PROSCIA®

Concentriq® Embeddings Accelerated Biomarker Prediction AI Development by 13x

How Proscia's AI R&D Team Leveraged Foundation Models at Scale to Build 80 Breast Cancer Biomarker Prediction Models in Under 24 Hours

Study Authors
Corey Chivers, Ph.D., Vaughn Spurrier, Ph.D.,
Kyriakos Toulgaridis, Jeff Baatz, Julianna Ianni, Ph.D.

Executive Summary

The field of digital pathology is increasingly realizing the power of foundation models to enable faster AI breakthroughs in pathology research and precision medicine discovery and development, with applications ranging from cancer detection to biomarker identification and prognostic analysis.[1] Despite the fast-paced development of new foundation models for vision applications, significant operational challenges still hinder their use for downstream AI development. This process remains time-consuming and operationally intensive with manual data transfers, external processing, and burdensome pipeline maintenance. As a result, the incorporation of foundation models for therapeutic R&D has been delayed, leaving considerable potential unrealized.

Concentriq® Embeddings offers a solution with a seamless way for data scientists to generate embeddings from whole slide images (WSIs) using a curated collection of leading vision and vision-language foundation models directly connected to the Concentriq pathology platform where their data is stored, enriched, and managed. This offering enables more efficient and scalable development of AI models that can be used across the entire R&D lifecycle, accelerating the advancement of precision therapies and diagnostics.

In this study, we demonstrate the efficiency of Concentriq Embeddings to rapidly prototype biomarker prediction models. Concentriq Embeddings dramatically accelerated prototyping times—achieving results 13 times faster than traditional on-premise hardware with a modest WSI dataset size. Specifically, Concentriq Embeddings processed WSIs from the IMPRESS dataset in 2.5 hours, compared to the 33.4 hours needed by a high-end Linux workstation. **This efficiency gain is projected to scale up to 100x for larger datasets that are more typical in core discovery and development activities.** Since Concentriq Embeddings scales effortlessly in the cloud, no additional time is needed to do the same prototyping with not one, but several foundation models. This allowed us to train 80 models in under 24 hours using only a consumer-grade laptop.

Life sciences organizations are adopting AI-powered, data-driven strategies, ushering in a new era of drug development where every stage of R&D becomes more intelligent and efficient. Concentriq Embeddings helps enable this transformation by reducing data processing times and enhancing scalability so that data scientists can advance AI-enabled precision medicine with unprecedented speed and efficiency.

Introduction

AI and deep learning are revolutionizing precision medicine, and computational pathology is emerging as one of the most promising areas for unlocking value. Life sciences organizations are building world-class data science teams to create innovative AI models that accelerate and enhance critical R&D capabilities—ranging from novel biomarker discovery[2] to optimized clinical trial design,[3] and enhanced trial execution with support for endpoint assessments by understanding how exposure to treatments affected tissues and cells.[4]

Foundation models trained on large image datasets have the potential to further accelerate this shift by generating embeddings—numerical representations of an image's essential features—that form the basis for downstream AI model development. New foundation models are being applied to digital pathology with the promise of reducing the long development cycles required to build novel AI models. **However, operationalizing foundation models in practice and at scale remains challenging, leaving considerable potential for AI-driven discovery and development unrealized.**

Traditional methods for leveraging foundation models involve complex processes, including foundation model deployment, extensive data processing, hands-on pipeline maintenance cycles, and cumbersome model prototyping. These challenges, combined with significant computational resource requirements and long processing times, hinder the pace of innovation despite the promise of acceleration offered by this technology.

Concentriq Embeddings solves these challenges by enabling data scientists to easily leverage foundation models for downstream AI model development. With this solution, they can efficiently generate WSI embeddings using a collection of leading vision and vision-language foundation models directly connected to the enterprise pathology platform where their data resides. Data scientists or AI/ML engineers simply assemble a WSI dataset in Concentriq and submit an API post to request embeddings from a chosen foundation model, specifying the resolution, and optionally, the region of interest. Tile embeddings are quickly generated and returned to the user, ready for downstream AI model development (Figure 1).

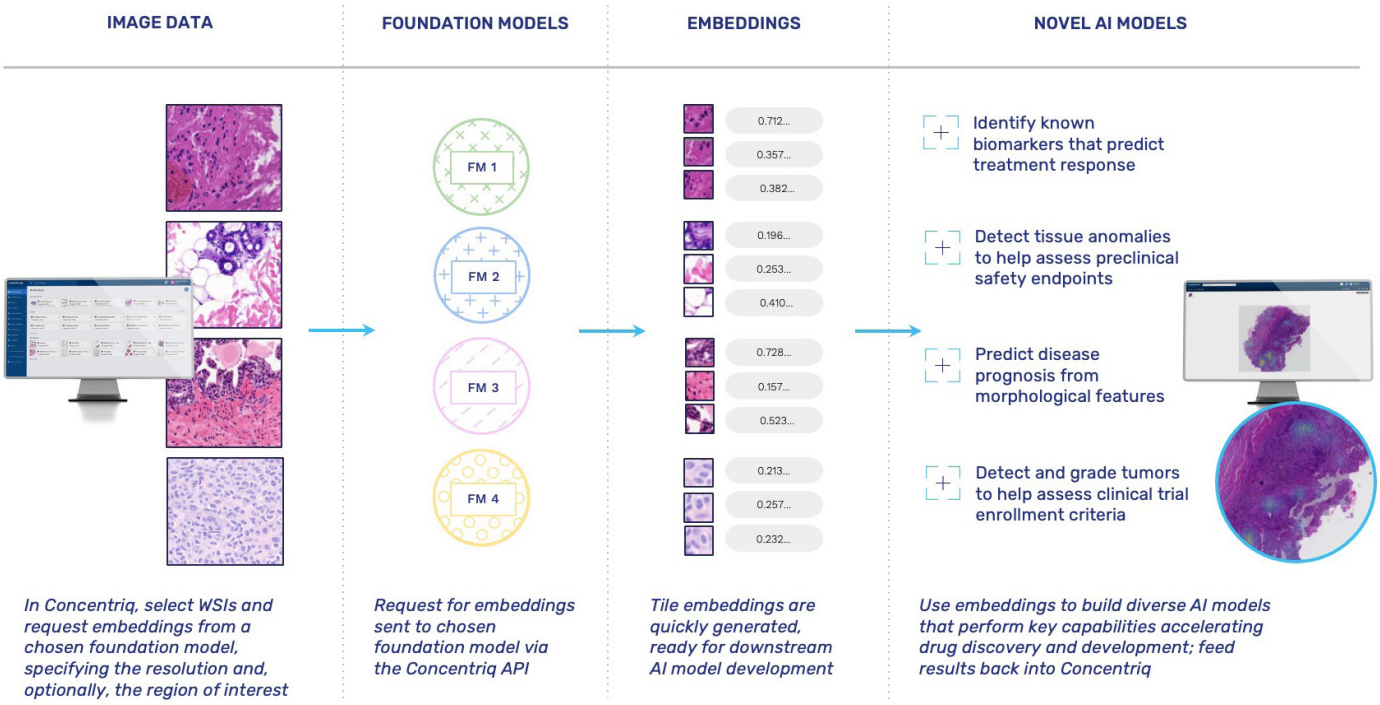

Figure 1: Workflow to generate WSI embeddings from leading vision and vision-language foundation models using Concentriq Embeddings.

By eliminating the need for extensive development efforts to set up, maintain, and optimize expensive data pipelines and infrastructure, data scientists can rapidly refine and iterate on model prototypes. This includes easily testing and swapping foundation models to identify the best fit for specific datasets and downstream applications. Ultimately, Concentriq Embeddings empowers life sciences organizations to fully harness their pathology data, accelerating AI-enabled precision medicine breakthroughs while cutting computational costs.

This study explores the efficiency and effectiveness of Concentriq Embeddings in producing high-quality embeddings for digital pathology, enabling faster prototyping and experimentation. We present a detailed comparison with conventional AI development systems to show how Concentriq Embeddings enabled our team to prototype multiple breast cancer biomarker prediction models, demonstrating its potential to drive impactful discoveries and advancements in precision medicine.

Study Methodology

To assess the performance of Concentriq Embeddings, we compared its speed against a traditional AI development system to prototype several breast cancer biomarker prediction models using the publicly available IMPRESS[5] dataset. This dataset consists of 126 H&E and 126 IHC slides from breast cancer biopsies, including 62 HER2-positive cases and 64 triple-negative breast cancer (TNBC) cases, with a total memory footprint of 49 GB. For this experiment we performed prototyping with DINOv2[6] embeddings.

We used two development approaches for this comparison:

1. **Manual Approach**—Utilizing conventional practices on a Linux workstation equipped with an NVIDIA V100 Tensor Core GPU with 32 GB of RAM. This system, housed in a data center in Philadelphia, PA, incurred costs for both acquisition (approximately $6,000) and monthly housing charges (approximately $600 per month).
2. **Automated Approach**—Utilizing Concentriq Embeddings.

Results and Comparative Analysis

Acquiring Embeddings Using a Manual Approach (Linux Workstation)

The Manual Approach first required downloading the IMPRESS WSIs, which, using 20 CPU worker processes, took 671 seconds, resulting in a download throughput of 49 GB in 671 seconds, equivalent to 584 Mbps. Once the slides were downloaded, OpenSlide-Python was used to cache 20X magnification thumbnail images of each slide again using 20 CPU parallel threads, a process that took 669 minutes. Starting from thumbnail images saved to disk, tiling and inference with DINOv2 required 725 minutes.

The total processing time, from the start of the data download to obtaining embeddings, was 1,405 minutes, or 23.4 hours. Local caching of 20X thumbnails increased the data size on disk from 49 GB for the slides to 203 GB when including both slides and thumbnails.

All times listed reflect processing durations and do not account for any code development required to execute these tasks. The active development time for generating embeddings on the Linux workstation was about 10 hours, bringing the total time spent producing embeddings using the Manual Approach to 2,005 minutes. This number assumes that the necessary code for this project was already available; developing from scratch or using a less optimized existing codebase could significantly increase the time investment for this task.

Efficiency Gains with Concentriq Embeddings

These same embeddings were obtained using the Automated Approach (Concentriq Embeddings). Since the IMPRESS images were already located in a repository in Concentriq, generating embeddings simply required submitting a request to the Concentriq Embeddings API with the repository ID. Concentriq Embeddings then processed the dataset in just 2.5 hours, with the ability to scale further as needed through cloud-based autoscaling.

When finished processing, Concentriq Embeddings saved 18 GB of embeddings and 1 GB of image thumbnails to cloud storage. From there, we accessed the embeddings in the cloud and completed the remaining processing on the same Linux workstation mentioned earlier. At a download speed of 584 Mbps, the Linux Workstation downloads 19 GB from cloud storage in 260 seconds. Thus, the total time to obtain embeddings using Concentriq Embeddings for the IMPRESS dataset was 2.5 hours.

> The study demonstrated that Concentriq Embeddings reduces the time to generate embeddings by 92% (a 13x decrease from 33.4 hours to 2.5 hours), and reduces storage requirements by 91% (an 11x decrease from 203 GB to 19 GB).

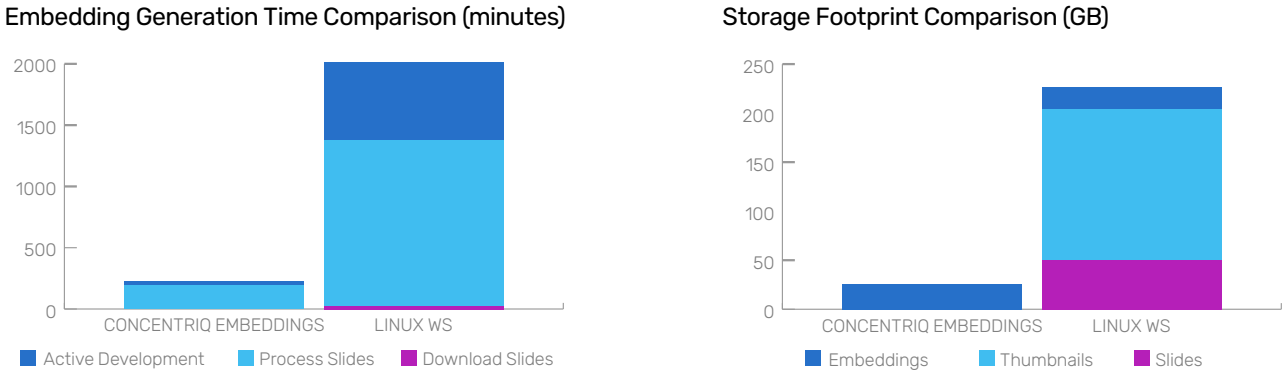

Figure 2: Left: Generating embeddings with Concentriq Embeddings takes approximately 8% of the time required by the Manual Approach (using a Linux Workstation). Right: The storage footprint of the embeddings is about 9% of the combined storage required for slides, thumbnails, and embeddings using the Manual Approach.

Accelerated Downstream Model Prototyping with Concentriq Embeddings

By streamlining the most computationally intensive part of modern deep learning model development—embeddings generation—Concentriq Embeddings enables running multiple experiments on simple hardware. After just a few minutes of active development to submit the requests, Concentriq Embeddings scales to produce embeddings from four foundation models.

We trained a Multiple Instance Learning (MIL) architecture model for each of 10 biomarkers present in the IMPRESS dataset using the four foundation models available in Concentriq Embeddings: DINOv2, PLIP[7], ConvNext[8], and CTransPath[9]. This prototyping exercise enabled a rapid assessment of the signal present for detecting breast biomarkers across different stains in the dataset, leveraging four foundation models in the process.

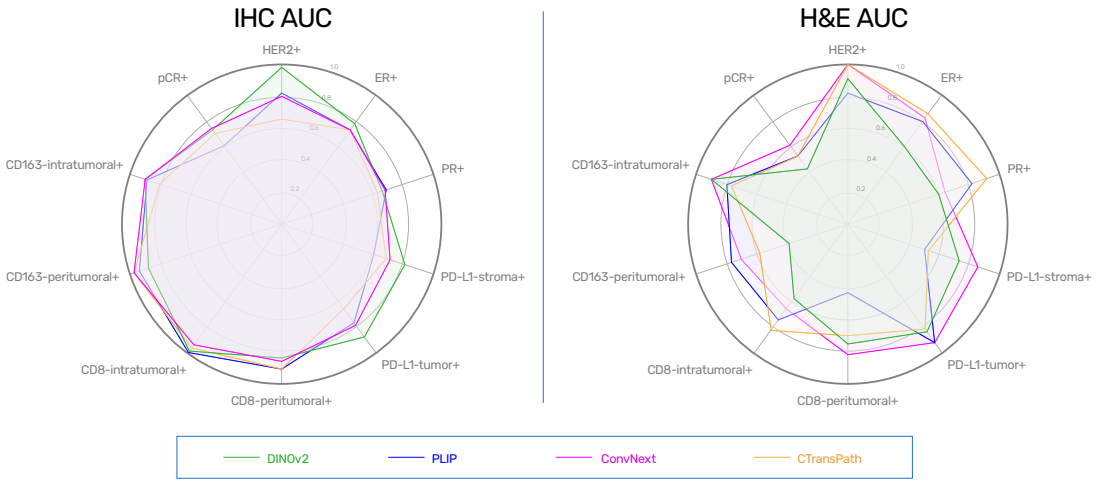

Figure 3: Concentriq Embeddings facilitates quick prototyping of models. Left: Foundation model detection of 10 breast cancer biomarkers using IMPRESS IHC slides. Right: Foundation model detection of the same 10 biomarkers using IMPRESS H&E slides. All model development was done on a consumer-grade laptop.

Using AI to screen images for predictive biomarkers can greatly enhance drug development, allowing for rapid and cost-effective screening of patient samples, quickly identifying those that require further molecular testing. Additionally, AI-based pre-screening tools can be used to stratify patients for clinical trials, potentially lowering costs and increasing success rates.

Discussion

With this study, we demonstrate drastically reduced model development time using the Automated Approach with Concentriq Embeddings compared to the traditional Manual Approach on a Linux Workstation. It is worth noting several challenges that this work does not address.

Dataset Size—First, the comparison is based on a dataset (IMPRESS) that is relatively lightweight when compared to standard pathology datasets. Applying Concentriq Embeddings to a heavier and more standard dataset (e.g., CAMELYON17, which contains 1,000 slides and is 3.0 TB in size), would further validate its scalability and efficiency. Using the same thumbnail creation speed as for IMPRESS (a conservative estimate given that an IMPRESS slide averages only 0.2 GB, while CAMELYON17 slides average 3.0 GB), creating CAMELYON17 thumbnails would take 44 hours, and tiling and inference would take 48 hours.

Moreover, since Concentriq Embeddings can horizontally scale, it can produce embeddings for CAMELYON17 in roughly the same amount of time as it took to process IMPRESS. This results in over a 100X speedup using Concentriq Embeddings when processing a more typically sized dataset.

Dependency Management—The active development time measured in this experiment is likely an underestimation compared to typical workflows, and the time required for a single prototyping experiment can be highly variable. If the developer has a Python environment already prepared with every processing dependency (e.g., packages required for opening slides and deep learning inference), the startup time for development of digital pathology AI prototypes might be minimal.

However, obtaining such an environment is often complex. For example, the work described here required two environments. OpenSlide and CUDA libraries can easily conflict, as can dependencies between OpenSlide-Python and Pytorch. To isolate OpenSlide and avoid environment conflicts in this work, thumbnails were extracted in a different Python environment from the other processing steps (downloading slides and tiled inference).

Solving problems with libraries or Python dependencies has a well known epithet: “Dependency Hell.” Dependency Hell can easily consume weeks of developer time to escape and adds significant, unforeseen risk to a project. Concentriq Embeddings alleviates these challenges through its cloud-based, automated environment, which minimizes the need for manual setup and reduces project risk associated with dependency conflicts.

Multiple Foundation Models—This work presented model development with the Manual Approach using a single foundation model. However, many prototyping experiments require embeddings from multiple foundation models, as demonstrated by the biomarker prediction results we achieved using Concentriq Embeddings.

Working with multiple foundation models increases the required active development time in traditional workflows, as each model must be processed in a series, increasing the tiling and embedding time. Concentriq Embeddings’ horizontal scalability in the cloud allows for simultaneous processing across multiple foundation models without increasing overall processing time, improving the likelihood of downstream model development producing impactful and actionable results.

Cost Efficiency—Beyond the time savings, Concentriq Embeddings offers cost benefits by reducing the need for high-end hardware, expensive compute and storage, and time-consuming activities to address operational challenges traditionally associated with leveraging foundation models. To learn more about the computational savings associated with Concentriq Embeddings, read our detailed analysis which presents the financial advantages of adopting this solution.

Conclusion

The findings of this study underscore the transformative impact of Concentriq Embeddings on the field of digital pathology. By drastically reducing processing times across multiple foundation models, facilitating rapid prototyping, and overcoming operational barriers such as dependency management and resource limitations, Concentriq Embeddings enables data scientists to accelerate the development of downstream AI models and enhance their likelihood of success. With this streamlined and scalable solution for modern AI development, life sciences organizations can push the boundaries of AI, driving forward the potential for groundbreaking discoveries and ultimately, improved patient outcomes.

About Proscia

Proscia is a software company accelerating pathology's transition to a digital, data-driven discipline and enabling AI to advance precision medicine. Its Concentriq enterprise pathology platform, precision medicine AI portfolio, and real-world data fuel the development and use of novel therapies and diagnostics to drive the fight against humanity's most challenging diseases, like cancer. 14 of the top 20 pharmaceutical companies and a global network of diagnostic laboratories rely on Proscia's solutions each day. The company has FDA 510(k) clearance and was the first to secure CE-IVDR certification to advance digital pathology primary diagnosis in the European Union.

Learn more about Concentriq Embeddings

Request a Demo

References

1. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850-862. doi:10.1038/s41591-024-02857-3
2. Tarantino P, Mazzarella L, Marra A, Trapani D, Curigliano G. The evolving paradigm of biomarker actionability: Histology-agnosticism as a spectrum, rather than a binary quality. Cancer Treat Rev. 2021;94:102169. doi:10.1016/j.ctrv.2021.102169
3. Giraldo NA, Kaunitz GJ, Cottrell TR, et al. The differential association of PD-1, PD-L1, and CD8 + cells with response to pembrolizumab and presence of Merkel cell polyomavirus (MCPyV) in patients with Merkel cell carcinoma (MCC). Cancer Res. 2017; 77; 662. doi: 10.1158/1538-7445.AM2017-662
4. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-477. doi:10.1038/s41573-019-0024-5
5. Huang Z, Shao W, Han Z, et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ Precis Oncol. 2023;7(1):14. Published 2023 Jan 27. doi:10.1038/s41698-023-00352-5
6. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. 2024. arXiv. doi:2304.07193. https://arxiv.org/abs/2304.07193
7. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307-2316. doi:10.1038/s41591-023-02504-3
8. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022. arXiv. doi:2201.03545. https://arxiv.org/abs/2201.03545
9. Wang X, Yang S, Zhang J, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal. 2022;81:102559. doi:10.1016/j.media.2022.102559

Document #1006 V1 10-2024