Deliverable D2.2

Deployment kit for user institutions to replicate the cloud infrastructure (accessible from the AI4Life website)

Project Title	Artificial Intelligence For Image Data Analysis In The Life Sciences
Project Acronym	AI4Life
Project Number	101057970
Project Start Date	01.09.2022
Project Duration	36 Months

WP N° & Title	WP2: User Services and Computing Infrastructure
WP Leaders	KTH and EMBL-EBI
Deliverable Lead Beneficiary	KTH
Dissemination Level	PU
Contractual Delivery Date	30.04.2024 (M20)
Actual Delivery Date	29.04.2024
Authors	Wei Ouyang, Craig Russell
Contributors	Weize Xu, Joanna Hård
Reviewers	Estibaliz Gómez de Mariscal, Beatriz Serrano-Solano, Arrate Muñoz-Barrutia, Dorothea Dörr

AI4Life has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101057970. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

Change Log

Version	Date	Author	Description of changes
v0.1	08.04.2024	Joanna Hård	Initial draft
v0.2	15.04.2024	Wei Ouyang	Edits and suggestions
v0.3	18.04.2024	Estibaliz Gómez de Mariscal	Edits and suggestions
v0.4	23.04.2024	Beatriz Serrano-Solano, Dorothea Dörr	Edits and suggestions
v0.5	24.04.2024	Weize Xu, Craig Russell	Edits and suggestions
v0.6	29.04.2024	Wei Ouyang	Final draft approved for submission

Acronyms and Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
D	Deliverable
de.NBI	German Network for Bioinformatics Infrastructure
GCP	Google Cloud Platform
GPU	Graphical Processing Unit
HPC	High-Performance Computing
K8s	Kubernetes
PU	Public
RI	Research Infrastructure
SLURM	Simple Linux Utility for Resource Management
v	Version

Table of contents

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

Executive Summary

AI4Life continues its mission to democratize access to state-of-the-art Artificial Intelligence (AI) methods in biological imaging by expanding the capabilities and reach of its initiatives. A cornerstone of this effort is the enhancement of infrastructure to support the widespread deployment and use of AI tools through the BioImage Model Zoo. This commitment is embodied in the current deliverable, which focuses on providing robust deployment toolkits for replicating and customizing the AI4Life computing infrastructure across diverse institutional environments.

The aim of **Deliverable (D) 2.2** is to create a deployment toolkit designed to allow user institutions to autonomously establish and manage their version of the BioEngine—our advanced AI model serving platform. This toolkit supports a variety of IT infrastructures, including those managed by High-Performance Computing (HPC) centers with Simple Linux Utility for Resource Management (SLURM) and production environments using Kubernetes (K8s). Featuring the Hypha Launcher, this versatile module facilitates integration and dynamic management of AI resources, ensuring that BioEngine can operate effectively in any setup. It includes features for managing containerized applications and supports automatic scaling to meet user demand efficiently.

Throughout the past months, the deployment toolkit has been refined and tested across several HPC clusters and Kubernetes environments, including Google Cloud Platform and the de.NBI cloud infrastructure. These implementations underscore our toolkit's readiness and reliability for broad adoption. Developed in collaboration with key AI4Life partners and designed for ease of use and scalability, the toolkit is set to significantly lower the entry barriers to advanced AI applications in the life sciences, aligning with AI4Life's overarching goal of making AI both accessible and effective across the global scientific community.

1. Introduction

The AI4Life project, funded by the European Union's Horizon Europe research and innovation program, is a collaborative effort coordinated by Euro-BioImaging and supported by 10 partners, including four European Research Infrastructures. Commencing in September 2022 and extending until September 2025, AI4Life aims to democratize access to advanced AI-based image analysis methods. Recognizing the transformative potential of AI in microscopy image restoration, segmentation, and analysis, AI4Life seeks to overcome technological barriers hindering widespread adoption. To achieve this, the project focuses on establishing services tailored to both life scientists and computational methods developers in the AI and computer vision fields. Central to AI4Life's mission is the universal accessibility of AI tools, particularly for life scientists lacking specialized training in programming or AI. By making state-of-the-art AI-based image analysis accessible, AI4Life aims to empower researchers in the life sciences, driving scientific progress and innovation.

One of the seminal outputs of the AI4Life project has been the development of the BioEngine[1]—a robust cloud-based platform launched as part of Deliverable D2.1. This infrastructure has already made significant strides in making AI models from the BioImage Model Zoo[2] accessible and operational via user-friendly interfaces such as ImageJ.JS[3]. The BioEngine has facilitated effortless, on-demand AI computations, serving as a pivotal tool for both novice and expert users aiming to harness the power of AI for biomedical image analysis.

While BioEngine excels in providing a centralized platform for model testing and exploration, its design as a remotely hosted service means that users must upload their data to our servers. This setup, while sufficient for individual tests and evaluations, is not optimized for large-scale data processing or continuous use due to bandwidth constraints and concerns over data privacy and security. Moreover, the operational capacity of our servers, constrained by finite GPU resources, imposes further limitations on the scale and scope of data processing tasks that can be handled concurrently.

Recognizing these challenges and aligning them with our commitment to sustainable, scalable, and accessible AI solutions, we have developed a deployment toolkit. This toolkit will enable institutions to replicate the AI4Life computing infrastructure on-premise. Doing so will alleviate the need to transmit large datasets externally and will address privacy concerns, thereby enhancing trust and user autonomy in utilizing AI tools. Furthermore, this move towards localization of resources is in direct response to

[1] https://ai4life.eurobioimaging.eu/announcing-bioengine/
[2] https://bioimage.io
[3] https://ij.imjoy.io/

the necessity for high-volume data processing and the strategic scaling of AI capabilities within user-controlled environments. The deployment toolkit, which is the focus of this deliverable, marks a critical step in our ongoing efforts to provide comprehensive, user-empowered AI solutions that transcend the initial scope of remote model testing and bring the transformative power of AI directly into the hands of researchers and institutions worldwide.

2. Description of work

As part of our continuous efforts to enhance the flexibility and accessibility of the BioEngine, we have extended its capabilities to support diverse IT infrastructures commonly found in research environments. To this end, we have developed a comprehensive deployment toolkit that suits both HPC environments managed by SLURM and Kubernetes (K8s) clusters, which are more suited to production deployments.

2.1 Deploying BioEngine to HPC

Central to this initiative is the 'Hypha Launcher', a versatile module designed to simplify the deployment and management of AI resources. Accessible via Hypha Launcher: https://github.com/aicell-lab/hypha-launcher, it enables dynamic job launching in HPC settings and integration with cluster management software. Key features of the Hypha Launcher include:

- CLI/API Capabilities: Simplifies tasks such as downloading models from S3 storage, pulling Docker images of the Triton server, and launching S3 and Triton servers.
- Container Engine Support: Compatible with Docker and Apptainer, facilitating broad adoption across various compute environments like local Linux setups and SLURM clusters.

The Hypha Launcher serves as a gateway for institutions to launch the BioEngine worker locally, manage model deployments, and initiate the Triton Inference Server to serve models seamlessly within their own infrastructure. This integration with Triton Inference Server allows maximizing model serving to many users with limited GPU. Detailed instructions for setting up and running the BioEngine worker are provided here: https://bioimage-io.github.io/bioengine/#/bioengine-hpc-worker.

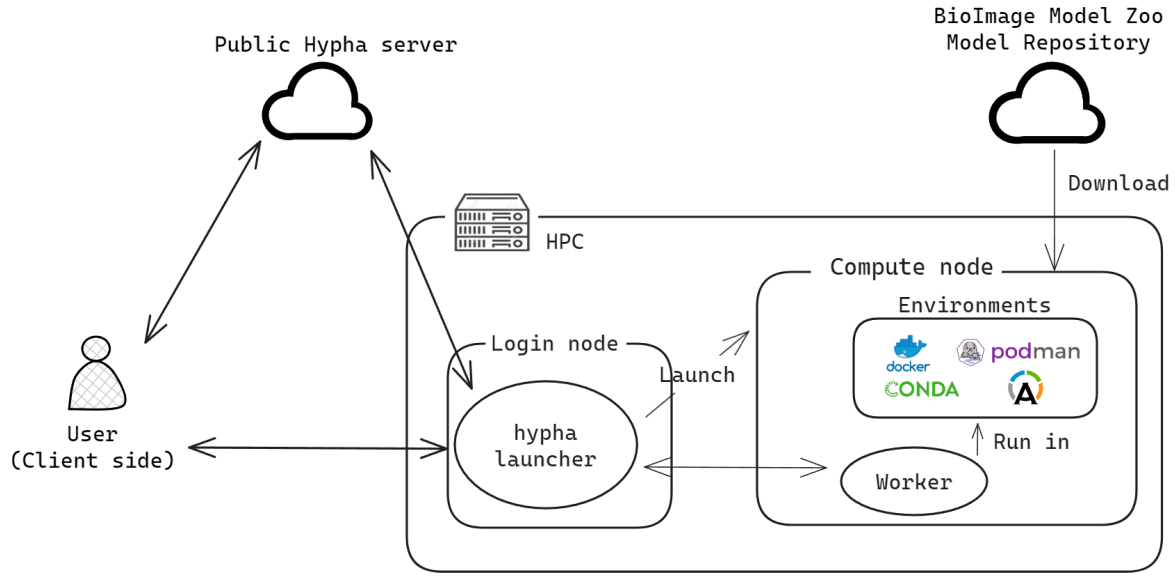

Figure 1. Diagram for the BioEngine worker on HPC. Users access the BioEngine either directly through the Hypha server running in the HPC or through our public Hypha server as a proxy. In a typical HPC setting, the BioEngine worker uses the Hypha launcher running on the HPC login node, which allows the launcher to dynamically initiate new Triton server containers on the compute node.

During our development, we have been testing the BioEngine worker among different HPC clusters, including:

- Alvis (an HPC cluster dedicated to AI research, hosted at Chalmers University of Technology),
- Berzelius (an AI/ML focused compute cluster, hosted at Linköping University)

Both of the HPC clusters utilize SLURM as their management software, yet the BioEngine platform is designed to function with other types of clusters as well.

To address potential network barriers and security concerns within institutional frameworks, we offer two distinct access options:

1. A local Hypha server providing an all-in-one solution, albeit with potential accessibility limitations outside institutional networks. This requires users to manage SSL certificates and domain names.
2. An alternative that connects local BioEngine instances to our public server, facilitating immediate online accessibility and external collaboration without complex network configurations.

2.2 Kubernetes Integration and Scalability

Recognizing the need for scalable and resilient AI applications, our toolkit includes dedicated resources for setting up Kubernetes environments:

- Helm Chart for BioEngine Deployment: We provide a Helm chart to streamline the Kubernetes deployment process. This and other useful materials are available in our Hypha Helm Chart's Repository: https://github.com/bioimage-io/hypha-helm-charts.
- Deployment Examples: Our Hypha Kubernetes Repository (https://github.com/bioimage-io/hypha-kubernetes) offers practical examples to assist users in configuring the BioEngine within a Kubernetes framework.

Kubernetes has been identified as a key target due to its exceptional flexibility and rapidly expanding user base in standard cloud deployments. It is supported by major private clouds and it is prevalent in most public clouds, particularly in Europe. Additionally, Kubernetes can be deployed locally for development purposes on workstations and small servers using K3s (a lightweight Kubernetes deployment from Rancher). As an open-source platform with a robust community, Kubernetes ensures long-term support.

Further guidance on Kubernetes deployment is available through our detailed Kubernetes toolkit guide: https://bioimage-io.github.io/bioengine/#/k8s-toolkit.

Our deployment toolkit has been tested under the following Kubernetes clusters:

- Google Kubernetes Engine provided at Google Cloud Platform,
- de.NBI Kubernetes cluster (a Kubernetes cluster provided by the German Network for Bioinformatics Infrastructure).

To enhance the scalability and resilience of the BioEngine within Kubernetes environments, significant advancements have been integrated into our Helm Chart and deployment strategies. These enhancements focus on ensuring high availability and cost efficiency across multiple data centers, leveraging advanced Kubernetes features and third-party tools.

Helm Chart and Deployment Strategy

1. Helmsman for Multi-Datacenter Deployment:
   For our deployment stack, we have utilized Helmsman, a Helm Charts as a code tool, to manage deployments across multiple data centers (de.NBI, EMBL-EBI Embassy cloud and Google Cloud). This approach not only simplifies the

management of Helm Chart deployments but also enhances disaster recovery by distributing resources geographically.

2. Version Control for Stability and User Deployment:
   Our Helm Charts are version-controlled within the Hypha Helm Charts Repository. This practice allows users to deploy specific versions of the BioEngine, ensuring stability and compatibility with diverse Kubernetes environments.

3. Resource Duplication for High Availability:
   To address the potential downtime associated with individual model runner failures, we deployed critical components with duplication. This redundancy ensures that the service remains operational, even if a single model runner instance fails.

Cost Optimisation Strategies

1. Use of Google Cloud Spot Instances:
   By utilizing Google Cloud Spot Instances, we significantly reduced operational costs. Spot Instances allow us to bid on unused Google Cloud capacity at a price lower than the standard rate.

2. Kubernetes Autopilot (Google Cloud Platform only) for Efficient Resource Utilization: Deploying the BioEngine on Google Kubernetes Autopilot enables us to optimize costs further. This managed service automatically adjusts the quantity and size of resources, charging only for the Spot Pods used. This flexibility is particularly beneficial in our context, where model runners may be sporadically terminated due to our use of cost-effective Spot Instances.

Resiliency

We have made further improvements to the deployment resiliency within the Helm Chart for the BioEngine and our own main BioEngine deployment on GCP.

1. The Helm Charts have been extended to include model-runner service replication.
2. We now run the BioEngine for the BioImage Model Zoo in dual replication so that if an individual model runner dies or fails, it is automatically replicated with no loss to service (green/blue deployment).
3. Our current deployment on GCP requires this as a cost-saving as well as, the more parsimonious compute is evicted daily.

These strategic enhancements to our Kubernetes deployment and Helm Chart configurations not only ensure that the BioEngine remains robust and scalable but also

optimize operational costs. The integration of Helmsman, careful versioning, redundancy in deployment, and cost-effective resource utilization exemplify our commitment to delivering a resilient and economically viable solution. These considerations will allow widespread adoption of the platform through ease of deployment, and service resiliency.

During the development, we have been focusing on optimizing the deployment to improve the robustness of the system while reducing running costs. Figure 2 shows a screenshot of the running cost for our deployment at Google Cloud Platform:

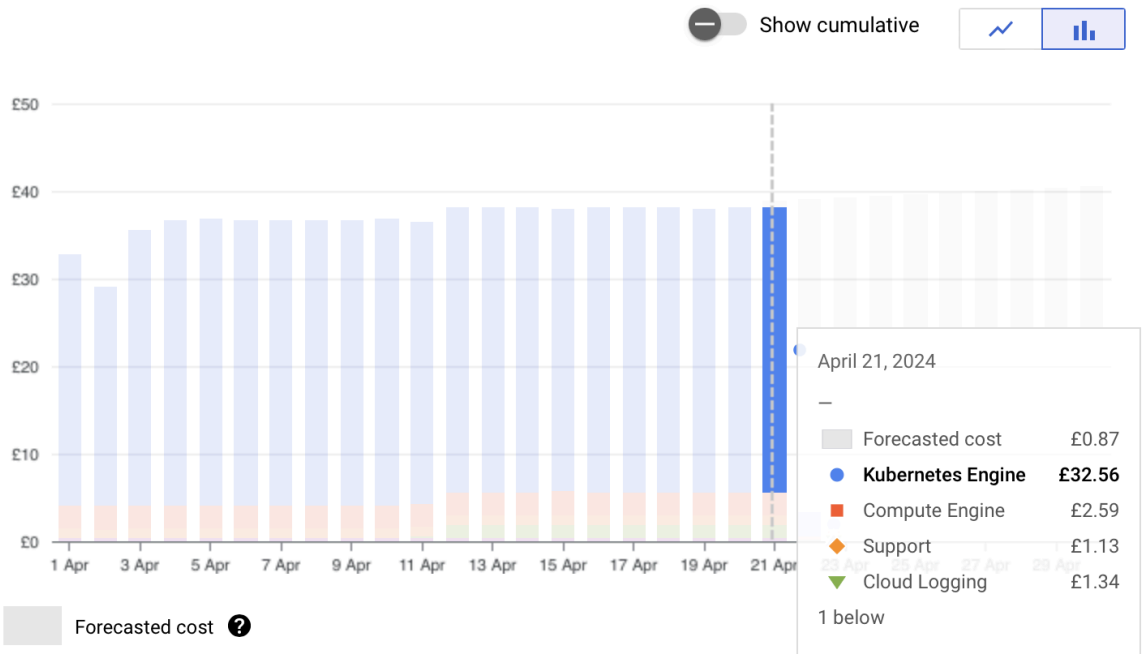

Figure 2. Operational cost for the BioEngine deployed on the Google Cloud Platform. While these costs fluctuate over time due to the dynamic nature of spot instances, the costs to run and provide BioEngine user inference access to all of the models in the BioImage Model Zoo with full resiliency and access to 2 NVIDIA T4 GPUs are low (~€47 per day).

2.3 Enhancing User Accessibility and Interactivity

To ensure that users can easily test and utilize the models within their own environments, we have developed the BioEngine Web Client. This web application allows users to connect to any custom BioEngine server and operate models from the BioImage Model Zoo directly from their browser via the ImageJ.js interface:

- Standalone Web app: https://bioimage-io.github.io/bioengine-web-client
- Integration to BioImage.IO: https://bioimage.io (on every model card)

- Source code: https://github.com/bioimage-io/bioengine-web-client

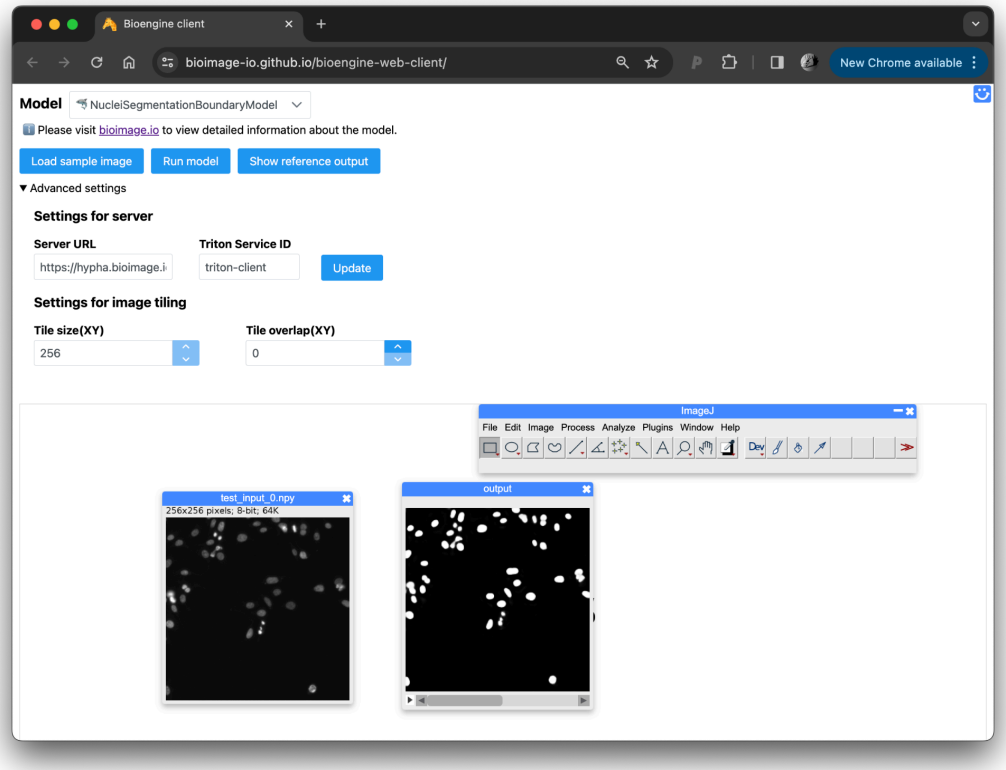

Figure 3. Screenshot of the BioEngine Web Client

The BioEngine Web Client is also integrated into the BioImage Model Zoo website to support model test-run features.

Additionally, BioEngine supports programmatic access through a documented API, enabling developers to integrate AI-powered image analysis tools into their applications effortlessly. The API documentation can be found here: https://bioimage-io.github.io/bioengine/#/api. This simplifies the integration process for developers, allowing them to focus more on application development rather than the complexities of underlying AI technologies.

2.4 Future Developments and Ecosystem Expansion

To further support the development and integration of AI tools, we are in the process of establishing a BioEngine application standard, which will be detailed in Deliverable D2.4. Concurrently, we are exploring the development of a desktop version of BioEngine using

Podman Desktop to provide a robust, containerized environment suitable for desktop applications. This initiative aims to complement our existing toolkit for HPC and Kubernetes environments, thereby broadening the potential for local and cloud-based deployments.

We are now in the process of deploying the BioEngine via Kubernetes to the European Grid Infrastructure as part of a trial agreement. The work is mostly ready but for some technical networking issues and security considerations.

3. Conclusion

The AI4Life project has made substantial progress in democratizing AI technology for image analysis in the life sciences, driven by the development of the BioEngine and its deployment toolkit. These innovations address critical barriers such as data privacy, operational cost, and scalability, and provide flexible infrastructure options through the Hypha Launcher for HPC and Kubernetes systems. This approach not only facilitates the broader adoption of AI tools but also allows better data security.

Looking ahead, AI4Life plans to further expand the BioEngine ecosystem with initiatives like the BioEngine application standard. These developments aim to simplify AI tool deployment and make advanced computational technologies more accessible to researchers and developers, irrespective of their infrastructure setup. We are actively reaching out to partners and users to test out the system, aiming to support a more diverse range of computing infrastructure. The ongoing enhancement of our platforms is expected to streamline operations and reduce costs, making AI applications more sustainable in research environments.

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam