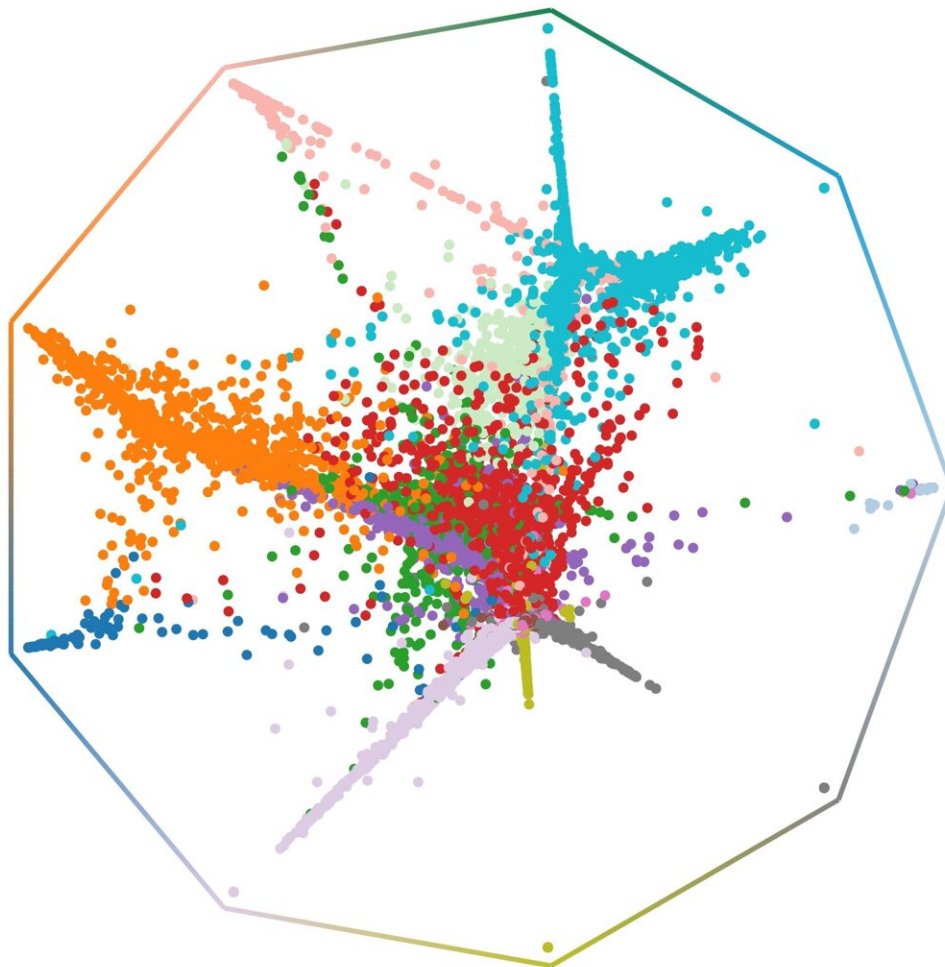# Predicting cell fates: Researchers develop AI solutions for next-gen biomedical research

February 2 2022

CellRank's fate probabilities for lung regeneration; each cell is placed at a position that reflects its probability of reaching any terminal state. Credit: Helmholtz Munich / Marius Lange

Data is not only the answer to numerous questions in the business world; the same applies to biomedical research. In order to develop new therapies or prevention strategies for diseases, scientists need more and better data, faster and faster. However, the quality is often very variable and the integration of different data sets often almost impossible. With the Computational Health Center at Helmholtz Munich, one of Europe's largest research centers for artificial intelligence in medical science is now being established under the direction of Fabian Theis. In close cooperation with the Technical University of Munich (TUM), more than one hundred scientists are using artificial intelligence and machine learning to discover solutions to precisely these problems, thus enabling medical innovations for a healthier society. In the latest issue of the journal *Nature Methods*, they present three articles with groundbreaking new solutions.

Fabian Theis, Head of the Computational Health Center at Helmholtz Munich and Professor for Mathematical Modelling of Biological Systems at TUM says that "it's been a crazy 4 weeks, with many of our scientific stories and methods coming to fruition in that same time window. Our research groups focuses on using single cell genomics to understand the origin of disease in a mechanistic fashion—for this we leverage and develop machine learning approaches to better represent this complex data. In the three new paper, we worked on single cell data integration, trajectory learning and spatial resolution, respectively. Besides the applications shown in the papers, we expect to support the

next generation of single-cell research towards disease understanding."

Here are the latest solutions developed by Helmholtz Munich and TUM researchers:

Solving the data integration challenge

To see whether an observation one makes in a single dataset can be generalized, you can check whether the same can be observed in other datasets of the same system. In single-cell data, so-called batch effects complicate combining datasets in this manner. These are differences in the molecular profiles between samples as they were generated at a different time, in a different place, or from a different person. Overcoming these effects is a central challenge in single-cell genomics with more than 50 proposed solutions. But which one is the best? A group of researchers around Malte Lücken carefully curated 86 datasets and compared 16 of the most popular data integration methods on 13 tasks. After over 55,000 hours of computation time and a detailed evaluation of 590 results, they built a guide for optimized data integration. This allows for improved observations on disease processes across datasets at a population scale.

Predicting cell states with open-source software

Many questions in biology revolve around continuous processes like development or regeneration. For any cell in such a process, single-cell RNA-sequencing measures gene expression. The method, however, is destructive to cells and scientists obtain only static snapshots. Thus, many algorithms have been developed to reconstruct continuous processes from snapshots of gene expression. A common limitation: These algorithms cannot tell us anything about the direction of the process. To overcome this limitation, Marius Lange and colleagues

developed a new algorithm called CellRank. It estimates directed cell-state trajectories by combining previous reconstruction approaches with RNA velocity, a concept to estimate gene up- or down-regulation. Across in-vitro and in-vivo applications, CellRank correctly inferred fate outcomes and recovered previously known genes. In a lung regeneration example, CellRank predicted novel intermediate cell states on a dedifferentiation trajectory whose existence was validated experimentally. CellRank is an open-source software package that is already used by biologists and bioinformaticians around the world to analyze complex cellular dynamics in situations like cancer, reprogramming or regeneration.

Visualizing spatial omics analysis

Recent years have seen a growing development of technologies to measure gene expression variation in tissue. The advantage of such technologies is that scientists can see cells in their context, thus being able to investigate principles of tissue organization and cellular communication. Researchers need flexible computational frameworks in order to store, integrate and visualize the growing diversity of such data. To tackle this challenge, Giovanni Palla, Hannah Spitzer, and colleagues developed a new computational framework, called Squidpy. It enables analysts and developers to handle spatial gene expression data. Squidpy integrates tools for gene expression and image analysis to efficiently manipulate and interactively visualize spatial omics data. Squidpy is extensible and can be interfaced with a variety of machine learning tools in the python ecosystem. Scientists around the world are already using it to analyze spatial molecular data.

Marius Lange et al, CellRank for directed single-cell fate mapping, *Nature Methods* (2022). DOI: 10.1038/s41592-021-01346-6

Giovanni Palla et al, Squidpy: a scalable framework for spatial omics analysis, *Nature Methods* (2022). DOI: 10.1038/s41592-021-01358-2