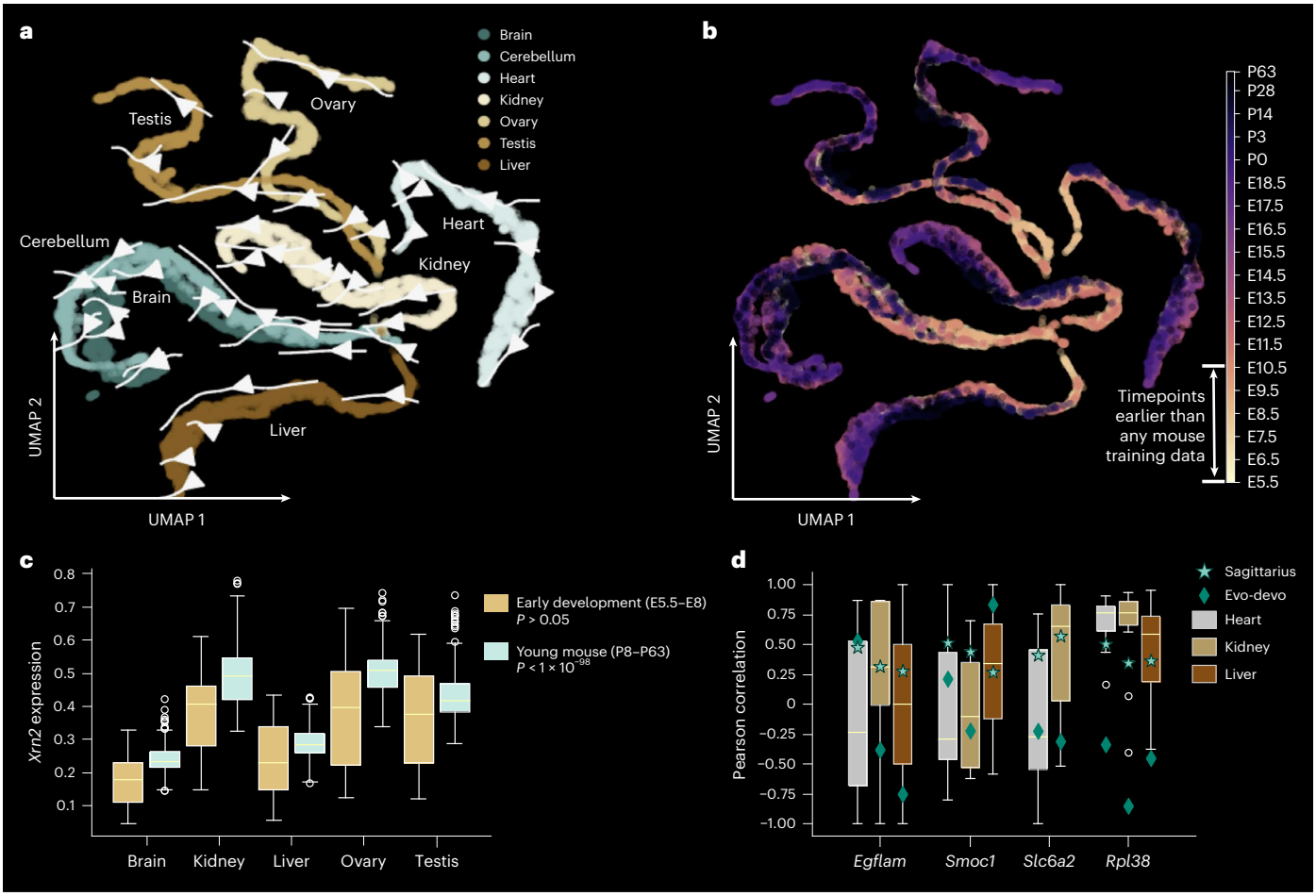


**Fig. 3 | Mouse transcriptomic velocity across organs. a**,**b**, UMAP plots showing simulated mouse gene expression from E5.5 to P63 for seven organs, coloured by organ (**a**) and time (**b**). The arrows in **a** indicate the transcriptomic velocity of each organ. **c**, Box plot comparing the simulated expression of *Xrn2* at early development (E5.5–E8) with young mouse (P8–P63) across five organs, with $n = 250$ simulated measurements per organ. *Xrn2* expression is not statistically different between the brain, kidney, liver, ovary and testis organs in early development (ANOVA $P = 0.83$), but differs between organs for the young mouse time range, in particular with lower expression levels in the liver relative to other organs (ANOVA $P = 9.37 \times 10^{-78}$). **d**, Box plot examining the consistency of gene expression temporal patterns between simulated data and single-cell RNA sequencing data for *Egflam*, *Smoc1*, *Slc6a2* and *Rpl38* in different mouse tissues over time. Boxes indicate the distribution of cell-type correlations for each tissue in Tabula Muris Senis, with $n = 7, 5, 2$ cell types in heart, kidney and liver respectively for *Egflam*; $n = 7, 6, 3$ respectively for *Smoc1*; $n = 5, 7$ in heart and kidney respectively for *Slc6a2* and $n = 9, 19, 8$ cell types in heart, kidney and liver for *Rpl38*. Better predictions are closer to the distribution of Tabula Muris Senis cell-type correlations for each tissue. The stars show the Pearson correlations from Sagittarius's simulated correlation for ageing mouse tissues (140 timepoints beginning at P14), and the diamonds show the correlation with respect to time of the younger mouse organ measurements in the Evo-devo dataset. In **c**,**d**, the box bounds show the interquartile range from quartiles 1 to 3, with the centreline indicating the median and the whiskers extending 1.5 times the interquartile range from the box.

To validate Sagittarius's ability to extrapolate to new perturbation experiments, we then designed three extrapolation tasks: complete generation, combination & dose and combination & time (Methods and Fig. 4b). For each task, we trained Sagittarius on a subset of the LINCS dataset, withholding the remaining measurements as test data. We compared Sagittarius's performance with a cVAE, the only comparison approach that could handle multiple continuous variables off the shelf. Evaluating model predictions on the basis of held-out test data, we found that Sagittarius had an average Spearman correlation of 0.93, 0.92 and 0.88 for the three tasks respectively, compared with 0.85, 0.88 and 0.81 for the cVAE (Fig. 4c–e, one-sided Fisher $z$-transformed test $P < 5 \times 10^{-225}$, $5 \times 10^{-92}$ and $5 \times 10^{-301}$ respectively). This indicates that aligning all perturbation experiments to the shared reference space enables Sagittarius to accurately extrapolate drug-induced gene expression vectors for unmeasured drug treatment experiments at doses and times that are not contained in the training data, enabling an easy, unbiased search approach to drug sensitivity markers. This may greatly increase our understanding of the molecular basis of cancer and of drug response.

## A drug sensitivity similarity network for drug repurposing

Gene expression has been widely used to identify the drug-induced and diseased-induced gene expression signatures in drug-repurposing studies[49–51], partly due to the scale at which analyses can be efficiently performed and validated. As Sagittarius can accurately predict expression for any perturbation combination, we applied Sagittarius to drug repurposing by constructing a similarity network of extrapolated perturbation experiments (Methods). We found that communities within the network demonstrated a pattern with respect to treatment sensitivity, with average half-maximal inhibitory concentration ($IC_{50}$) doses of 1.68, 1.83, 1.90 and 2.40 μM in the Genomics of Drug Sensitivity in Cancer (GDSC) dataset[52] (Fig. 5a). To further investigate the potential for drug-repurposing opportunities, we conducted a case study on an eight-experiment subgraph from the sensitive community, shown in the inset of Fig. 5a.

https://doi.org/10.1038/s42256-023-00679-5

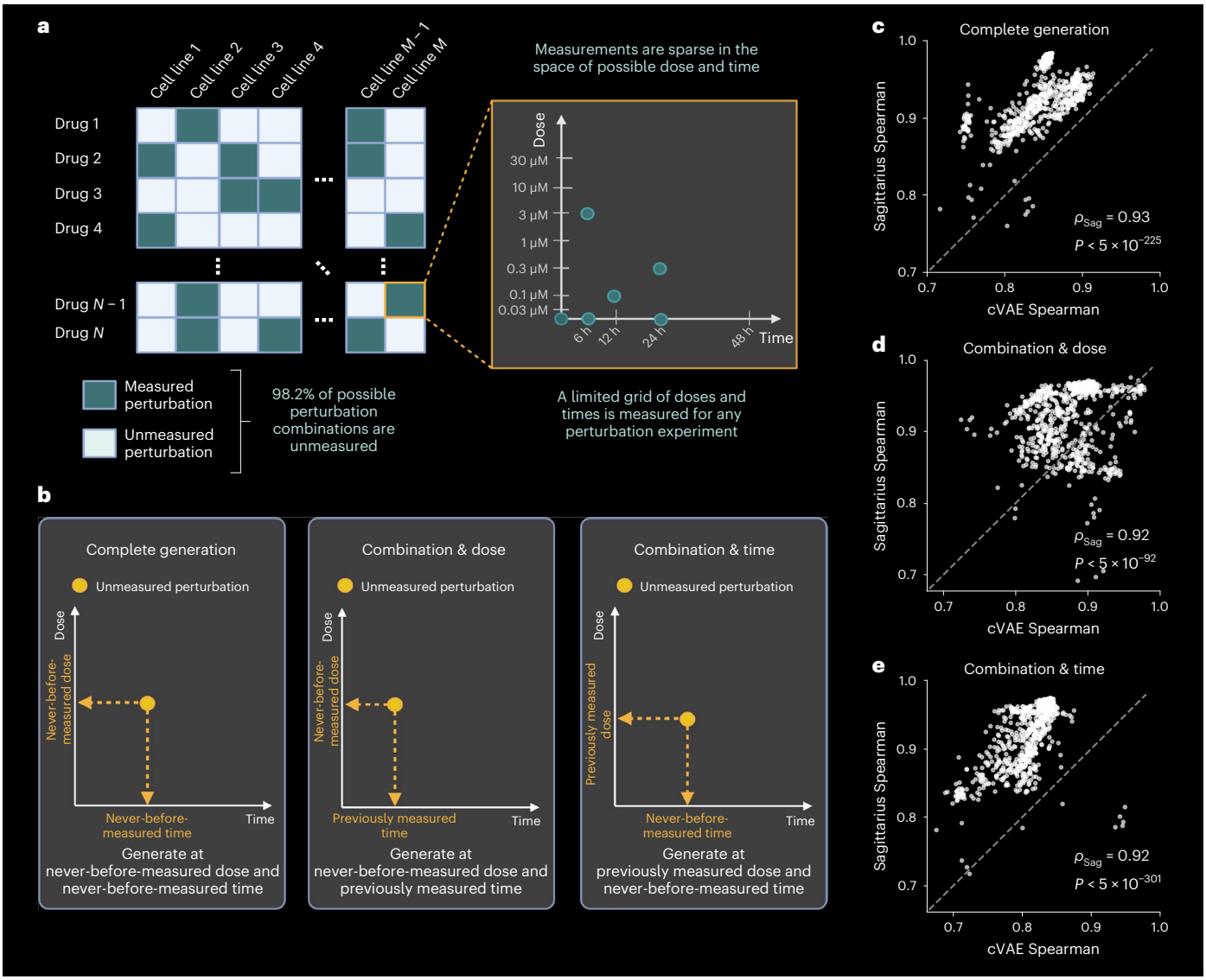


**Fig. 4 | Drug-induced gene expression extrapolation at unmeasured experimental combinations, doses and times. a**, The LINCS pharmacogenomic dataset contains gene expression measurements from a set of experiments where a cancer cell line is treated with a therapeutic compound. The set of measured cell lines and compounds is sparse, with less than 1.77% of possible experiments measured. The measured experiments are also only measured at selected dose and treatment times, and the entire dataset includes a limited number of dose and treatment times. **b**, Illustration of the three extrapolation tasks we evaluate for the LINCS dataset: complete generation, where we predict an unmeasured cell line and compound experiment at both a dose and a time that are unmeasured by any experiment in the dataset; combination & dose, where we predict an unmeasured cell line and compound experiment at a time that has been measured in the dataset but a dose that is unmeasured by all experiments; and combination & time, where we predict an unmeasured cell line and compound experiment at a dose that has been measured in the dataset but a time that is unmeasured by all experiments. **c**–**e**, Scatter plots comparing the average Spearman correlation of simulated test combinations from Sagittarius and the existing cVAE model for each test drug on the complete generation (**c**), combination & dose (**d**) and combination & time (**e**) extrapolation tasks. Sagittarius's average correlation across test datapoints is reported, along with the one-sided Fisher *z*-transformed test *P* values testing whether Sagittarius outperforms cVAE in each setting, with $P = 3.94 \times 10^{-225}$, $P = 2.73 \times 10^{-92}$ and $P = 2.50 \times 10^{-301}$ respectively.

The subnetwork includes six perturbations for the breast carcinoma cell line MCF7 and non-small cell lung cancer cell line A549, all of which are measured in the GDSC dataset. A549 is sensitive to treatment with vorinostat, gefitinib and selumetinib ($IC_{50}$ of 0.49, 0.67 a 0.83 μM respectively), and MCF7 is sensitive to treatment with palbociclib, MK-2206 and gefitinib ($IC_{50}$ of 0.40, 0.89 and 1.03 μM respectively). The existence of edges between different drugs for the same cell line and the edge between two cell lines for the same drug connects to drug-repurposing work based on cell line gene expression signatures[53] and drug mechanisms of action[54].

Importantly, by comparing extrapolated differential expression signatures with those of successful treatments, Sagittarius enables drug-repurposing recommendations where neither the drug nor cell line needs to occur in a known successful therapy. The eight-perturbation subnetwork also includes the prostate adenocarcinoma cell line PC3 treated with pictilisib and the colorectal adenocarcinoma cell line HT29 treated with nintedanib, although these drugs and cell lines do not appear elsewhere in the subnetwork. The GDSC dataset does not screen either of these treatment combinations, but previous work has found that pictilisib inhibited PC3 proliferation ($IC_{50}$ = 0.28 μM)[55] and nintedanib demonstrated antitumour efficacy in HT29 xenograft models[56,57] and cell lines ($IC_{50}$ = 1.40 μM)[57]. This implies that Sagittarius can extrapolate perturbation experiments to identify candidate drug-repurposing targets across cell lines, cancer types and therapeutic compounds, creating new

https://doi.org/10.1038/s42256-023-00679-5

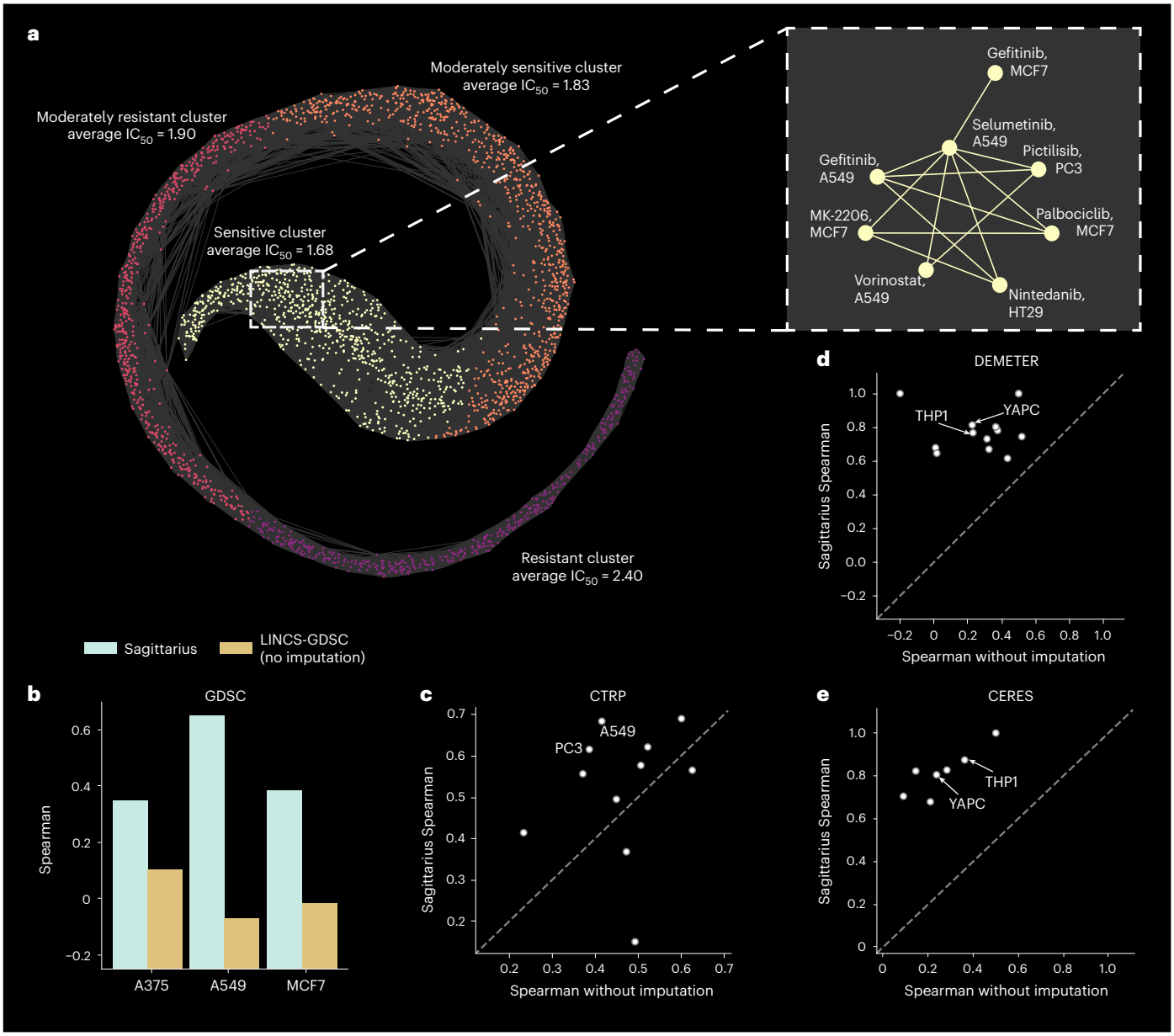


**Fig. 5 | Drug and cell-line treatment efficacy extrapolation analysis.** **a**, *k*-nearest-neighbour network where each node represents a drug and cell-line combination, with edges between the most similar drug-induced expression effects. The four communities in the graph are shown in different colours and labelled according to the average GDSC-measured $IC_{50}$ dose of that community, measured in micromolar. The inset shows a connected eight-node subgraph from the sensitive community, made up of the non-small-cell lung cancer cell line A549 treated with selumetinib, gefitinib and vorinostat; the breast cancer cell line MCF7 treated with gefitinib, MK-2206 and palbociclib; the prostate carcinoma cell line PC3 treated with pictilisib and the colorectal adenocarcinoma cell line HT29 treated with nintedanib. **b**,**c**, Bar plot (**b**) and scatter plot (**c**) of Spearman correlation between predicted and GDSC-measured (**b**) or CTRP-measured (**c**) $IC_{50}$ doses per cell line, comparing a neural network trained with imputed data from Sagittarius and a neural network trained with only the GDSC or CTRP treatment combinations that are also measured in LINCS. Each cell line has one correlation coefficient, meaning $n = 1$ cell-line sensitivity ranking for the bar plot in **b**. **c**, Points above the $y = x$ line are cell lines for which Sagittarius's imputed dataset improved the downstream prediction accuracy. **d**,**e**, Scatter plots of Spearman correlation between predicted and DepMap-measured cancer gene essentiality scores for each cancer line, with the DEMETER (**d**) and CERES (**e**) DepMap versions. A neural network trained with imputed data from Sagittarius is compared with a model trained on the LINCS treatment combinations that correspond to cell-line and gene pairs in the DEMETER or CERES datasets. All points are above the $y = x$ line, meaning that Sagittarius improved downstream gene essentiality prediction performance for all cell lines on both DepMap versions.

opportunities for inexpensive and unbiased drug screening as an initial step in the precision medicine pipeline.

## Perturbation augmentation improves drug response prediction

Given its drug-repurposing potential, we systematically evaluated Sagittarius on two large-scale cell line drug response prediction datasets, GDSC and the Cancer Therapeutic Response Portal (CTRP) dataset[58]. Drug-induced expression profiles have been useful for drug response prediction[59], but are expensive to measure compared with basal cell line expression, making Sagittarius's extrapolated data especially valuable. We used a downstream neural network to predict the $IC_{50}$ label for treatment perturbations, and compared the performance of a model trained on extrapolated data from Sagittarius with a model

trained on the available measured perturbations in LINCS (Methods). When trained on data from Sagittarius, the downstream model had an average Spearman correlation of 0.46 and 0.52 per test cell line, comparing the predicted drug sensitivities with the measured sensitivities in GDSC and CTRP respectively (Fig. 5b,c). In comparison, the model trained on the available experimentally measured data had an average correlation of 0.004 in GDSC and 0.42 in CTRP. We attribute the poor GDSC performance to the low overlap in treatments screened by the two datasets and, consequently, to the extremely small size of the training dataset. Sagittarius, in contrast, was able to extrapolate missing profiles and improve downstream performance, particularly for cell lines and drugs that were among the most frequently measured in the LINCS dataset (Supplementary Figs. 14 and 15). This shows that Sagittarius can take advantage of the many perturbation experiments to inform better predictions for each drug and cell line, even when applied to unmeasured or sparsely measured combinations.

### Improved cancer-essential gene prediction using Sagittarius

In addition to drug response, we analysed Sagittarius's ability to predict cancer gene essentiality. We used the Cancer Dependency Map (DepMap), considering both the DEMETER[60] and the CERES[61] versions. We used a downstream neural network to predict gene essentiality in a given cell line, where we featurized each cell line and gene pair using a drug-induced expression vector from the given cell line and a LINCS drug with the target gene of interest (Methods). We trained one version of the downstream model on extrapolated data from Sagittarius and another on the available experimentally measured drug-induced profiles in LINCS. Comparing the predicted gene essentiality scores with the labels in DepMap, the model trained on data from Sagittarius obtained an average test Spearman correlation of 0.789 and 0.816 per cell line for DEMETER and CERES respectively, compared with 0.278 and 0.261 for the model trained solely on the in vitro LINCS data (Fig. 5d,e). Again, we found that the Sagittarius-backed model particularly improved predictions for well measured LINCS cell lines including the THP1 leukaemia and YAPC pancreatic cell lines (Supplementary Fig. 16). We attribute the strong performance across many different cancer types and drug targets to the shared reference space, where dose and treatment-time response can be compared across cancer cell lines and compounds.

### Simulating mutation profiles for early-stage cancer patients

Having extrapolated gene expression time series in one and two continuous dimensions, we then sought to apply Sagittarius to cancer survival time data. We focus on extrapolating somatic mutation profiles as strong signals of disease, but also validate our results on patient gene expression profiles (Methods and Supplementary Fig. 18). It remains very challenging to measure genomic profiles from patients with nascent tumours, as they are rarely diagnosed at this stage[62], yet these initial mutations can be the most informative as to the cancer's mechanisms and potential early-intervention therapies before other passenger mutations accumulate[63]. We therefore designed a time-series formulation for patient data from 24 cancer types in the TCGA dataset[31], where extrapolation to later timepoints corresponds to the mutation profiles of patients with longer survival times (Fig. 6a). To account for censored event times, we leverage recent machine learning techniques[64,65] to exclude censored patients whose event time probably differs from their survival time (Methods). As a result, the sarcoma cancer type time series contains 115 patients, 31 of whom had a censored death event (Fig. 6b).

Focusing on the sarcoma and thyroid carcinoma (THCA) cancer types as case studies, we designated the longest-surviving patients as test data and used Sagittarius to extrapolate mutation profiles with the same survival times. Varying the number of test patients to examine the models' performance in different settings, we found that Sagittarius had an average test set area under the receiver operating characteristic (AUROC) of 0.72 and 0.73 between the extrapolated and measured mutation profiles for THCA and sarcoma respectively (Methods and Fig. 6c,d), representing 45% and 11% improvements respectively over classical methods, and also improved over other deep learning methods (Supplementary Fig. 21). As a further case study, we focused on a sarcoma patient with a 76 month postdiagnosis survival time, where Sagittarius had particularly improved the test AUROC (Fig. 6e). The patient had a mutation in *LRP1B*, but the mean method assigned zero probability to this mutation, reflecting the observed distribution of sarcoma patients with worse prognosis. Meanwhile, Sagittarius predicts *LRP1B* as the fourth most likely mutation, perhaps learning that *LRP1B* mutations are associated with good therapeutic response in many cancer types[66]. Sagittarius also assigns higher likelihood to common sarcoma mutations such as *ADGRV1*[67], suggesting that the model can jointly leverage patterns within the sarcoma training data and patterns from other cancer types to improve extrapolation.

### HH signalling pathway in simulated early-stage sarcoma

Having confirmed our ability to extrapolate mutation profiles for sarcoma patients with longer survival times, we retrained Sagittarius on all cancer type time series and then extrapolated gene mutation profiles for 27 early-stage sarcoma patients, resulting in the most likely mutated gene set *DNAH17, PREX1, EGFLAM, FAM47B, DSEL, ARID2, TRPM1, NLGN1, PTCH1* and *MYCBP2* (Methods and Supplementary Fig. 23). We found that many of these genes are related to the HH signalling pathway and improper activation of the *GLI* oncogene (Fig. 6f), which has been connected to improved survival outcomes in sarcoma patients[68,69]. For instance, *PTCH1* is a tumour suppressor gene[70] in the HH pathway connected to some sarcomas[70–72]; *MYCBP2* encodes a protein that has been shown to interact with *GLI*[73] via *MYC* upregulation[74,75]; the protein encoded by *ARID2* directly interacts with *GLI1*[76,77]; *DNAH17* encodes a protein that affects the HH pathway through its role in the primary cilia[78,79] and *PREX1* encodes a protein whose pathway has been associated with *GLI* code regulation[80] and cross-talk with the HH pathway in melanoma[72,81]. In addition to these connections to the *GLI* oncogene, *EGFLAM* has been shown to induce activation of *PREX1*[82], and *NLGN1* encodes a protein that was found to be enriched with the HH pathway in colorectal carcinoma[83] (Supplementary Table 2). Furthermore, previous gene expression analyses found that the HH signalling pathway was enriched for differentially expressed genes in multiple sarcomas[84,85]. Sagittarius's extrapolated mutation profiles therefore point to the HH signalling pathway and particularly the hyperactivation of the *GLI* oncogene as potentially important sources of tumorigenesis in sarcomas.

## Discussion

Sagittarius enables extrapolation of genomic profiles from sparse, heterogeneous time series without requiring aligned timepoints or batch correction between different experimental conditions. Although Sagittarius can extrapolate to unseen timepoints, the model still struggles with large domain shifts between developmental stages in training and test, as we identify in the Evo-devo human extrapolation task. Similarly, Sagittarius is unable to extrapolate to precise timepoints. The learned mapping to and from the shared reference space enables comparison between diverse and unaligned time series, but also warps the queried and measured timepoints to align with biological age and thereby alters the timepoints outside the range of the dataset in potentially unforeseen ways. In the future, Sagittarius could be integrated with models to predict the chronological and biological age associated with a new sample. Furthermore, Sagittarius could benefit from work in model calibration[86] to output a model confidence score along with a predicted profile.

Sagittarius is inspired by decades of work in modelling cell dynamics. One key difference between Sagittarius and pseudotime cell fate models such as Monocle[13], Palantir[87] and Slingshot[88] is that these models reconstruct cell lineage within the bounds of one or more originator states and one or more terminal states, while Sagittarius is able to

https://doi.org/10.1038/s42256-023-00679-5

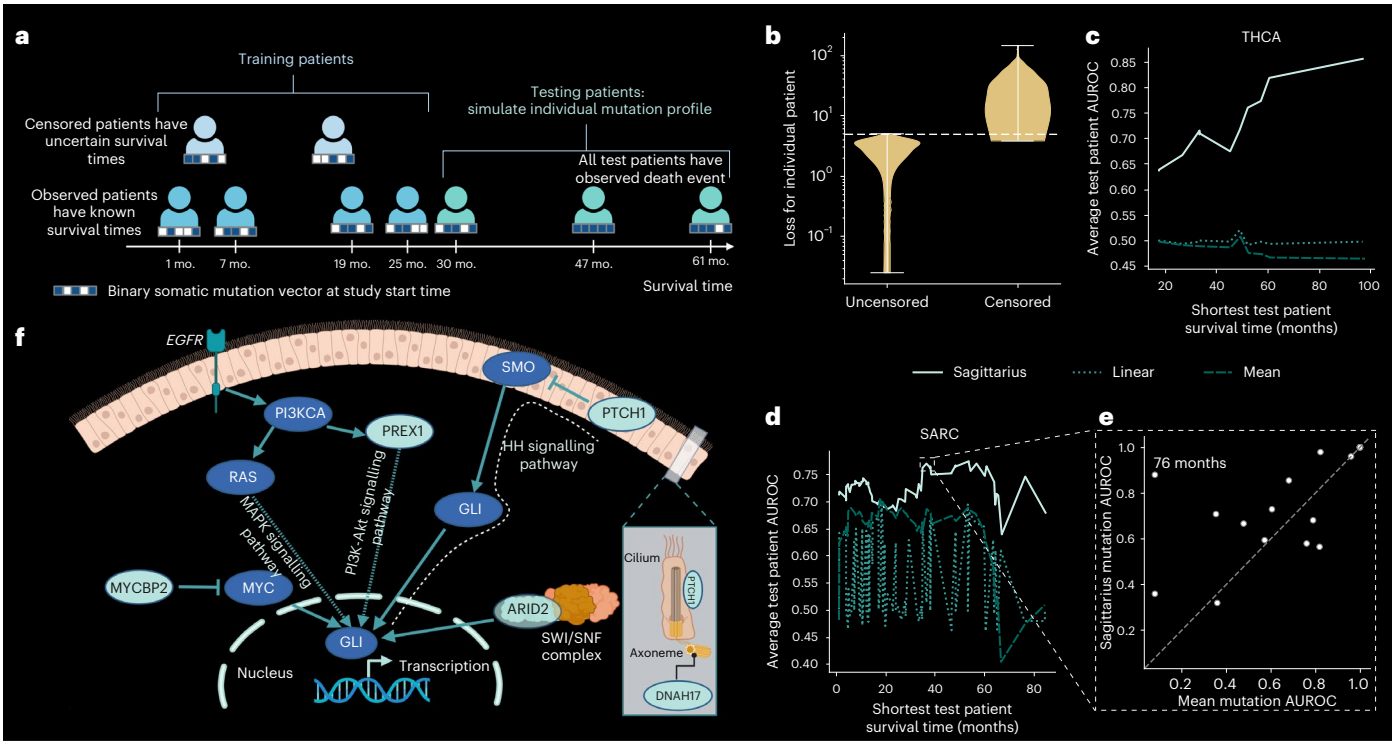


**Fig. 6 | Early cancer patient mutation profile extrapolation. a**, Illustration of the training and testing splits for a given cancer type in the TCGA extrapolation task, where training patients have the shortest survival times and test patients have longer survival times for that cancer type. mo., months. **b**, Violin plot of the survival time regression model's absolute error per patient for the sarcoma cancer type, divided according to the patient's censoring label. We remove all patients with a loss above the dashed line from the dataset, and train Sagittarius only on the patients below the dashed line. The filled area shows the distribution of the full data range, with additional ticks at the distribution extrema for clarity. **c**,**d**, Plot of the average predicted mutation profile AUROC for each of the THCA (**c**) and sarcoma (SARC) (**d**) cancer type test splits, ordered according to the shortest survival time in months in that test split. **e**, Scatter plot comparing the per-patient predicted mutation profile AUROC from Sagittarius and the mean comparison approach for the sarcoma test split including patients with an observed death event more than 37 months after diagnosis. Points above the $y = x$ line indicate that Sagittarius had a better predicted mutation profile than the comparison approach. **f**, Illustration of the ties between the *GLI* oncogene in the HH signalling pathway and the *PTCH1*, *PREX1*, *MYCBP2*, *ARID2* and *DNAH17* genes that Sagittarius predicted as among the most likely mutations in early-stage sarcoma patients.

extrapolate beyond the range of timepoints in the data. Pseudodynamics[89] later extended cell fate modelling to time-resolved steps, enabling extrapolation, but operates in a low-dimensional cell state space. Most recently, PRESCIENT[32] enabled single-cell transcriptomic trajectory modelling, but suffers worse performance in bulk expression modelling. Sagittarius strives to build upon these works by leveraging time series from related yet heterogeneous biological time series for accurate high-dimensional profile extrapolation.

## Methods

The Sagittarius model is divided into encoder and decoder modules around the shared reference space (Supplementary Fig. 1). For a set of $N$ heterogeneous input time gene expression time series $\{x_i \in \mathbb{R}^{T_i \times M}\}_{i=1}^{N}$, each associated with $C \geq 1$ categorical environmental labels $\left\{y_i \in \{1, \cdots, C_j\}_{j=1}^{C}\right\}_{i=1}^{N}$ and $B \geq 1$ continuous variables $\{t_i \in \mathbb{R}^{T_i \times B}\}_{i=1}^{N}$, the encoder first embeds each expression measurement, disentangling the environmental conditions from the embedded representation. The encoder also embeds the timepoint of each measurement. First, we use a high-frequency sine wave that both maps each timepoint to the range $[0, 1]$, mitigating potential out-of-domain challenges with increasingly large scalar inputs, and helps Sagittarius learn high-frequency patterns in the data, which neural networks have been shown to learn more slowly than low-frequency patterns[27,30]. We then learn a full timepoint embedding conditioned on the time series' environmental conditions, effectively learning a representation of the biological age of each measurement. Finally, we map the embedded expression measurements, along with their inferred biological age embeddings, to the reference space using a transformer encoder. This maps each input or source time series $i$ to the common trajectory that the model learns for the heterogeneous dataset (Supplementary Fig. 2).

The Sagittarius decoder proceeds in reverse. First, we use the decoder layers of the transformer and learned representations of biological age in each environmental condition to map from the reference space back to embedded measurements for the queried timepoints and condition of interest. Then we decode the embeddings, conditioned on the queried embeddings, to predict the full output expression vectors for each timepoint (Supplementary Fig. 3). A detailed discussion of the model is presented in Supplementary Note 1.

### Dataset processing

Evo-devo. The Evo-devo dataset[12] contains gene expression vectors for seven species and seven organs measured at multiple pre- and postnatal timepoints. We associated each measurement with the environmental variables $y_i = [\text{species}_i, \text{organ}_i]$. We restricted the dataset to orthologous genes in all seven species, identified with the Python pybiomart package[90] and the provided Ensembl gene identifiers, resulting in 5,037 genes. The measured timepoints for each species were given as strings with different units by species. As a preprocessing step, we computed each timepoint's rank within the ordered timepoints for that species to use as the timepoint label. We then randomly selected the rabbit heart time series and used the augmented Dickey–Fuller test, which tests for stationarity, and only retained genes with $P > 0.05$, resulting in 4,533 genes.

**LINCS.** We used the LINCS L1000 platform level 3 pharmacogenomic dataset[15]. We restricted the data experiments to micromolar dose measurements that did not exceed 20 μM. We then further restricted the dataset to include only drug and cell line treatment combinations that had more than 15 measurements in the restricted dataset, resulting in 2,687 retained combinations that each had between 16 and 78 measurements (Supplementary Fig. 13). We interpreted these as time series in two continuous variables, with $y_i = [\text{drug}_i, \text{cell line}_i]$, $t_i = [\text{dose}_i, \text{time}_i]$ and $x_i[\text{dose } d, \text{time } t] \in \mathbb{R}^{978}$.

**TCGA.** We used the TCGA Firehose legacy dataset[31], independently considering the somatic mutation and RNA-Seq datasets. Both datasets measure 20,501 genes: we restricted the dataset to the 1,000 most frequently mutated genes for the mutation experiments and the 1,000 most highly variable genes for the gene expression experiments, jointly considering all cancer types. We removed all patients with missing event times, and excluded mutation patients with no mutations in the 1,000 remaining genes. Finally, we excluded cancer types with fewer than 12 remaining patients. We finally constructed a time series for each cancer type $y_i$, where the $r$th patient in the time series had event time $t_i[r]$ and mutation or gene expression profile $x_i[r]$. Although this formulation enables extrapolation to cancer patients with good prognosis, censored patients may confound the temporal ordering of patients. For the mutation experiments, we exclude all censored patients whose reported event times probably differ markedly from their survival times, where this probability is acquired with a machine learning model (Supplementary Note 1).

### Quantitative experiments

For each experiment, we define the test set $\widetilde{T} = \{\widetilde{T}_i\}_{i=1}^{N}$, which we use to evaluate model performance (Supplementary Note 1).

**Evo-devo.** For the late developmental stage extrapolation experiment, we defined the test set $\widetilde{T}_i$ to contain the final four measurements from each time series $i$; analogously, we set $\widetilde{T}_i$ to contain the first four measurements from each time series for the early development extrapolation task. In both settings, this resulted in 192 test measurements and 471 measurements available for training or validation. At evaluation time, we used the models to predict the expression vectors at the test timepoints $\widetilde{T}$ and compared the predicted results with the masked measurements in the Evo-devo dataset.

**LINCS.** We defined three main quantitative LINCS experiments: complete generation, combination & dose, and combination & time. For each setting, we first randomly select five drug and cell-line treatment combinations to mask from the training data, requiring that both the drug and the cell line appear at least once somewhere else in the dataset. We always add all measurements from these five treatments as test points $\widetilde{T}$. In complete generation, we also randomly select three non-zero doses and one non-zero treatment time; in combination & dose, we randomly select three non-zero doses; in combination & time we select one non-zero treatment time. In all settings, any measurement at any selected dose or time is also added to the test set $\widetilde{T}$. This resulted in 2,144 training sequences and 269 validation sequences for each of the three tasks. There were 7,651, 27,242 and 10,417 total training measurements and 924, 3,326 and 1,202 validation measurements for the complete generation, combination & dose, and combination & time tasks respectively. Finally, the three tasks had 7,441, 7,377 and 7,395 test sequences with 15,068, 14,905 and 14,966 test measurements respectively.

**TCGA.** For cancer type $i$ with $\Gamma$ observed patients, we defined a training split of the $k$ observed patients with shortest survival time, as well as all censored patients with a shorter event time than the $k$th-longest observed survival time. We then used the remaining $\Gamma - k$ observed patients as the test patients $\widetilde{T}_i$. We further varied $k = 1, 2, \ldots, \Gamma - 1$ to study the effect of different training and test split sizes. In the TCGA gene expression extrapolation experiment, this leads to 13 distinct THCA test splits and 92 sarcoma test splits. In the TCGA mutation extrapolation experiment, this leads to 9 unique THCA test splits and 61 sarcoma test splits.

### Evo-devo gene-level predictions

To evaluate Sagittarius's performance at the level of a single gene, we used the extrapolation to late timepoints experiment and computed the Pearson correlation comparing timepoints for each gene. To identify housekeeping genes[91] (HKGs), we first stratified the Evo-devo dataset by species and computed the s.d. of each gene across all organs and timepoints. We then identified 114 HKGs per species using the one-sided rank-sum test with $P < 0.05$. To test whether HKGs and non-HKGs had different performance, we used Fisher's exact test at a 0.4 correlation threshold (Supplementary Table 1). Excluding Gene Ontology[92] terms with 1,000 or more associated genes, we then ran a gene set enrichment analysis for the gene set with correlation of at least 0.4 for each time series, retaining the Gene Ontology terms with $P < 0.05$ after Bonferroni adjustment (Supplementary Dataset 1).

### Mouse developmental analysis

After training Sagittarius on the complete Evo-devo dataset, we used Sagittarius to simulate 10 gene expression time series for each of the seven mouse organs, using the original mouse organ time series as source sequences for the model. To study early development and young mice, we extrapolated to timepoint indices ranging from −5 to 13 with granularity 0.1, resulting in 180 predicted measurements per organ per time series. To study ageing mice, we extrapolated to timepoint indices ranging from 11 to 25 with granularity 0.1, resulting in 140 predicted measurements per organ per time series. We smoothed the trajectories by computing the moving average with a window size of 10 (averaging over five measurements in each direction).

**Transcriptomic velocity.** We computed the UMAP[37] embeddings of the smoothed trajectories for early development and young mice. We also computed the developmental velocity in the UMAP space as the vector between the predicted measurements at time $t + 0.1$ and time $t$ in the embedded space for all trajectories and all $t$. We then smoothed the velocity vectors by taking the average of the velocity at $t$ and $t - 0.1$. We then took the normalized average (mean) over the smoothed velocities associated with each of the ten simulated trajectories for a given organ. To decrease clutter in the organ development plot, we took the resulting velocity at integer timepoints. We then projected all gene expression measurements to a grid in the UMAP space, and defined the velocity at each point as the weighted average of the 100 velocity vectors nearest to that grid point using the NearestNeighbors module from sklearn.neighbors[93]. Finally, we discarded the 5% of velocities with the smallest magnitude to simplify the plot. We repeated these steps using a principal component analysis embedding space for the analogous principal component analysis visualization (Supplementary Fig. 12).

**Organ development genes.** To identify genes that had a very similar expression at early developmental stages but differing expression levels at later developmental stages, we considered the first and last 25 timepoints from each of the smoothed organ trajectories corresponding to early development and young mice, resulting in a total of 250 timepoints each for the early development and young mouse time ranges across all of the samples. We used the ANOVA test with Bonferroni multiple hypothesis testing correction to compare the gene expression values for each gene across organs in the early development time ranges and again to compare expression of each gene across organs in the young mouse development time ranges.

**Tabula Muris Senis gene evaluation.** To evaluate whether Sagittarius could accurately predict gene expression patterns in ageing mice, we computed the Pearson correlation over time for each of the genes in the ageing mouse extrapolated time series. We also computed the Pearson correlation over time for the mouse organ time-series measurements in the Evo-devo dataset, which end at P63. Finally, we used the tissue data from the single-cell Tabula Muris Senis droplet dataset[16] for heart and aorta, kidney, and liver, which were the three tissues that aligned with the Evo-devo organs. We computed the average expression at each timepoint for each cell type in the tissue data, and then took the Pearson correlation of the average cell-type expression over time. We compare the correlation over time for Sagittarius's extrapolated data and the measured Evo-devo data with the distribution of cell-type correlation in the Tabula Muris Senis dataset.

## Drug dosage similarity network

After training Sagittarius on the complete LINCS dataset we randomly selected 78 distinct doses from the dataset, which ranged from $8.33 \times 10^{-5}$ to 19.9998, and selected a treatment time of 6 h. For each drug and cell-line treatment in the dataset, we then used Sagittarius to predict the drug-induced expression profile for the treatment at each of the 78 doses with a 6 h treatment time, using the actual treatment measurements in the dataset as the source sequence. To remove the strong cell-type-specific signal in the profiles, we subtracted Sagittarius's predicted basal cell line expression from the drug-induced expression vectors.

We took the average over all 78 doses of the differential expression vectors to produce a single 978-dimensional vector for each of the 2,687 treatment combinations. We then computed a similarity matrix $\Sigma$ as

$$\sigma_{i,j} = 1 - \left| x_i^{\Delta} - x_j^{\Delta} \right|,$$

$$\Sigma_{i,j} = \min_{i',j'} \left( \max \left( \frac{\sigma_{i,j} - \sigma_{i',j'}}{\sigma_{i,i'} - \sigma_{i',j'}} \right) \right)$$

where $x_n^{\Delta}$ indicates the differential expression of treatment combination $n$. We constructed an average differential expression $k$-nearest-neighbour network $G_{\mathrm{KNN}}$, beginning from a fully connected graph with edge weights $\Sigma$, by first removing all edges with $\Sigma_{ij} < 0.95$, then removing the degree($i$) − 50 edges with lowest weight for each node $i$, removing all nodes with degree less than 30 and finally reducing the remaining graph to its largest connected subgraph. We then ran Louvain community detection[94] from the Python community package[95] (python-louvain) to identify communities in $G_{\mathrm{KNN}}$. To simplify the analysis, we combined neighbouring communities until four remained, and then calculated the community $IC_{50}$ by averaging the individual treatment $IC_{50}$ over all treatments in the community that were also measured in the GDSC dataset. We visualized $G_{\mathrm{KNN}}$ using the edge-weight spring embedded layout in Cytoscape[96], with minimum, maximum and default edge weights of 0, 1 and 0.5 respectively. We ran 200 average iterations for each node. The spring strength parameter was set to 15, the spring rest length to 45, the disconnected spring strength to 0.05 and the disconnected spring rest length to 2,000. We did not add any spring strength to avoid collisions, and used two layout passes. Finally, we randomized the graph before computing the layout.

## Drug sensitivity prediction

To evaluate Sagittarius's utility for drug sensitivity prediction, we again randomly selected 78 doses in the LINCS dataset and fixed the treatment time at 6 h. We then used Sagittarius's learned weights to compute the transformer encoder's average key representation over the doses for a given drug and cell-line combination, corresponding to the average treatment efficacy relative to the reference space. We fix the treatment time while varying dose to best capture the impact of dose on the treatment response, as $IC_{50}$ is based on drug dose–response curves[52,58,97].

**GDSC dataset.** For the GDSC-based prediction, we computed the key representations for each GDSC[52]-measured combination of drug and cell line, provided that both the drug and cell line appeared somewhere in the LINCS dataset (although not necessarily together). We then constructed one dataset with 271 datapoints using Sagittarius's representations, and another dataset with 151 datapoints that contained the drug-induced gene expression profiles for treatment combinations that also appeared in the GDSC dataset, denoted Sagittarius and LINCS-GDSC respectively. We ran threefold cross-validation on the LINCS-GDSC model, where the test set made up two-thirds of the data in each fold. We divided Sagittarius's simulated dataset such that the test split for each fold matched the LINCS-based model split, and used the remaining data for training.

**CTRP dataset.** For this experiment we computed Sagittarius's key representations for each CTRP[58]-measured experimental combination. The Sagittarius dataset had 2,929 datapoints and the LINCS-CTRP dataset had 625 datapoints.

**Dataset evaluation.** To evaluate the quality of the Sagittarius and LINCS-based $IC_{50}$ prediction models, we computed the average Spearman correlation between the model's $IC_{50}$ predictions and the dataset labels (either from GDSC or CTRP) for each test cell line. To compare overall test performance, we restricted our analysis to cell lines where at least one of the models had significant correlation (Spearman rank-order $P < 0.05$). We used the Spearman correlation between all predicted and measured validation data to quantify validation set performance.

**Model hyperparameters.** We used threefold cross-validation on the LINCS-based dataset, where two-thirds of the data were used as test for each fold. We defined test splits in the Sagittarius dataset to match the LINCS-based test split for each fold. We held out 10% of the remaining training data for both the Sagittarius and LINCS-based datasets to determine the best regression model for the drug sensitivity prediction task in each dataset. We considered both support vector regression and multilayer perceptron (MLP)-based regressors: for support vector regression configurations we considered linear, polynomial and radial basis function kernels; for MLP regression configurations, we considered a regularization weight $\alpha \in [1 \times 10^{-4}, 1 \times 10^{-2}, 1, 10]$. All other hyperparameters maintained the defaults in sklearn[93]. For the LINCS-based dataset model, the best-performing configuration on the validation data used a support vector regressor with a radial basis function kernel on the GDSC dataset and polynomial kernel on the CTRP dataset. For the Sagittarius dataset model, the best-performing configuration used the MLP model with $\alpha = 10$ on the GDSC dataset and $\alpha = 0.01$ on the CTRP dataset.

## Cancer gene essentiality prediction

**Dataset construction.** We used the DEMETER[60] and CERES[61] versions of the DepMap dataset, which quantify gene essentiality via short hairpin RNAs and CRISPR–Cas9 essentiality screens respectively. For each gene and cell line combination in DepMap, we searched for a drug in the LINCS dataset that listed the given gene as its target, hypothesizing that the drug's inhibitory effect on a cell line is related to the cell line's dependence on the target gene[98]. We constructed a dataset from Sagittarius on the basis of the transformer's average key representation as in the drug sensitivity prediction dataset, using 78 random doses and a treatment combination in the given cell line and targeting the gene of interest for each DepMap essentiality pair. This resulted in 4,216 and 1,666 datapoints from Sagittarius for the DEMETER and CERES versions respectively. The analogously constructed LINCS-DepMap dataset used the average drug-induced expression across doses for DepMap pairs that matched available

LINCS experiments, resulting in 765 and 353 datapoints for the two versions.

**Dataset evaluation.** To evaluate the quality of the Sagittarius- and LINCS-based datasets, we computed the Spearman correlation between the gene essentiality scores measured in the dataset and those predicted by the model for each test cell line.

**Model training.** We trained a two-layer MLP regressor with 200 and 100 hidden nodes respectively, ReLU activation functions, MSE loss and the Adam optimizer[99] with a learning rate of $1 \times 10^{-3}$. We used fivefold cross-validation, where 20% of the LINCS-DepMap dataset was used as the test set, and we aligned the Sagittarius dataset's test set to match the LINCS-DepMap test set. We further held out 10% of the resulting training set for each of the five splits to use as a validation set for early model training termination.

### Early cancer patient simulation

**Mutation profile simulation.** To simulate the early-stage sarcoma patient mutation profiles, we trained Sagittarius on all available TCGA mutation data and then predicted mutation probability profiles at 27 survival timepoints, ranging from 203 to 283 months. Specifically, we selected the longest 27 survival times that appeared somewhere in the initial TCGA mutation dataset, with

$$t \in \{\ 203.12,\ 204.01,\ 260.70,\ 208.23,\ 209.43,\ 210.51,\ 210.81,\ 211.01,\ 211.73,\ 212.09,\ 216.59,\ 216.75,\ 225.43,\ 229.04,\ 230.72,\ 232.00,\ 232.62,\ 233.44,\ 234.10,\ 238.11,\ 244.32,\ 244.91,\ 255.49,\ 263.07,\ 275.66,\ 281.08,\ 282.69\ \}\ \text{months}$$

(Supplementary Fig. 23). We then averaged the mutation profile predictions of the 27 timepoints and identified the 10 genes the model predicted as most likely to be mutated.

**Differentially expressed gene simulation.** To investigate the transcriptional differences between early- and late-stage sarcoma patients, we trained Sagittarius on all available TCGA RNA-Seq profiles and then simulated new expression profiles for two synthetic sarcoma patients, one with postdiagnosis survival $t = 0$ and the other with $t = 244.91$, representing late-stage and early-stage tumours respectively. We then identified the $k = 10$ genes that were most overexpressed in the late-stage tumour compared with the early-stage tumour as *PMP2, NDRG1, JUN, SEPT9, PHC2, NCAM1, GFAP, APP, WNK1* and *RAB10*. Previous work identified *JUN*[100], *NCAM1*[101], *NDRG1*[102–105] and *PMP2*[106,107] in aggressive sarcomas with poor prognosis and short time to metastasis. Furthermore, other work has found *RAB10*[108], *SEPT9*[109], *PHC2*[110], *WNK1*[111] and *APP*[112] overexpression in sarcomas.

### Reporting summary

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

## Data availability

The data necessary for reproducing the paper are available in the figshare repository at https://figshare.com/projects/Sagittarius/144771. We provide a more detailed note describing the datasets provided in the figshare repository. Namely, the main experiments can be run with the preprocessed data files provided, and we also include data for the cross-dataset analyses such as the application in drug repurposing. Finally, we provide large-scale simulated data from Sagittarius for the Evo-devo and TCGA datasets.

## Code availability

A Python repository including the Sagittarius implementation and code to reproduce the results in this paper is available at https://github.com/addiewc/Sagittarius[113], with additional details for reproducing results provided in the repository (https://doi.org/10.5281/zenodo.7879454)[114]. We ran experiments on Linux 8.7 with RTX A4000 GPU. We used Python 3.9.7, PyTorch[115] 1.9.1, anndata[116] 0.8.0, cudatoolkit 11.1.74, matplotlib[117] 3.4.3, NumPy[118] 1.21.2, pandas[119] 1.3.3, pip 21.3.1, pybiomart[90] 0.2.0, python-louvain[95] 0.15, Scanpy[120] 1.8.2, SciPy[121] 1.7.1, seaborn[122] 0.11.2, sklearn[93] 0.0, statsmodels[123] 0.13.0, tqdm[124] 4.62.3, umap-learn[37] 0.5.1, yaml 0.2.5, lifelines[125] 0.26.5, BioRender Student Plan and Adobe Illustrator 25.2.3.

## References


1. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
2. Arbeitman, M. N. et al. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275 (2002).
3. Zheng, L. et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
4. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
5. Lee, J. S. et al. Single-cell transcriptome of bronchoalveolar lavage fluid reveals sequential change of macrophages during SARS-CoV-2 infection in ferrets. *Nat. Commun.* **12**, 4567 (2021).
6. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
7. Douglass, E. F. Jr et al. A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. *Cell Rep. Med.* **3**, 100492 (2022).
8. Kohonen, P. et al. A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat. Commun.* **8**, 15932 (2017).
9. Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.29.493900 (2022).
10. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
11. Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
12. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
13. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
14. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020).
15. Subramanian, A. et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
16. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
17. Schaum, N. et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**, 596–602 (2020).
18. Wang, W. et al. Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nat. Med.* **26**, 1644–1653 (2020).
19. Sunkin, S. M. et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).

20. Radovic, A., He, J., Ramanan, J., Brubaker, M. A. & Lehrmann, A. M. Agent forecasting at flexible horizons using ODE flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (2021).
21. Peng, G., Cui, G., Ke, J. & Jing, N. Using single-cell and spatial transcriptomes to understand stem cell lineage specification during early embryo development. *Annu. Rev. Genomics Hum. Genet.* **21**, 163–181 (2020).
22. Haniffa, M. et al. A roadmap for the Human Developmental Cell Atlas. *Nature* **597**, 196–205 (2021).
23. Sohn, K., Lee, H. & Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. in *Advances in Neural Information Processing Systems* (eds. Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R.) vol. 28 3483–3491 (Curran Associates, Inc., 2015).
24. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* e11517 (2023).
25. Cho, K. et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) (eds. Moschitti, A., Pang, B. & Daelemans, W.) 1724–1734 (Association for Computational Linguistics, 2014).
26. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural Ordinary Differential Equations. in *Advances in Neural Information Processing Systems* (eds. Bengio, S. et al.) vol. 31 6571–6583 (Curran Associates, Inc., 2018).
27. Shukla, S. N. & Marlin, B. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations* (ICLR, 2021).
28. Chen, R. T. Q., Amos, B. & Nickel, M. Learning neural event functions for ordinary differential equations. *International Conference on Learning Representations* (ICLR, 2021).
29. Vaswani, A. et al. Attention is All you Need. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 5998–6008 (Curran Associates, Inc., 2017).
30. Rahaman, N. et al. On the spectral bias of neural networks. *Proc. Mach. Learning Res.* **97**, 5301–5310 (2019).
31. Cancer Genome Atlas Research Networket al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
32. Yeo, G. H. T., Saksena, S. D. & Gifford, D. K. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat. Commun.* **12**, 3222 (2021).
33. Tam, P. P. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.* **68**, 3–25 (1997).
34. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
35. Qiu, C. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54**, 328–341 (2022).
36. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
37. McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3**, 861 (2018).
38. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
39. Viegas, J. O. et al. RNA degradation eliminates developmental transcripts during murine embryonic stem cell differentiation via CAPRIN1-XRN2. *Dev. Cell* **57**, 2731–2744.e5 (2022).
40. Tomecki, R., Sikorski, P. J. & Zakrzewska-Placzek, M. Comparison of preribosomal RNA processing pathways in yeast, plant and human cells—focus on coordinated action of endo- and exoribonucleases. *FEBS Lett.* **591**, 1801–1850 (2017).
41. Watada, E. et al. Age-dependent ribosomal DNA variations in mice. *Mol. Cell. Biol.* **40**, e00368-20 (2020).
42. Nimura, K. et al. Regulation of alternative polyadenylation by Nkx2-5 and Xrn2 during mouse heart development. *eLife* **5**, e16030 (2016).
43. Chatterjee, S. & Grosshans, H. Active turnover modulates mature microRNA activity in *Caenorhabditis elegans*. *Nature* **461**, 546–549 (2009).
44. Chatterjee, S., Fasler, M., Büssing, I. & Grosshans, H. Target-mediated protection of endogenous microRNAs in *C. elegans*. *Dev. Cell* **20**, 388–396 (2011).
45. Chowdhury, T., Samajdar, A., Sardar, M. & Chatterjee, S. Dauer quiescence as well as continuity of the life cycle after dauer-exit in *Caenorhabditis elegans* are dependent on the endoribonuclease activity of XRN-2. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.02.489690 (2022).
46. Kato, M., de Lencastre, A., Pincus, Z. & Slack, F. J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol.* **10**, R54 (2009).
47. Qiao, G.-J., Chen, L., Wu, J.-C. & Li, Z.-R. Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. *PeerJ* **7**, e6548 (2019).
48. Takada, H. & Kurisaki, A. Emerging roles of nucleolar and ribosomal proteins in cancer, development, and aging. *Cell. Mol. Life Sci.* **72**, 4015–4025 (2015).
49. Loganathan, T., Ramachandran, S., Shankaran, P., Nagarajan, D. & Mohan S, S. Host transcriptome-guided drug repurposing for COVID-19 treatment: a meta-analysis based approach. *PeerJ* **8**, e9357 (2020).
50. Belyaeva, A. et al. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. *Nat. Commun.* **12**, 1024 (2021).
51. Minamiyama, M. et al. Naratriptan mitigates CGRP1-associated motor neuron degeneration caused by an expanded polyglutamine repeat tract. *Nat. Med.* **18**, 1531–1538 (2012).
52. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
53. Yang, C. et al. A survey of optimal strategy for signature-based drug repositioning and an application to liver cancer. *eLife* **11**, e71880 (2022).
54. Cheng, X. et al. Drug repurposing for cancer treatment through global propagation with a greedy algorithm in a multilayer network. *Cancer Biol. Med.* **19**, 74–89 (2022).
55. Folkes, A. J. et al. The identification of 2-(1H-indazol-4-yl)-6-(4-methanesulfonyl-piperazin-1-ylmethyl)-4-morpholin-4-yl-thieno[3,2-d]pyrimidine (GDC-0941) as a potent, selective, orally bioavailable inhibitor of class I PI3 kinase for the treatment of cancer. *J. Med. Chem.* **51**, 5522–5532 (2008).
56. Roth, G. J. et al. Nintedanib: from discovery to the clinic. *J. Med. Chem.* **58**, 1053–1063 (2015).
57. Suzuki, N., Nakagawa, F., Matsuoka, K. & Takechi, T. Effect of a novel oral chemotherapeutic agent containing a combination of trifluridine, tipiracil and the novel triple angiokinase inhibitor nintedanib, on human colorectal cancer xenografts. *Oncol. Rep.* **36**, 3123–3130 (2016).
58. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).

59. Menden, M. P. et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**, 2674 (2019).
60. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
61. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
62. Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **11**, 3475 (2020).
63. Bozic, I. et al. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107**, 18545–18550 (2010).
64. Arazo, E., Ortego, D., Paul, A., O'Connor, N. E. & McGuinness, K. Unsupervised label noise modeling and loss correction. *Proc. Mach. Learning Res.* **97**, 312–321 (2019).
65. Li, J., Socher, R. & Hoi, S. C. H. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations* (2020).
66. Brown, L. C. et al. LRP1B mutations are associated with favorable outcomes to immune checkpoint inhibitors across multiple cancer types. *J. Immunother. Cancer* **9**, e001792 (2021).
67. Arang, N. & Gutkind, J. S. G protein-coupled receptors and heterotrimeric G proteins as cancer drivers. *FEBS Lett.* **594**, 4201–4232 (2020).
68. Ichikawa, D. et al. Integrated diagnosis based on transcriptome analysis in suspected pediatric sarcomas. *NPJ Genom. Med.* **6**, 49 (2021).
69. Pietrobono, S., Gagliardi, S. & Stecca, B. Non-canonical Hedgehog signaling pathway in cancer: activation of GLI transcription factors beyond Smoothened. *Front. Genet.* **10**, 556 (2019).
70. Lo, W. W., Pinnaduwage, D., Gokgoz, N., Wunder, J. S. & Andrulis, I. L. Aberrant hedgehog signaling and clinical outcome in osteosarcoma. *Sarcoma* **2014**, 261804 (2014).
71. Banerjee, S. et al. Loss of the PTCH1 tumor suppressor defines a new subset of plexiform fibromyxoma. *J. Transl. Med.* **17**, 246 (2019).
72. Martinez, M. F. et al. Nevoid basal cell carcinoma syndrome: PTCH1 mutation profile and expression of genes involved in the Hedgehog pathway in Argentinian patients. *Cells* **8**, 144 (2019).
73. Ge, Z. et al. Clinical significance of high c-MYC and low MYCBP2 expression and their association with Ikaros dysfunction in adult acute lymphoblastic leukemia. *Oncotarget* **6**, 42300–42311 (2015).
74. Vatapalli, R. et al. Histone methyltransferase DOT1L coordinates AR and MYC stability in prostate cancer. *Nat. Commun.* **11**, 4153 (2020).
75. Yoon, J. W. et al. Noncanonical regulation of the Hedgehog mediator GLI1 by c-MYC in Burkitt lymphoma. *Mol. Cancer Res.* **11**, 604–615 (2013).
76. Tazzari, M. et al. Molecular determinants of soft tissue sarcoma immunity: targets for immune intervention. *Int. J. Mol. Sci.* **22**, 7518 (2021).
77. Wang, X., Haswell, J. R. & Roberts, C. W. M. Molecular pathways: SWI/SNF (BAF) complexes are frequently mutated in cancer—mechanisms and potential therapeutic insights. *Clin. Cancer Res.* **20**, 21–27 (2014).
78. Fan, X. et al. The association between methylation patterns of DNAH17 and clinicopathological factors in hepatocellular carcinoma. *Cancer Med.* **8**, 337–350 (2019).
79. Hassounah, N. B., Bunch, T. A. & McDermott, K. M. Molecular pathways: the role of primary cilia in cancer progression and therapeutics with a focus on Hedgehog signaling. *Clin. Cancer Res.* **18**, 2429–2435 (2012).
80. Stecca, B. & Ruiz i Altaba, A. Context-dependent regulation of the GLI code in cancer by HEDGEHOG and non-HEDGEHOG signals. *J. Mol. Cell. Biol.* **2**, 84–95 (2010).
81. Brechbiel, J., Miller-Moslin, K. & Adjei, A. A. Crosstalk between hedgehog and other signaling pathways as a basis for combination therapies in cancer. *Cancer Treat. Rev.* **40**, 750–759 (2014).
82. Chen, J., Zhang, J., Hong, L. & Zhou, Y. EGFLAM correlates with cell proliferation, migration, invasion and poor prognosis in glioblastoma. *Cancer Biomark.* **24**, 343–350 (2019).
83. Yu, Q. et al. Upregulated NLGN1 predicts poor survival in colorectal cancer. *BMC Cancer* **21**, 884 (2021).
84. Ren, Y.-M. et al. Exploring the key genes and pathways of side population cells in human osteosarcoma using gene expression array analysis. *J. Orthop. Surg. Res.* **13**, 153 (2018).
85. Cutcliffe, C. et al. Clear cell sarcoma of the kidney: up-regulation of neural markers with activation of the Sonic hedgehog and Akt pathways. *Clin. Cancer Res.* **11**, 7986–7994 (2005).
86. Wald, Y., Feder, A., Greenfeld, D. & Shalit, U. On Calibration and Out-of-Domain Generalization. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 2215–2227 (Curran Associates, Inc., 2021).
87. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
88. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **19**, 477 (2018).
89. Fischer, D. S. et al. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461–468 (2019).
90. de Ruiter, J. pybiomart: a simple Pythonic interface to BioMart. *GitHub* https://github.com/jrderuiter/pybiomart (2018).
91. Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J. & Lewis, N. E. What are housekeeping genes? *PLoS Comput. Biol.* **18**, e1010295 (2022).
92. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
93. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
94. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
95. Aynaud, T. python-louvain 0.15: Louvain algorithm for community detection. *GitHub* https://github.com/taynaud/python-louvain (2020).
96. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
97. Berry, L. M. & Zhao, Z. An examination of IC50 and IC50-shift experiments in assessing time-dependent inhibition of CYP3A4, CYP2D6 and CYP2C9 in human liver microsomes. *Drug Metab. Lett.* **2**, 51–59 (2008).
98. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
99. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. in *3rd International Conference on Learning Representations*, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (eds. Bengio, Y. & LeCun, Y.) (2015).
100. Mariani, O. et al. JUN oncogene amplification and overexpression block adipocytic differentiation in highly aggressive sarcomas. *Cancer Cell* **11**, 361–374 (2007).

101. Bae, J. Y. et al. Evaluation of immune-biomarker expression in high-grade soft-tissue sarcoma: HLA-DQA1 expression as a prognostic marker. *Exp. Ther. Med.* **20**, 107 (2020).
102. Wang, H. et al. HER4 promotes cell survival and chemoresistance in osteosarcoma via interaction with NDRG1. *Biochim. Biophys. Acta Mol. Basis Dis.* **1864**, 1839–1849 (2018).
103. Yan, X., Chua, M.-S., Sun, H. & So, S. N-Myc down-regulated gene 1 mediates proliferation, invasion, and apoptosis of hepatocellular carcinoma cells. *Cancer Lett.* **262**, 133–142 (2008).
104. Cheng, J. et al. NDRG1 as a biomarker for metastasis, recurrence and of poor prognosis in hepatocellular carcinoma. *Cancer Lett.* **310**, 35–45 (2011).
105. Hua, Y. et al. Plasma membrane proteomic analysis of human osteosarcoma and osteoblastic cells: revealing NDRG1 as a marker for osteosarcoma. *Tumour Biol.* **32**, 1013–1021 (2011).
106. Graf, S. A. et al. The myelin protein PMP2 is regulated by SOX10 and drives melanoma cell invasion. *Pigment Cell Melanoma Res.* **32**, 424–434 (2019).
107. Cheng, L. et al. Integration of genomic copy number variations and chemotherapy-response biomarkers in pediatric sarcoma. *BMC Med. Genom.* **12**, 23 (2019).
108. Guo, Q., Sun, H., Zheng, K., Yin, S. & Niu, J. Long non-coding RNA DLX6-AS1/miR-141-3p axis regulates osteosarcoma proliferation, migration and invasion through regulating Rab10. *RSC Adv.* **9**, 33823–33833 (2019).
109. International Cancer Genome Consortiumet al. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
110. Mito, J. K. et al. Cross species genomic analysis identifies a mouse model as undifferentiated pleomorphic sarcoma/malignant fibrous histiocytoma. *PLoS ONE* **4**, e8075 (2009).
111. Capra, M. et al. Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer Res.* **66**, 8147–8154 (2006).
112. Pandey, P. et al. Amyloid precursor protein and amyloid precursor-like protein 2 in cancer. *Oncotarget* **7**, 19430–19444 (2016).
113. Woicik, A. addiewc/Sagittarius: Sagittarius. *Zenodo* https://doi.org/10.5281/zenodo.7879454 (2023).
114. Woicik, A. Simulated EvoDevo dataset. *figshare* https://doi.org/10.6084/m9.figshare.20425572 (2022).
115. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 8026–8037 (Curran Associates, Inc., 2019).
116. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: annotated data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.12.16.473007 (2021).
117. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
118. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
119. The Pandas Development Team. pandas-dev/pandas: pandas. *Zenodo* https://doi.org/10.5281/zenodo.7857418 (2023).
120. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
121. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
122. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
123. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92–96 https://doi.org/10.25080/majora-92bf1922-011 (*SciPy*, 2010).
124. da Costa-Luis, C. et al. tqdm: a fast, extensible progress bar for Python and CLI. *Zenodo* https://doi.org/10.5281/zenodo.7697295 (2023).
125. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).


## Acknowledgements

S.W. is supported by the Sony Research Award. Figures 1, 4a,b and 6a,f were created with BioRender.com.

## Author contributions

A.W. and S.W. conceptualized the work and designed the method. A.W., S.W. and J.M. designed the experiments. A.W. and M.Z. ran the experiments, and A.W., M.Z. and J.C. developed computational tools for Sagittarius. A.W. and S.W. wrote the manuscript and designed the figures.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00679-5.

**Correspondence and requests for materials** should be addressed to Sheng Wang.

**Peer review information** *Nature Machine Intelligence* thanks Chenling Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Sheng Wang

Last updated by author(s): Apr 30, 2023

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used for data collection. |
| Data analysis | We used Python 3.9.7, pytorch 1.9.1, anndata 0.8.0, cudatoolkit 11.1.74, matplotlib 3.4.3, numpy 1.21.2, pandas 1.3.3, pip 21.3.1, pybiomart 0.2.0, python-louvain 0.15, scanpy 1.8.2, scipy 1.7.1, seaborn 0.11.2, sklearn 0.0, statsmodels 0.13.0, tqdm 4.62.3, umap-learn 0.5.1, yaml 0.2.5, lifelines 0.26.5, BioRender Student Plan, and Adobe Illustrator 25.2.3. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data necessary for reproducing the paper are available in the figshare repository at https://figshare.com/projects/Sagittarius/144771. We provide a more detailed note describing the datasets provided in the figshare repository. Namely, the main experiments can be run with the pre-processed data files provided, and

we also include data for the cross-dataset analyses such as the application in drug repurposing. Finally, we provide large-scale simulated data from Sagittarius for the Evo-devo and TCGA datasets.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | *Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.*<br>*Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.*<br>*Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.* |
| Reporting on race, ethnicity, or other socially relevant groupings | *Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).*<br>*Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)*<br>*Please provide details about how you controlled for confounding variables in your analyses.* |
| Population characteristics | *Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."* |
| Recruitment | *Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.* |
| Ethics oversight | *Identify the organization(s) that approved the study protocol.* |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We used the available samples in the Evo-devo, LINCS, and TCGA datasets. We determined test splits that had an apparently sufficient size to be robust to different random seeds. |
| Data exclusions | We excluded LINCS perturbation experiments with fewer than 15 dose and time measurements. This reduced the size of the dataset to speed model training and evaluation. We also excluded some TCGA patients with censored survival times so as to maintain a reasonable patient ordering for the time series evaluation. |
| Replication | We set random seeds for data loading and model initialization in order to reproduce the experimental findings. Repeated experiments were able to replicate the computational results when using the same hardware (Linux 8.7; RTX A4000 GPU). |
| Randomization | For the Evo-devo experiments, we divided the test split based on the final four measured timepoints in each time series for the late extrapolation task (and the first four timepoints for the early extrapolation task). We then divided train and validation splits randomly from the remaining data, with 20% of the available measurements used for validation.<br><br>For the LINCS experiments, we randomly selected 5 treatment combinations of a drug and cell line, and used all measurements from these treatments as test data. We also randomly selected 3 non-zero doses from and additionally treated all measurements at one of these doses as test data for the combination & dose task; we randomly selected one non-zero treatment time and additionally treated all measurements at this treatment time as test data for the combination & time task; we randomly selected 3 non-zero doses and one non-zero treatment time and additionally treated all measurements at one of these doses or this treatment time as test data for the complete generation task.<br><br>For the TCGA experiment, we used SARC and THCA as our cancer types for case study. We used 20% of the data from the remaining cancer types as validation data. For the studied cancer type (either SARC or THCA in each of the two studies), we defined a deterministic train/test split where we treated the first k patients profiles (k patients with the shortest survival times) as train data and the remaining N-k patient profiles as test data, with N the total number of patients for the cancer type. We then varied k from k=1 to k=N-1 for a set of experiments from that study. |

| | |
|---|---|
| Blinding | Blinding was irrelevant to this study, as the train and test splits for each experiment were either determined by the type of analysis (e.g., extrapolation for Evo-devo and TCGA experiments) and all were fixed after random assignment. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |