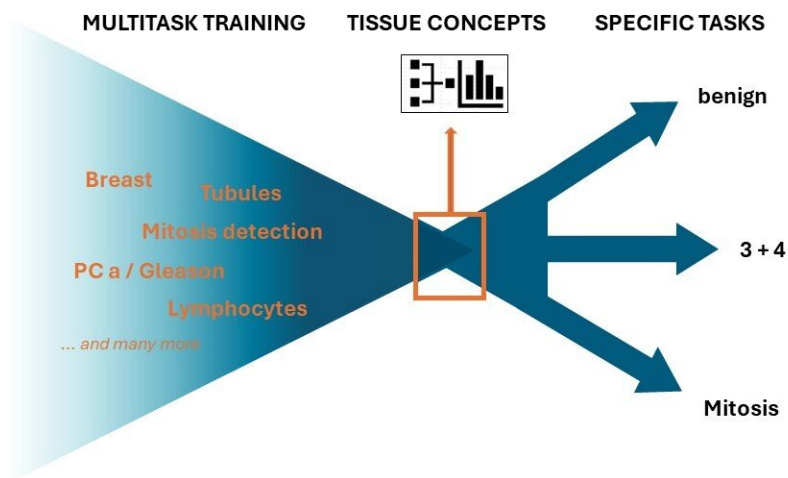# A data-efficient foundation model for biomarker detection

September 2 2024, by Stefanie Lavik

The tissue concepts foundation model developed by the Fraunhofer MEVIS experts: The foundation model is simultaneously pre-trained for various tasks (multitasking). Specific applications come later. Credit: Fraunhofer MEVIS

The use of artificial intelligence (AI) systems shows promise in medicine, where they can be used to detect diseases earlier, improve treatments, and ease staff workloads. But their performance depends on how well the AI is trained.

A new multi-task approach to training AI makes it possible to train foundation models quicker and more cost-effectively, with less data. Researchers are turning to this approach to compensate for the shortage of data in medical imaging—and ultimately save lives.

According to the World Health Organization (WHO), there has been a significant increase in cases of cancer worldwide. Clear indicators, known as biomarkers, are key to reliable diagnosis and successful treatment. AI systems can help identify these kinds of measurable parameters in pathological images.

Researchers from the Fraunhofer Institute for Digital Medicine MEVIS teamed up with RWTH Aachen University, the University of Regensburg, and Hannover Medical School to develop a foundation model for this. The resource-efficient model analyzes tissue samples quickly and reliably, based on just a fraction of the usual training data.

Moving away from large volumes of data and self-supervised learning

Standard foundation models, like the large language models used for ChatGPT, are trained using large and diverse data sets, supervising themselves as they learn. But for medical image analysis, data is generally scarce, and in fact, the small amounts of data available in clinical studies pose a major challenge for the use of AI.

In addition, clinical centers differ in how they process pathological preparations and in their patient populations—even before the specific form and characteristics of diseases are considered.

All of these factors make it harder to reliably detect existing patterns, and thus diagnostically relevant characteristics. To train AI effectively,

this means large volumes of training images from different origins are typically needed. But each cross-sectional image of tissue is typically several gigabytes in size, containing thousands of different cells but only reflecting a tiny fraction of the variability present.

Specialization follows solid foundational training

Fraunhofer MEVIS has devised a solution based on supervised pre-training. "We're developing a training strategy for foundational AI modeled on the training that pathologists undergo. They don't have to relearn what a nucleus is all over again in each case. That's textbook knowledge. Once these concepts have been covered, they're present as a foundation and can be applied to various diseases," explains Dr. Johannes Lotz, an expert from Fraunhofer MEVIS.

In much the same way, their AI model undergoes foundational training, learning general characteristics and laws known as tissue concepts from a broad collection of tissue section images created with various tasks. Combining these tasks gives rise to the large volumes of data needed to train a robust large AI model.

The learned tissue concepts are then applied to a specific task in a second step. In this way, the algorithms can identify biomarkers distinguishing different types of tumors, for example—all with much less data.

"In our solution, every data set has been annotated by a specially trained human with the information that needs to be learned," explains Jan Raphael Schäfer, an AI expert at Fraunhofer MEVIS who works in Lotz's team. "We give our model the image and provide the answer at the same time. And we do it for numerous different tasks simultaneously, using a multi-task approach."

The team also uses an image registration method developed at the institute: HistokatFusion. This method makes it possible to generate automatically annotated training data from tissue studies such as immunohistochemical staining, thereby using marked antibodies to visualize proteins or other structures.

To do this, this method combines information from multiple histopathological images. The experts incorporate these automatically generated annotations into the training of their model, which accelerates data collection.

Outstanding results with just 6% of the resources

Compared to models that do not involve supervised training, the Fraunhofer researchers' approach achieves similar results with only 6% of the training data. "Since the amount of training data in deep learning correlates with training effort and processing power, we found that we needed about 6% of the resources typically required.

"Furthermore, we only need about 160 hours of training, which is a crucial cost factor. This means we can train an equivalent model with much less effort," Lotz explains.

The Fraunhofer experts' participation in the international SemiCOL (Semi-supervised learning for colorectal cancer detection) competition for cancer classification and segmentation showed how well these pre-trained models can be generalized.

The team won the classification part of the challenge without having to undertake expensive adjustments to their model and ultimately came in second out of nine participating teams.

Tests of interactive image segmentation, in which tissue structures are

automatically detected and measured in an image, also show that this method has great potential. The model needs only a few sample image sections to extend concepts that it has already learned.

But that isn't all. "Models based on our solution make it possible to develop new interactive medical AI training tools that let specialists interact directly with AI solutions and train relevant models quickly, even without any technical background knowledge," says Schäfer.

Freely accessible and transferable

The researchers publish the pre-trained model and the code for further learning on various platforms. This lets specialists use it for non-commercial purposes, developing their own solutions. The team is also working with clinical partners to have the solution approved for medical applications and to systematically validate it.

The experts at Fraunhofer MEVIS are certain that once in day-to-day clinical practice, systems involving their foundation model will reduce workloads in pathology and improve the success of treatment.

Provided by Fraunhofer-Gesellschaft