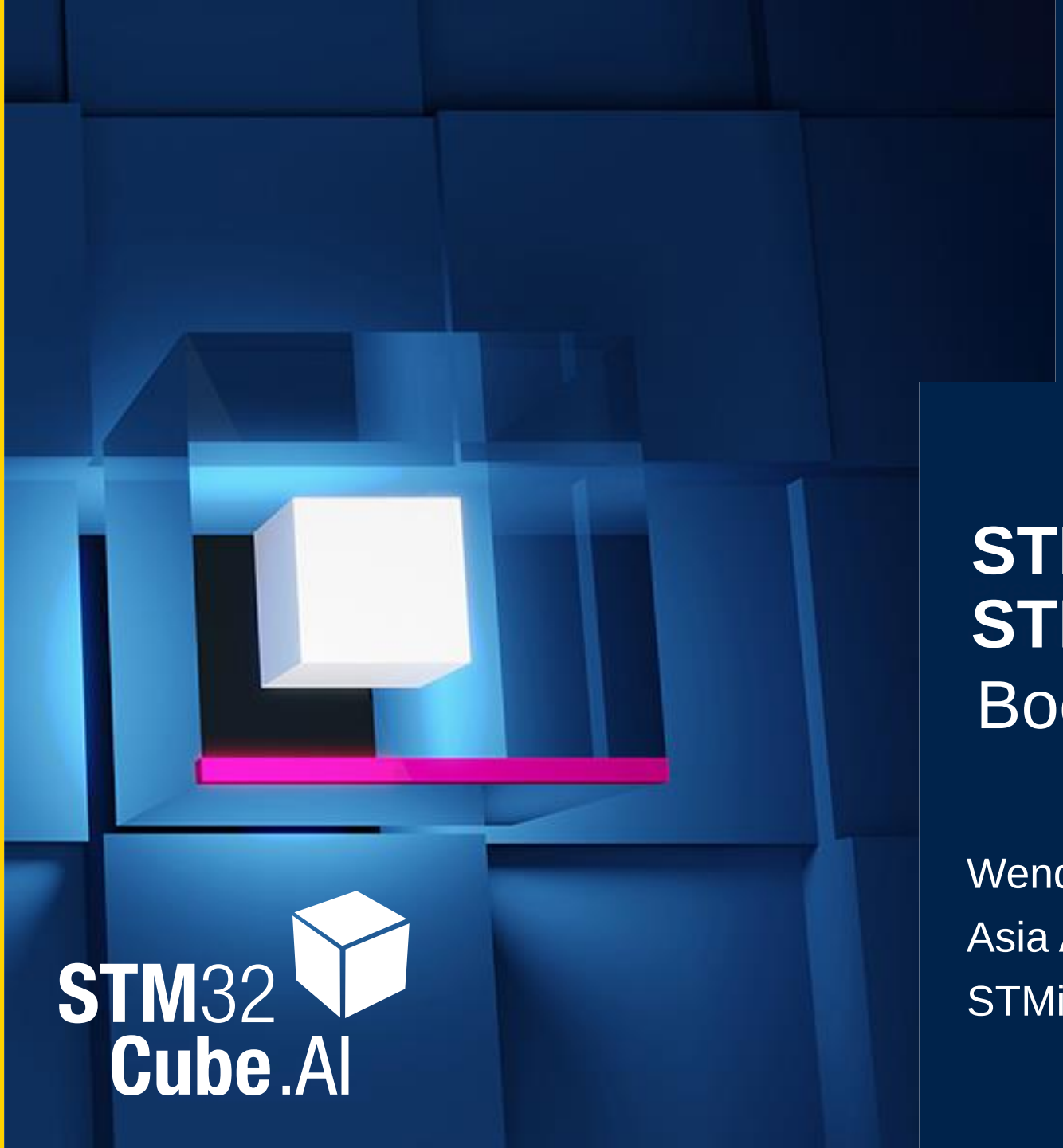


life.augmented

INDUSTRIAL SUMMIT 2023

SHENZHEN, CHINA | 28 SEPTEMBER

STM32Cube.AI
STM32Cube.AI Develop Cloud
Boost Your Edge AI Application

Wendy LI

Asia AI Competence Center

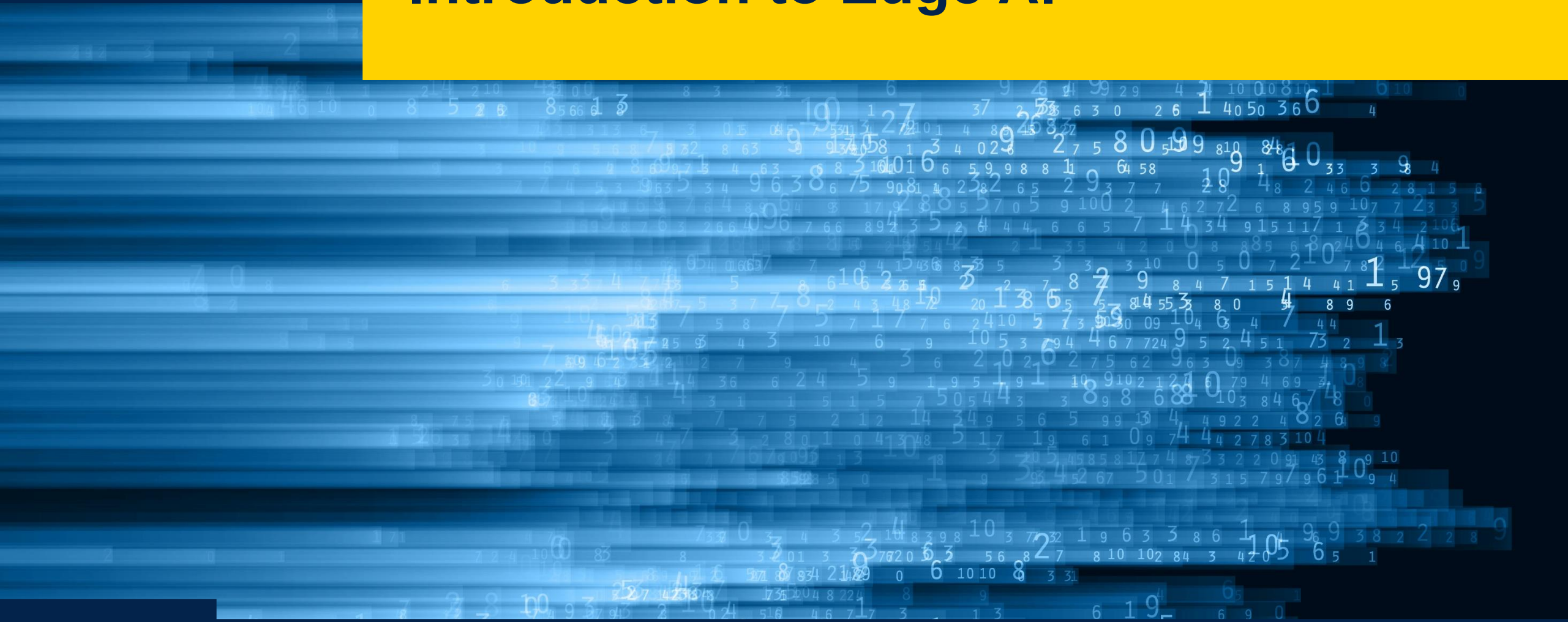
STMicroelectronics

STM32 Cube.AI

Agenda

1. Introduction to Edge AI
2. STM32Cube.AI- AI optimization tool for STM32
3. STM32Cube.AI for Desktop
4. STM32Cube.AI Developer Cloud

Introduction to Edge AI



life.augmented

Signals turning into data

Embedded applications will collect more data in the future

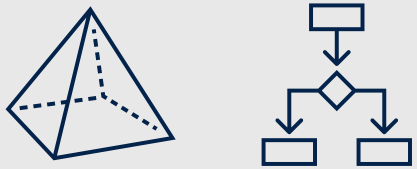
Growing demand for data-driven insights

Increasing use of sensors

Proliferation of IoT devices

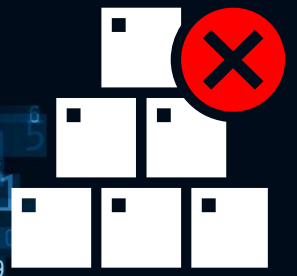
life.augmented

AI is offering the best approach to process this growing amount of data



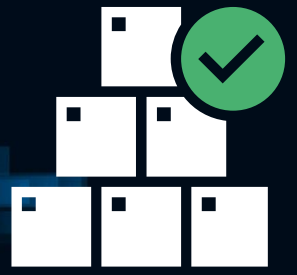
Algorithms and **predefined models** to analyze data and make predictions or decisions

Machine learning algorithms to automatically **learn** patterns and relationships **from the data**



Traditional approaches show their limitations:

- when dealing with **large datasets**
- when the **phenomena are too complex**



AI-based data processing offers a more flexible and powerful approach to analyzing and making decisions from large data collection

life.augmented

The raise of Edge AI

Ultra-low latency
Real-time applications

Reduced data transmission
Generate meaningful information

Enhanced privacy and security
No data sharing in the cloud

Sustainable on energy
Low-data / Low-power

Improved accuracy
analyze data from a wide range of sensors and sources

Edge AI will benefit to many application domains:

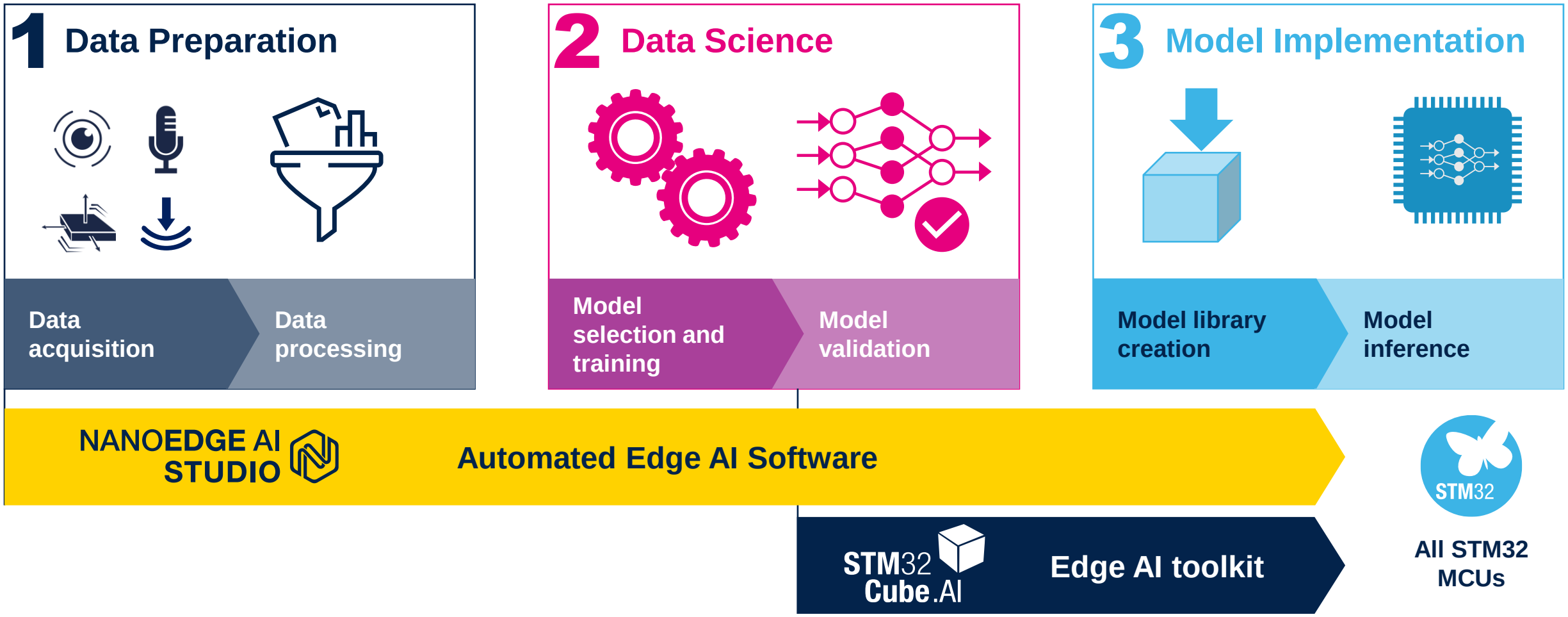
Industrial maintenance
Condition monitoring
Predictive maintenance

Control systems
From home heating systems to industrial machines

Internet of Things (IoT)
smart cities, smart buildings, connected homes, and industrial automation

life.augmented

AI development workflow – ST software offering



life.augmented

ST ecosystem ease your AI to reach production level

	STM32 Cube.AI **Edge AI toolkit** for model optimization on STM32	NANOEDGE AI STUDIO **Automated ML Software** for end-to-end Edge AI solution design on STM32
Key benefits	✓ Get optimized C-code from your trained model ✓ Desktop and online versions ✓ Benchmark service on remote hardware (online version) ✓ On-device performance validation	✓ The easiest way to integrate AI into your system ✓ Save resource and development cost ✓ Reach the highest performance with the automated model finder embedded in the tool
Application domain	All	Time series (except voice and speech)
Business model	Free of charge	Free for prototyping on STM32 dev boards Production requires right of use

life.augmented

A proven technology already adopted by multiple clients

NANOEDGE AI STUDIO

INDUSTRIAL | DEMO

Fan anomaly detection based on vibrations

Learn to detect abnormal behavior at the edge on a vibrating machine.

nke WATTECO

NANOEDGE AI STUDIO

INDUSTRIAL | CUSTOMER

AI solution for industrial predictive maintenance with NKE Watteco

Predictive maintenance solution for industrial equipment.

CROUZET

NANOEDGE AI STUDIO

TRANSPORTATION | CUSTOMER

AI solution for monitoring automatic doors with Crouzet

Predictive maintenance on motors for automatic door motors.

NANOEDGE AI STUDIO

INDUSTRIAL | DEMO

Anomaly detection in an electric motor

Current sensing to detect abnormal behaviors in motors.

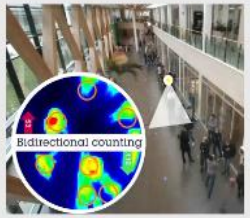
OXYtronic

NANOEDGE AI STUDIO

INDUSTRIAL | CUSTOMER

AI solution for industrial predictive maintenance with Oxytronic

Predictive maintenance solution for industrial equipment.



Schneider Electric

STM32 Cube.AI

SMART OFFICE | CUSTOMER

People flow counting Sensor with Schneider Electric

An innovative approach to measure people flows using an in-house thermal sensor.

STM32 Cube.AI

SMART CITY | DEMO

Acoustic scene classification

Identify different environments (indoor, outdoor, in-car) using a simple microphone.

STM32 Cube.AI

WEARABLES | DEMO

Human Activity Recognition

Easily identify 5 different activities with a 3D accelerometer.

STM32 Cube.AI

INDUSTRIAL | DEMO

People presence detection (visual wake word)

Human detection on high-performance MCU.

STM32 Cube.AI

INDUSTRIAL | DEMO

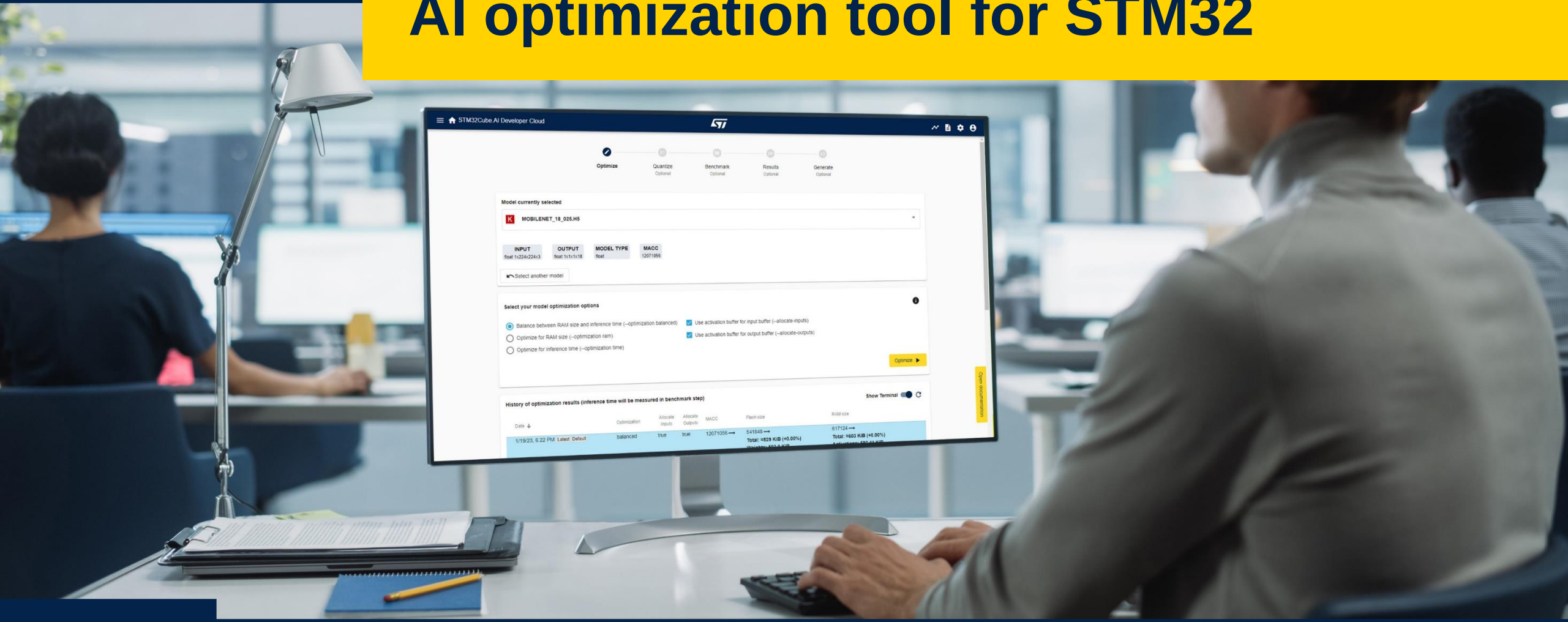
Aftermarket wireless digit reader

Equip meters with aftermarket wireless & low-power readers.

More case studies in ST website: https://stm32ai.st.com/use-cases/

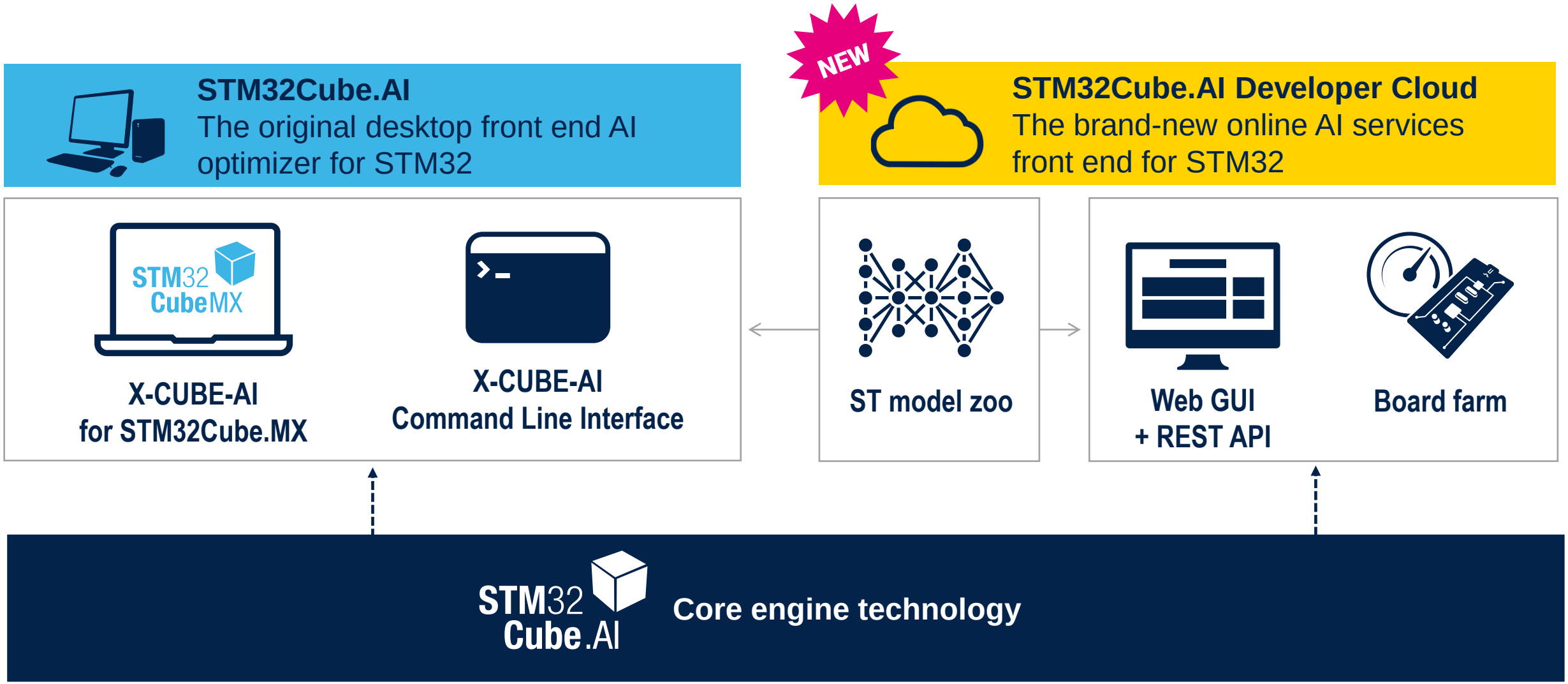
life.augmented

STM32Cube.AI
AI optimization tool for STM32



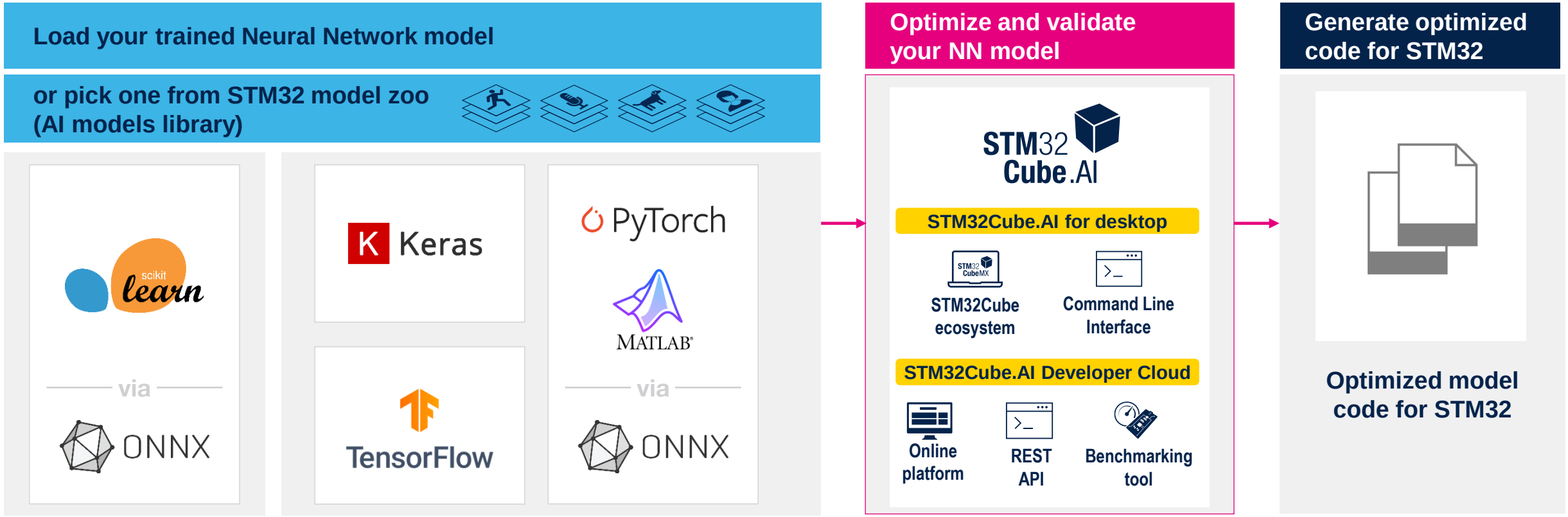
life.augmented

STM32Cube.AI overview



life.augmented

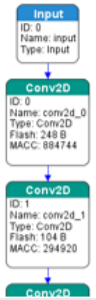
One tool – two versions to deploy AI on STM32



The 3 pillars of STM32Cube.AI

Graph optimizer

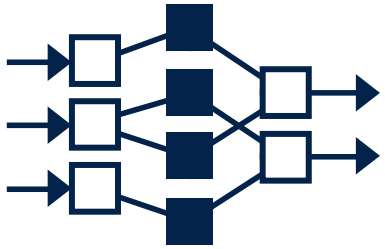
Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding…
- Operator-level info to fine-tune memory footprint and computation

Quantized model support

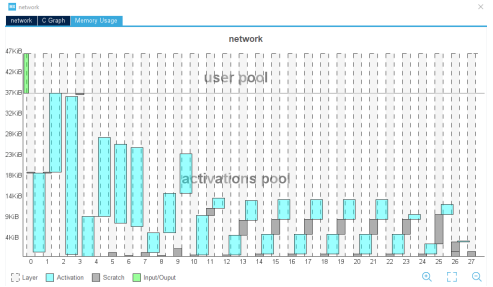
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8 or mixed-precision
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory footprint

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design



- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

life.augmented

Start with edge AI optimized models STM32 model zoo

A collection of application-oriented models optimized for STM32

Human activity

Motion Sensing

Image classification

Computer vision

Audio event detection

Audio classification

Object detection

Computer vision

Hosted on Github

Model training scripts

- Scripts to generate and validate

Application code example

- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation

life.augmented

What's new in STM32Cube.AI v8.0.0?

v8.0.0

Bringing higher degree of versatility with STM32Cube.AI

ONNX quantized models support

Introducing the support of **ONNX Tensor-oriented file format (QDQ):**

Initial model converted in ONNX QDQ can be:
- ONNX models
- Quantization-Aware training (QAT) models from Tensorflow or exported from PyTorch
- Quantized models converted from TFLite and other frameworks

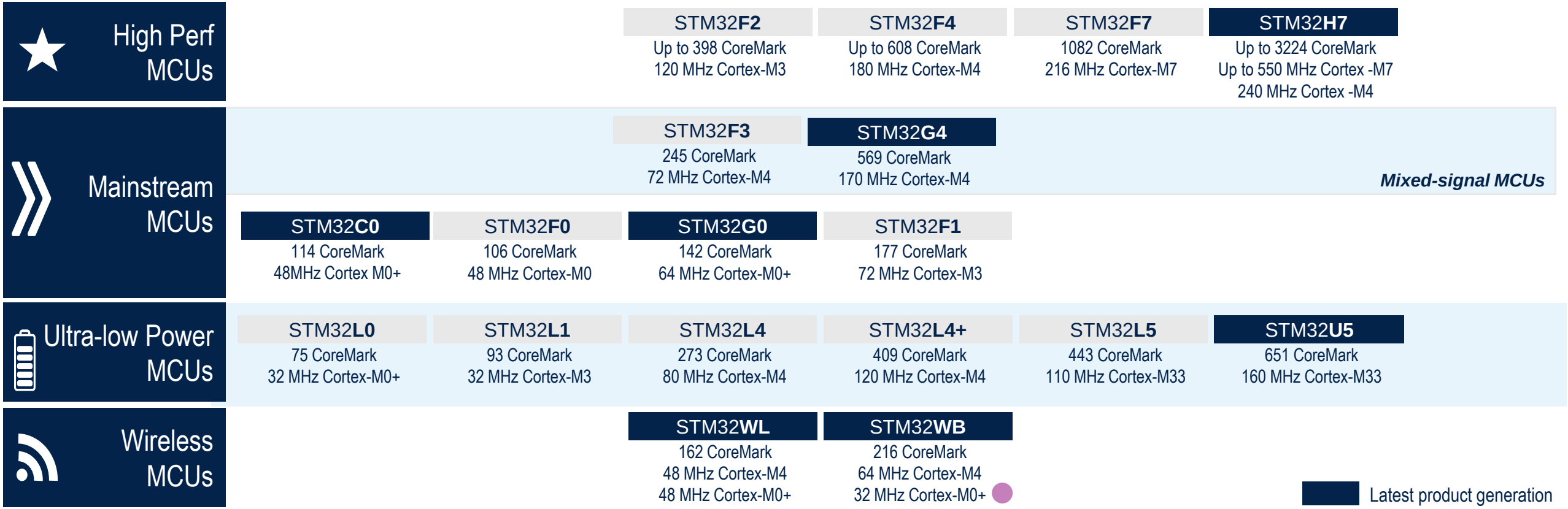
Up-to-date and improved code generation

- **Support for TensorFlow 2.11 models**
- **Support Keras.io 2.11**
- **Support ONNX Runtime 1.12.1**
- New kernel performance improvements.

life.augmented

Making Edge AI accessible to all STM32 portfolio

Take advantage of STM32Cube.AI on all STM32 series

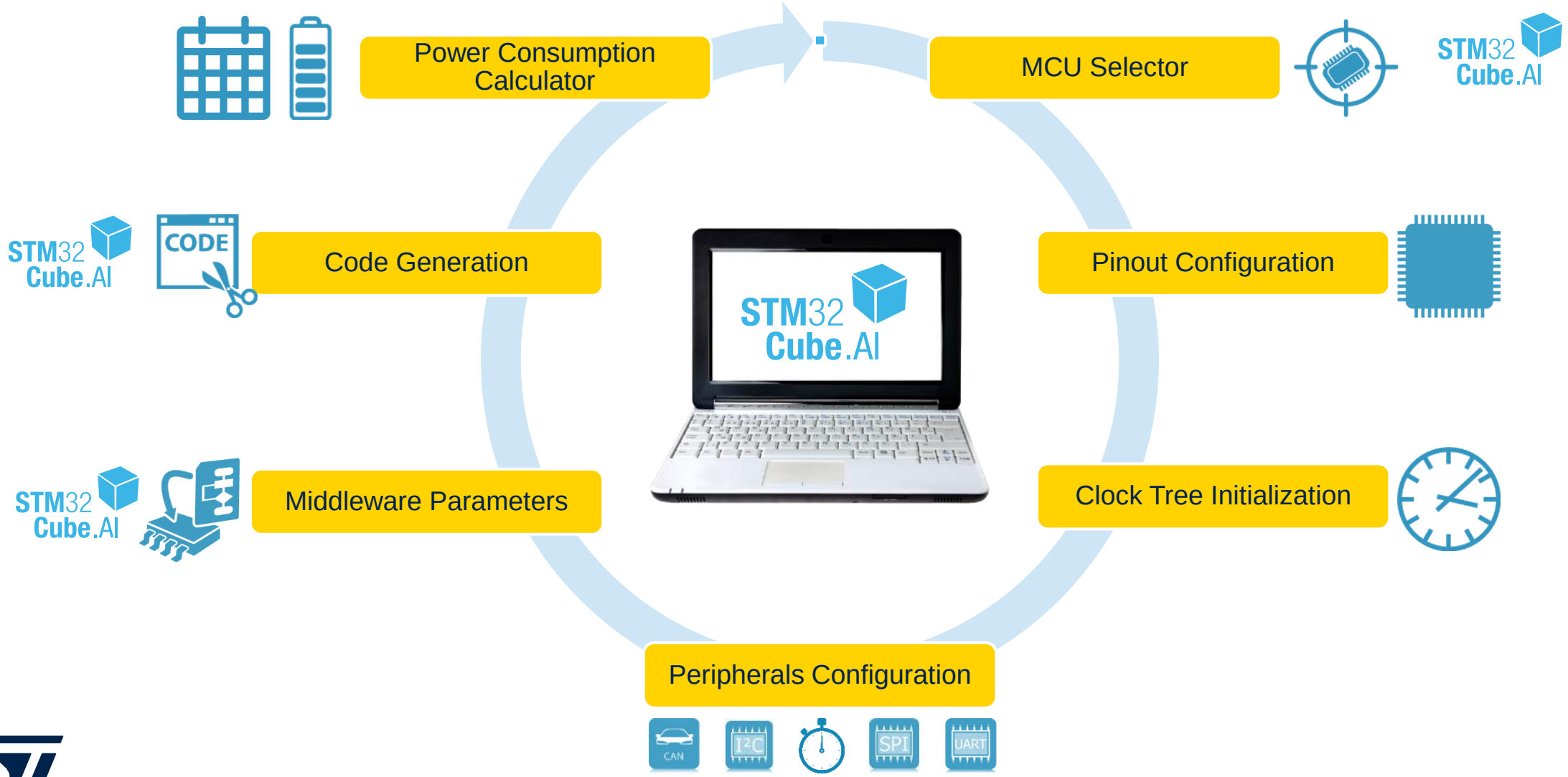


life.augmented

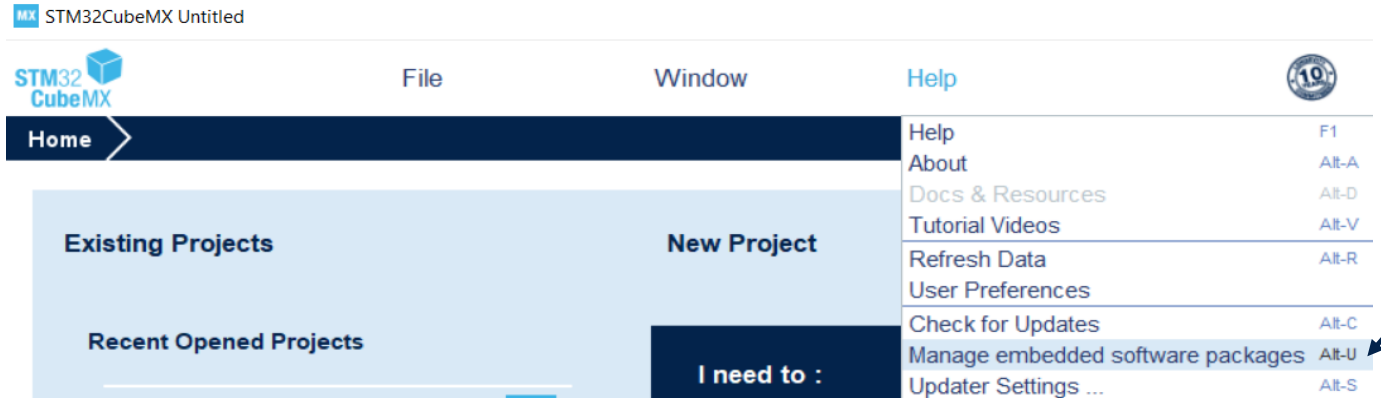
STM32Cube.AI for Desktop

life.augmented

STM32Cube.AI, an STM32CubeMX expansion

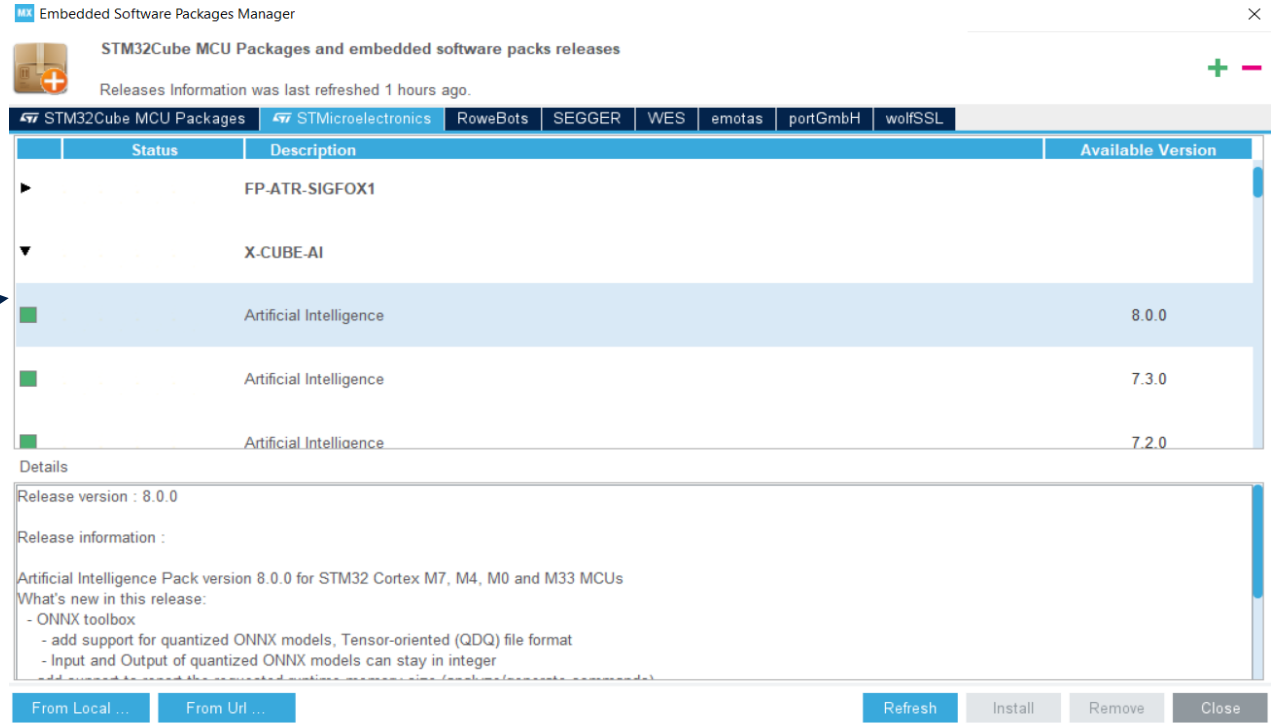


Install X-Cube-AI in STM32CubeMX



- Select "**Manage embedded software packages**" in **Help** item.

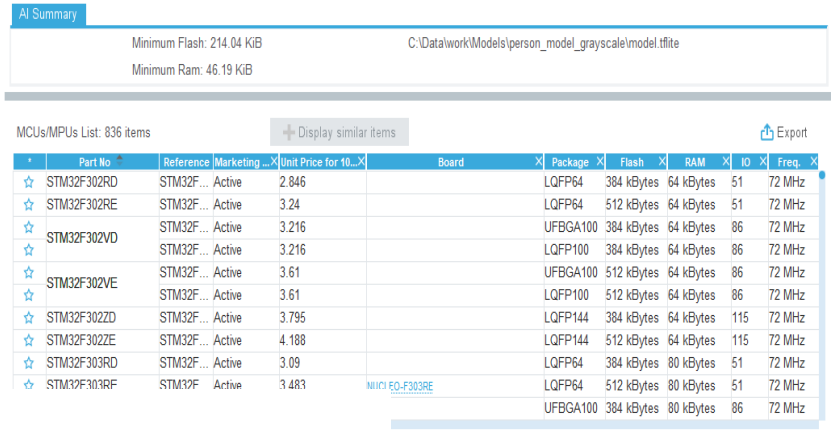
- find **X-CUBE-AI** in STMicroelectronics
- Install **the latest version(8.0.0)** in X-CUBE-AI.
- After installed, the square will turn on green in front of Artificial intelligence.



life.augmented

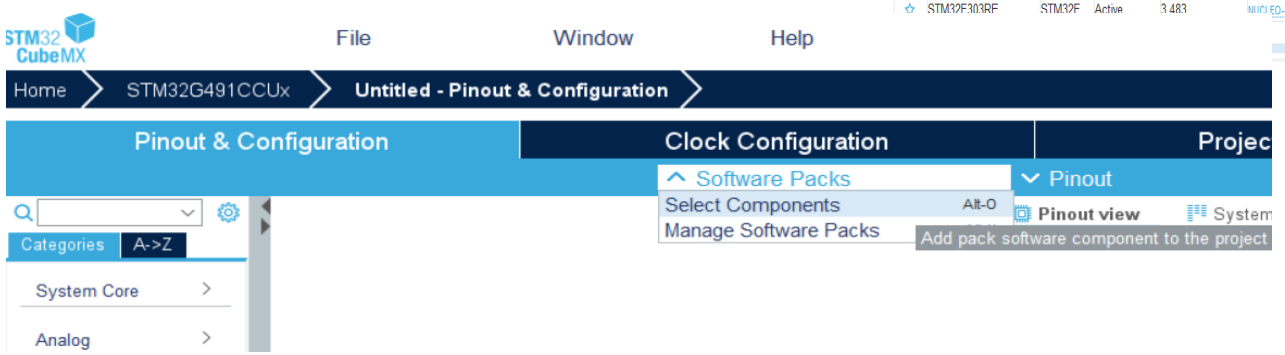
STM32Cube.AI user flow (1/4)

1. **Select MCU, add X-CUBE-AI pack**
   - Select corresponding STM32 MCUs
   - Select components X-CUBE-AI
   - Find X-CUBE-AI in left software packs.

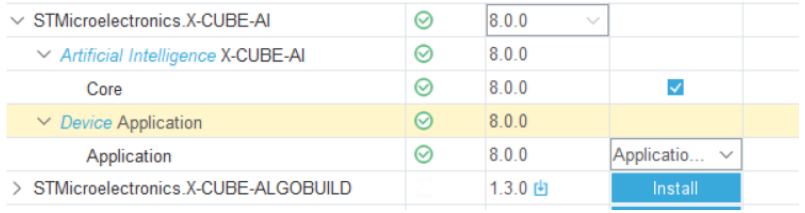


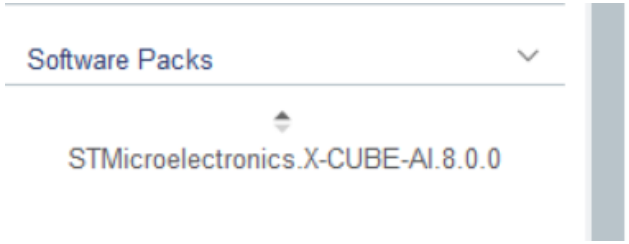
2. **Load and analyze NN Model**

3. **Optimize and validate**



4. **Generate project and integrate**

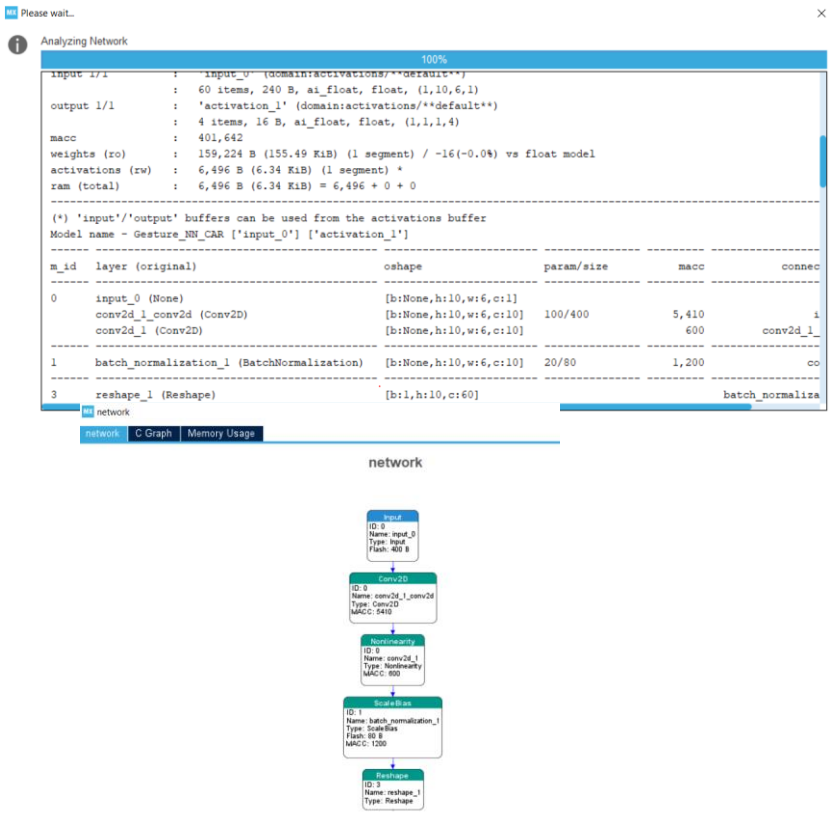




STM32Cube.AI user flow (2/4)

1. Select MCU, add X-CUBE-AI pack
2. Load and analyze NN Model
3. Optimize and validate
4. Generate project and integrate

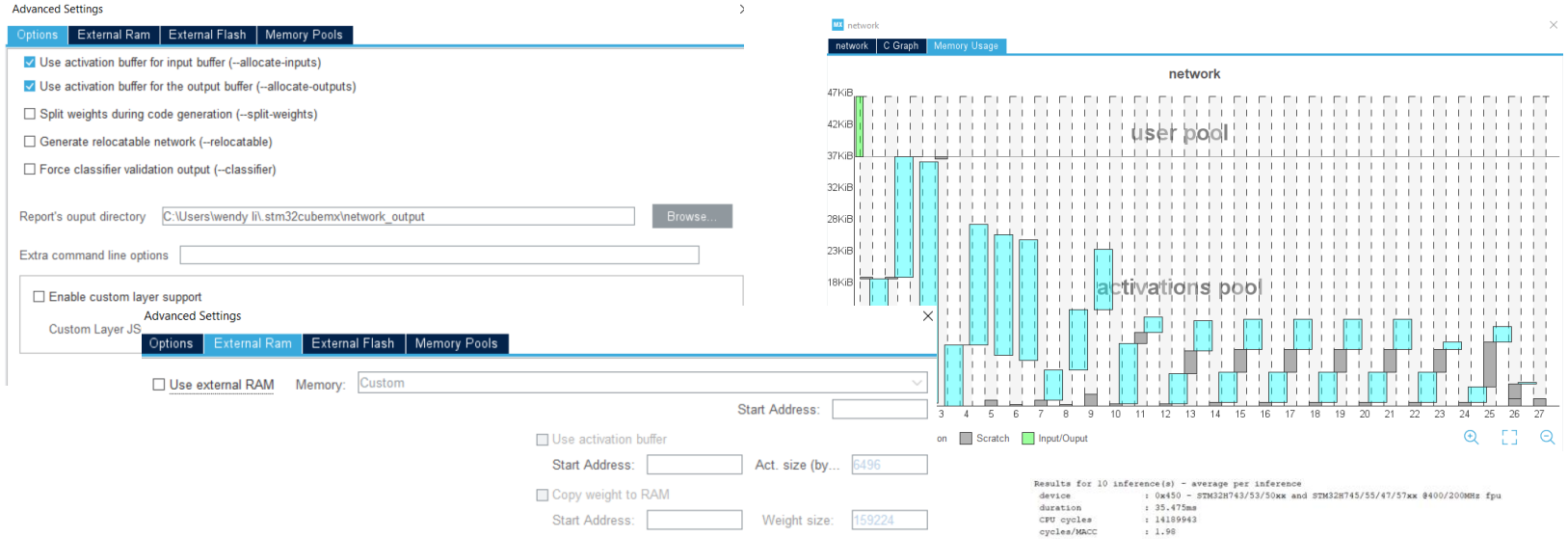
- Load NN model. (h5, tflite, onnx)
- Model complexity and footprint analysis
- Check Flash, Ram usage.



STM32Cube.AI user flow (3/4)

1. Select MCU, add X-CUBE-AI pack
2. Load and analyze NN Model
3. Optimize and validate
4. Generate project and integrate

- Fine tune memory allocation with optimizations and GUI
- Optimize system parameters and clock tree
- Extend model with your own customer layers



- Validate on desktop with your own dataset
- Validate on target and check inference time

Validate on desktop

Validate on target

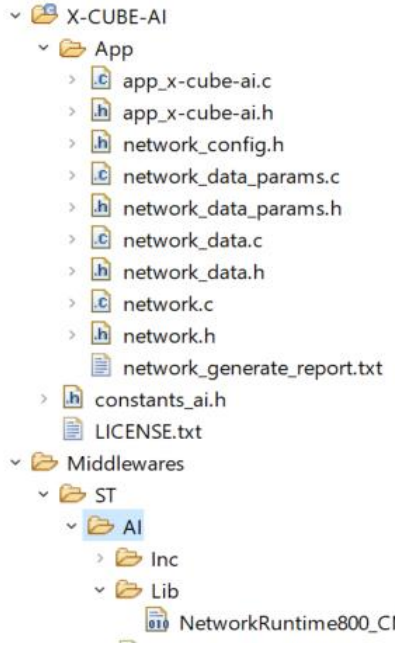
```
Results for 10 inference(s) - average per inference
 device          : 0x450 - STM32H743/53/50xx and STM32H745/55/47/57xx @400/200MHz fpu
 duration        : 35.475ms
 CPU cycles      : 14189943
 cycles/MACC     : 1.98
 c_nodes         : 28
c_id  m_id  desc                 oshape          fmt    ms          %
------------------------------------------------------------------------
0     0     Conv2D (0x103)       (1, 48, 48, 8)  uint8       3.453    9.7%
1     1     Conv2D (0x103)       (1, 48, 48, 8)  uint8       4.312   12.2%
2     2     Conv2D (0x103)       (1, 48, 48, 16) uint8       2.104    5.9%
3     3     Conv2D (0x103)       (1, 24, 24, 16) uint8       1.157    3.3%
4     4     Conv2D (0x103)       (1, 24, 24, 32) uint8       1.250    3.5%
5     5     Conv2D (0x103)       (1, 24, 24, 32) uint8       2.404    6.8%
6     6     Conv2D (0x103)       (1, 24, 24, 32) uint8       1.911    5.4%
7     7     Conv2D (0x103)       (1, 12, 12, 32) uint8       0.526    1.5%
8     8     Conv2D (0x103)       (1, 12, 12, 64) uint8       0.884    2.5%
9     9     Conv2D (0x103)       (1, 12, 12, 64) uint8       1.050    3.0%
10    10    Conv2D (0x103)       (1, 12, 12, 64) uint8       1.544    4.4%
11    11    Conv2D (0x103)       (1, 6, 6, 64)   uint8       0.256    0.7%
12    12    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.793    2.2%
13    13    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.495    1.4%
14    14    Conv2D (0x103)       (1, 6, 6, 128)  uint8       1.453    4.1%
15    15    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.497    1.4%
16    16    Conv2D (0x103)       (1, 6, 6, 128)  uint8       1.454    4.1%
17    17    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.495    1.4%
18    18    Conv2D (0x103)       (1, 6, 6, 128)  uint8       1.453    4.1%
19    19    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.495    1.4%
20    20    Conv2D (0x103)       (1, 6, 6, 128)  uint8       1.454    4.1%
21    21    Conv2D (0x103)       (1, 6, 6, 128)  uint8       0.496    1.4%
22    22    Conv2D (0x103)       (1, 6, 6, 128)  uint8       1.453    4.1%
23    23    Conv2D (0x103)       (1, 3, 3, 128)  uint8       0.134    0.4%
24    24    Conv2D (0x103)       (1, 3, 3, 256)  uint8       0.917    2.6%
25    25    Conv2D (0x103)       (1, 3, 3, 256)  uint8       0.259    0.7%
26    27    Conv2dPool (0x109)   (1, 1, 1, 256)  uint8       2.763    7.8%
27    28    Conv2D (0x103)       (1, 1, 1, 3)    uint8       0.012    0.0%
------------------------------------------------------------------------
                                                             35.475 ms
```

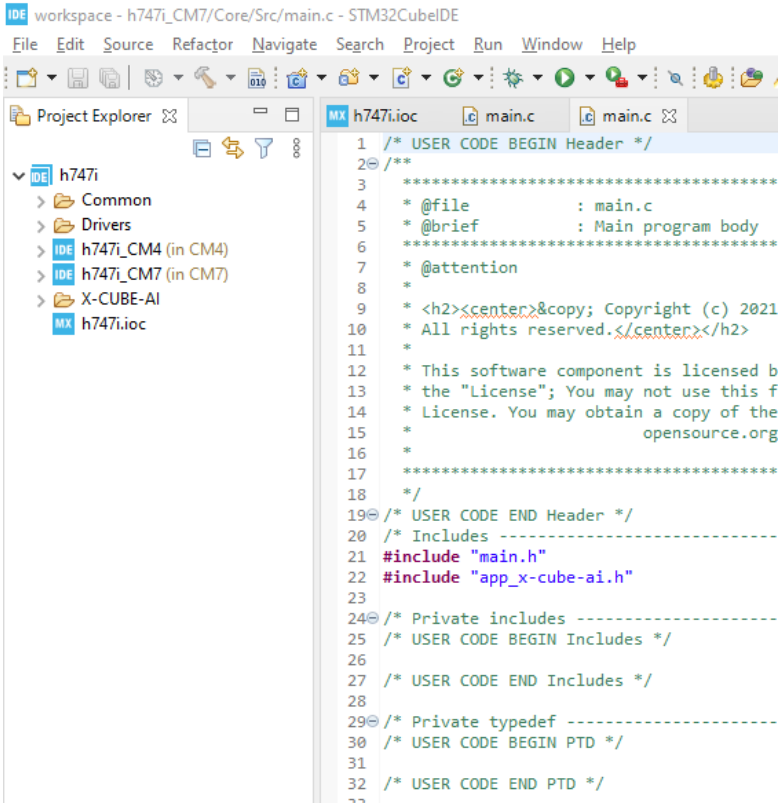
STM32Cube.AI user flow (4/4)

1. Select MCU, add X-CUBE-AI pack
2. Load and analyze NN Model
3. Optimize and validate
4. Generate project and integrate

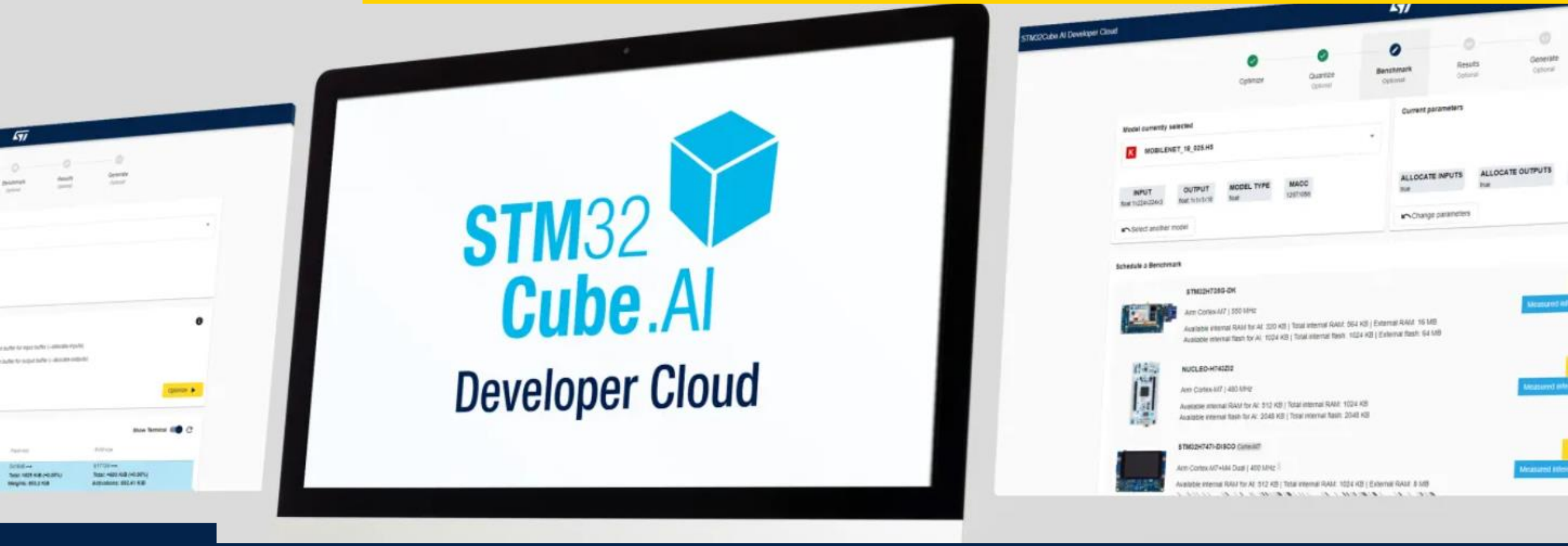
GENERATE CODE

- Generate Application Template
- Integrate with your application-specific code in your favorite IDE
- Perform system tests





STM32Cube.AI Developer Cloud



life.augmented

Cube AI vs Developer Cloud

1. Select MCU, add X-CUBE-AI pack
2. Load and analyze NN Model
3. Optimize and validate
4. Generate project and integrate

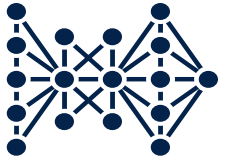
STM32 Cube.AI

1. Load NN Model
2. Optimize NN model
3. Quantize
4. Benchmark & Result
5. Generate project

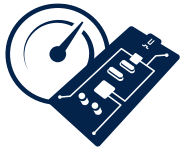
STM32 Cube.AI Developer Cloud

STM32Cube. AI Developer Cloud has these new benefits:

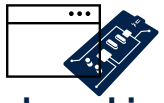
- Model Zoo
- Quantized tool
- Board farm
- Online, No need for tool installation



ST model zoo



Board farm



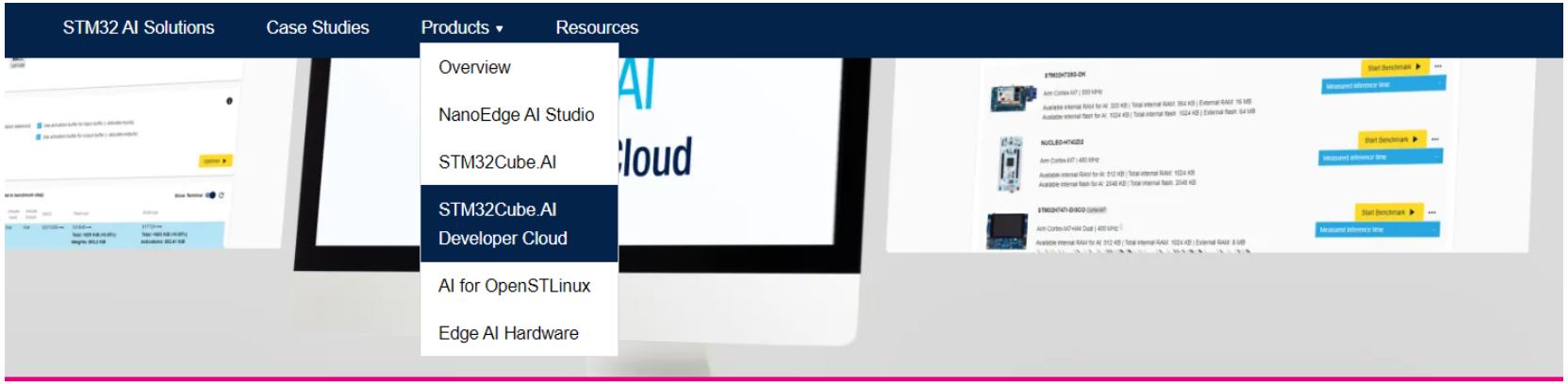
Benchmarking tool

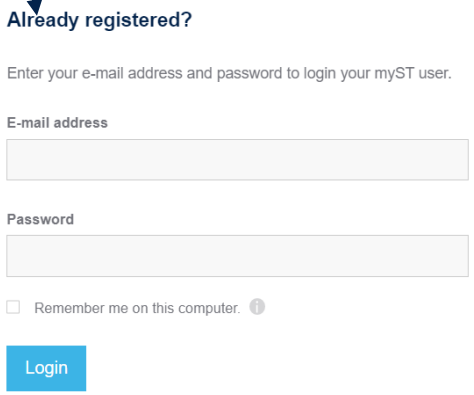
life.augmented

How to Find STM32Cube.AI Developer Cloud?

- Website: https://stm32ai.st.com/stm32-cube-ai-dc/

- Open online platform through ST website.
- Login in with your email, password





Online tool for fast AI optimization and benchmark

Open online platform

STM32Cube.AI Developer Cloud is a free-of-charge online platform and services allowing the creation, optimization, benchmarking, and generation of AI for the STM32 microcontrollers. It is based on the STM32Cube.AI core technology.
Benefits include access to online service to benchmark AI on STM32 boards. It comes with the ST Model zoo including reference model, training script and application examples.

Main features

Online GUI secured by ST user credentials
AI optimizer and visualizer
AI benchmark service on various STM32
STM32 model zoo with training scripts
Application code examples
Model quantization tool

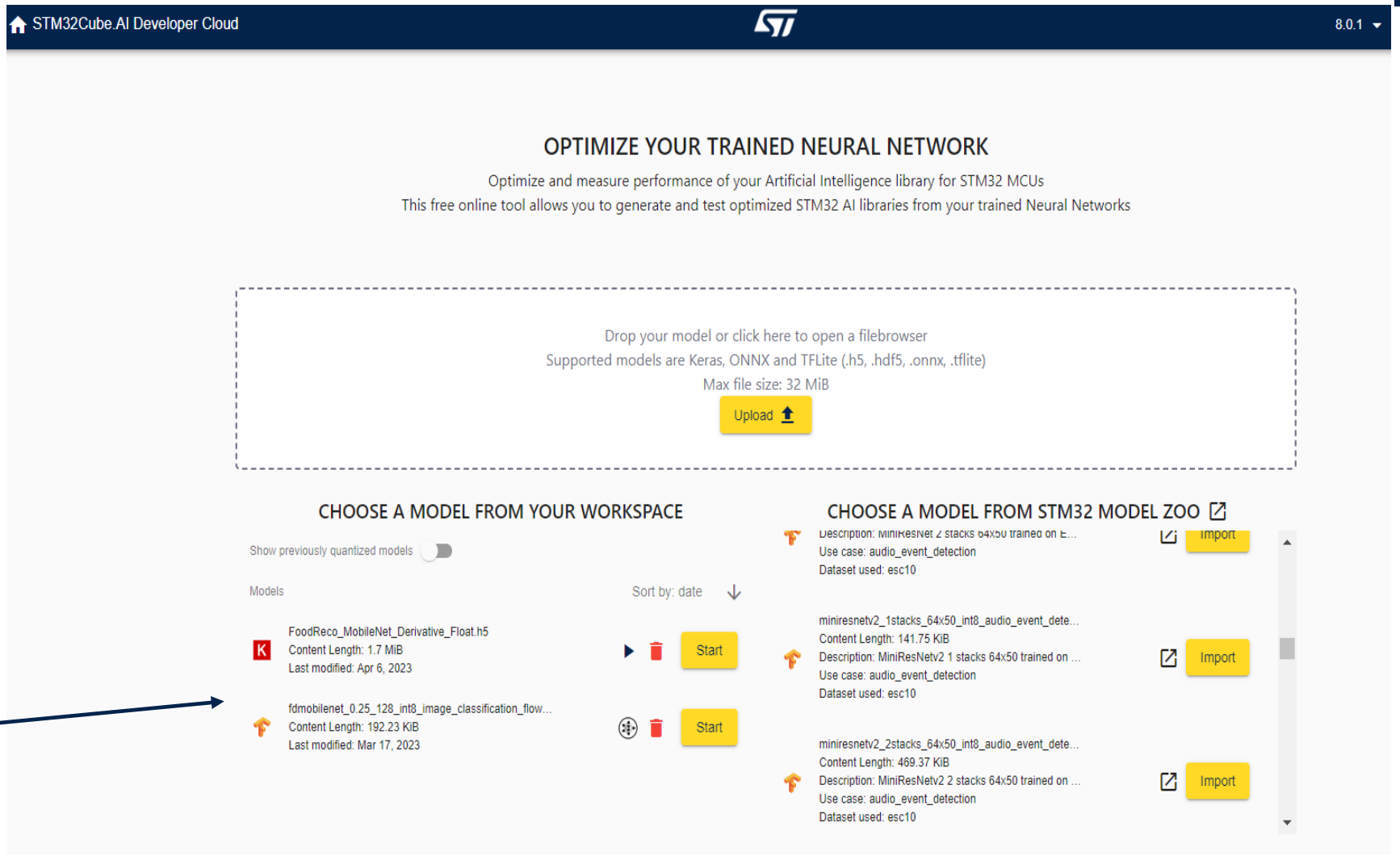
life.augmented

STM32 Cube.AI Developer Cloud

Load NN Model

- Load a model from **ST model zoo** or **own your trained model** from your favorite AI framework: Tensorflow lite, Keras,Pytorch,ONNX,scikit Learn,…
- Supported format are FLOAT32 or quantized INT8

The model you selected or uploaded before

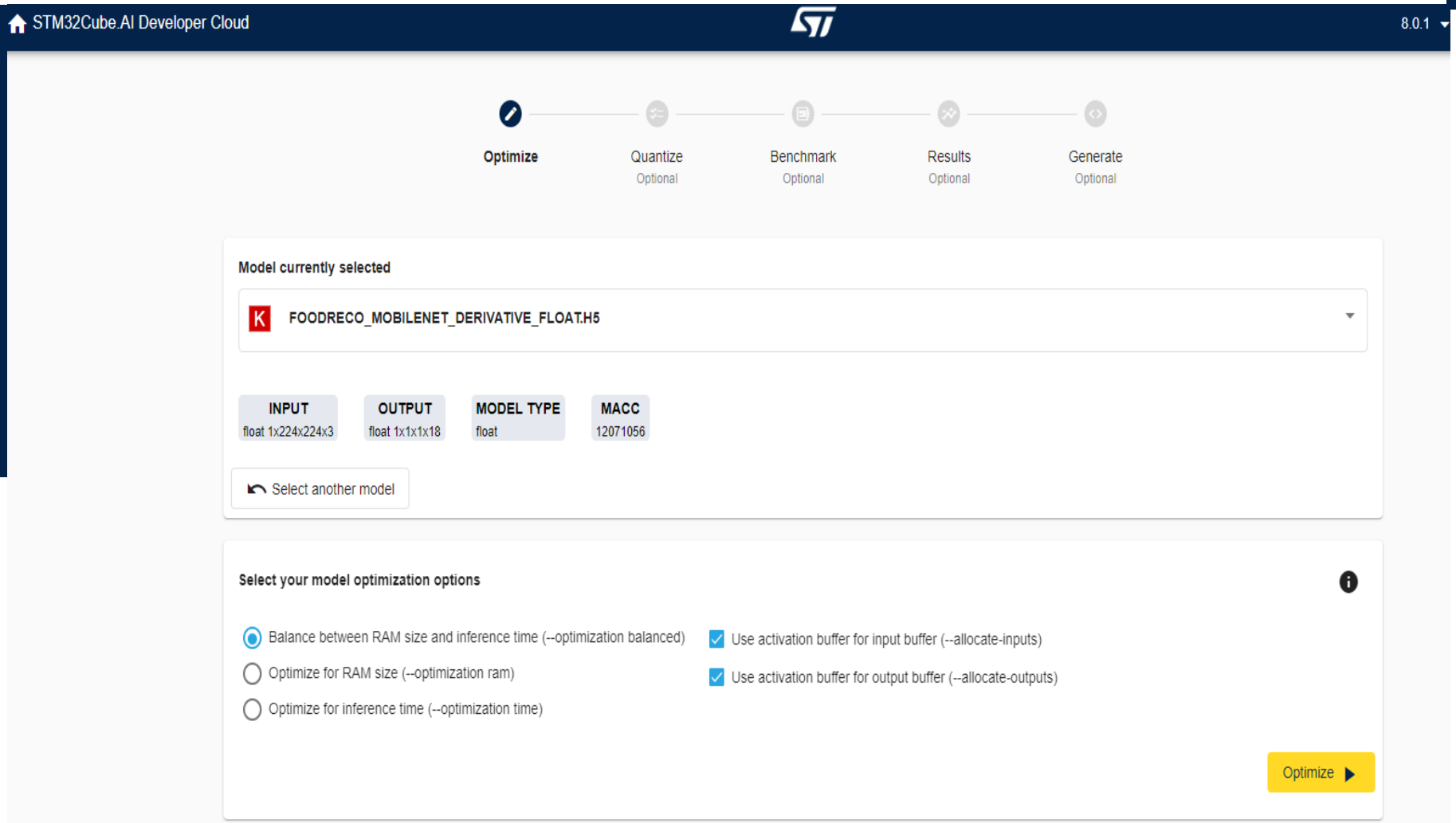


life.augmented

STM32 Cube.AI Developer Cloud

Optimize NN model(1/2)

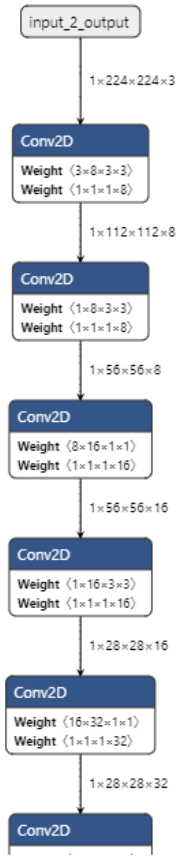
- Select your preferred model optimization options between the Ram or latency most critical parameter.
- Click “Optimize” to check the optimization result



life.augmented

STM32 Cube.AI Developer Cloud

Optimize NN model(2/2)



History of optimization results (inference time will be measured in benchmark step)

Show terminal

Date	Version	Optimization	Allocate Inputs	Allocate Outputs	MACC	Flash size	RAM size
4/6/23, 10:53 AM	8.0.1	time	true	false	12071056	541740	1014084
4/6/23, 10:54 AM Latest	8.0.1	balanced	true	true	12071056	541784 **Total: ≈529 KiB (+0.00%)** **Weights: 503.2 KiB** **Library (estimation) ≈26 KiB**	617120 **Total: ≈603 KiB (+0.00%)** **Activations: 592.41 KiB** **Library (estimation) ≈10 KiB** **Input/Output: 0 B / 0 B**
4/6/23, 10:53 AM	8.0.1	time	true	true	12071056	541784	1014016
4/6/23, 10:53 AM	8.0.1	ram	true	true	12071056	541784	617120
4/6/23, 10:51 AM Default	8.0.1	balanced	true	true	12071056	541784	617120

Show graph | Delete | Compare with default | Go to quantize | Go to benchmark

Items per page: 5 | 1 – 5 of 5

Terminal

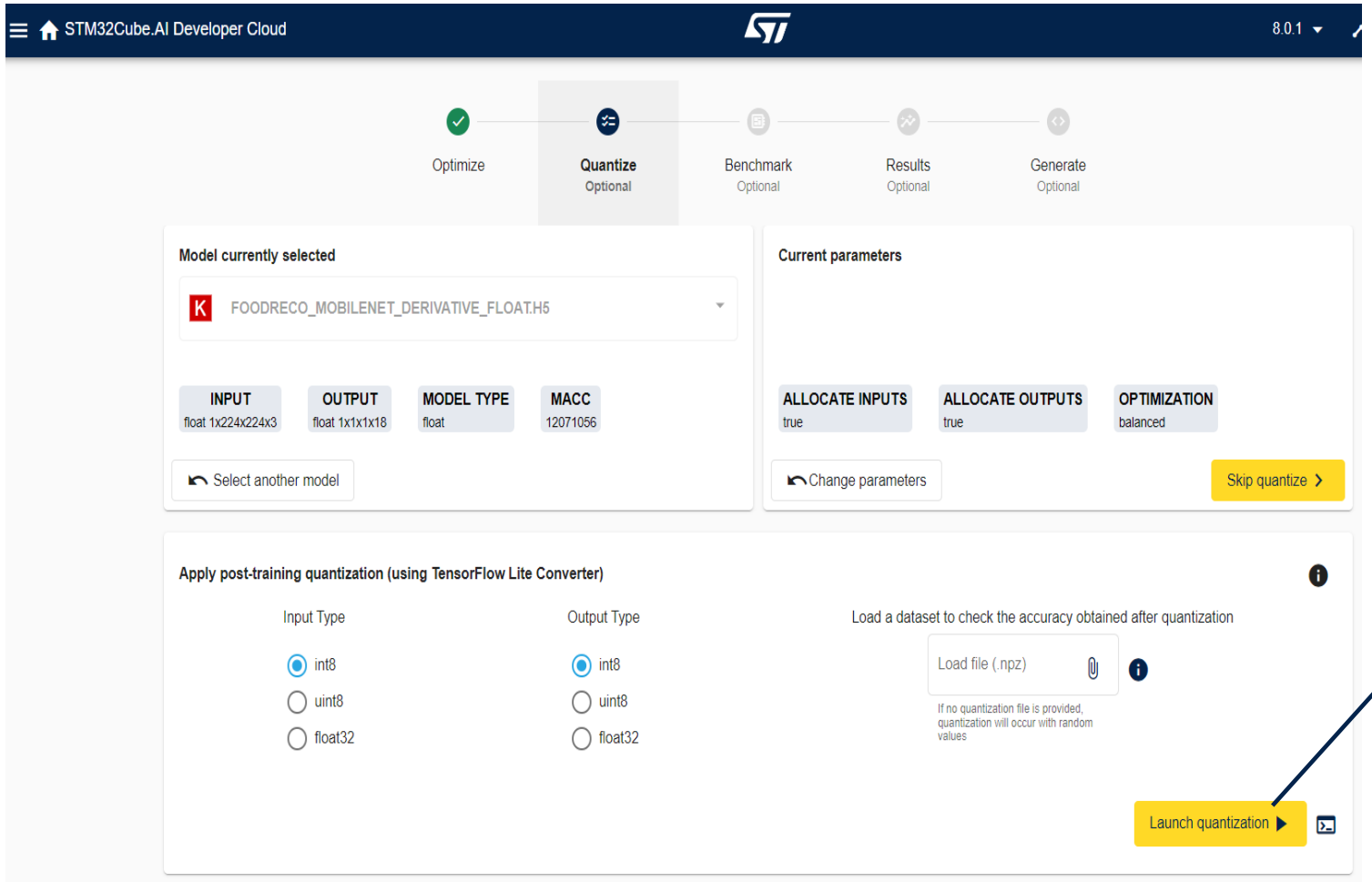
```
>>> stm32ai validate --model FoodReco_MobileNet_Derivative_Float.h5 --workspace workspace --output output --allocate-inputs --allocate-outputs --relo
Neural Network Tools for STM32AI v1.7.0 (STM.ai v8.0.1-19409)
Applied 6 of 7 optimizations
Copying the AI runtime files to the user workspace: workspace

 Exec/report summary (validate)
 ---------------------------------------------------------------------------------------------
 model file         :   FoodReco_MobileNet_Derivative_Float.h5
```

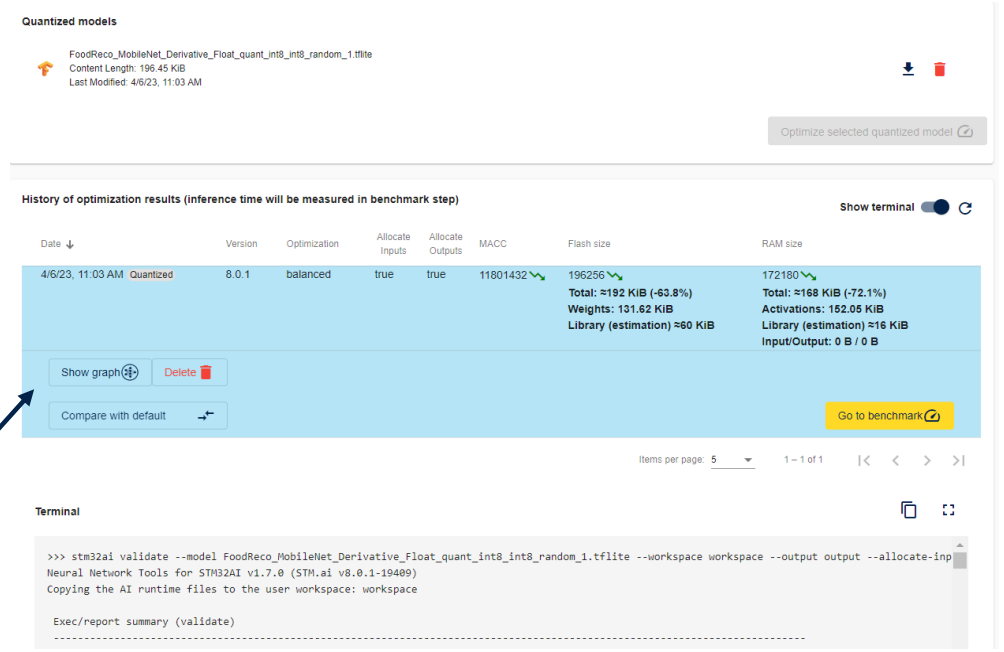
life.augmented

STM32 Cube.AI Developer Cloud

Quantize(optional)

- Perform post-training quantization to further reduce your NN model parameter size

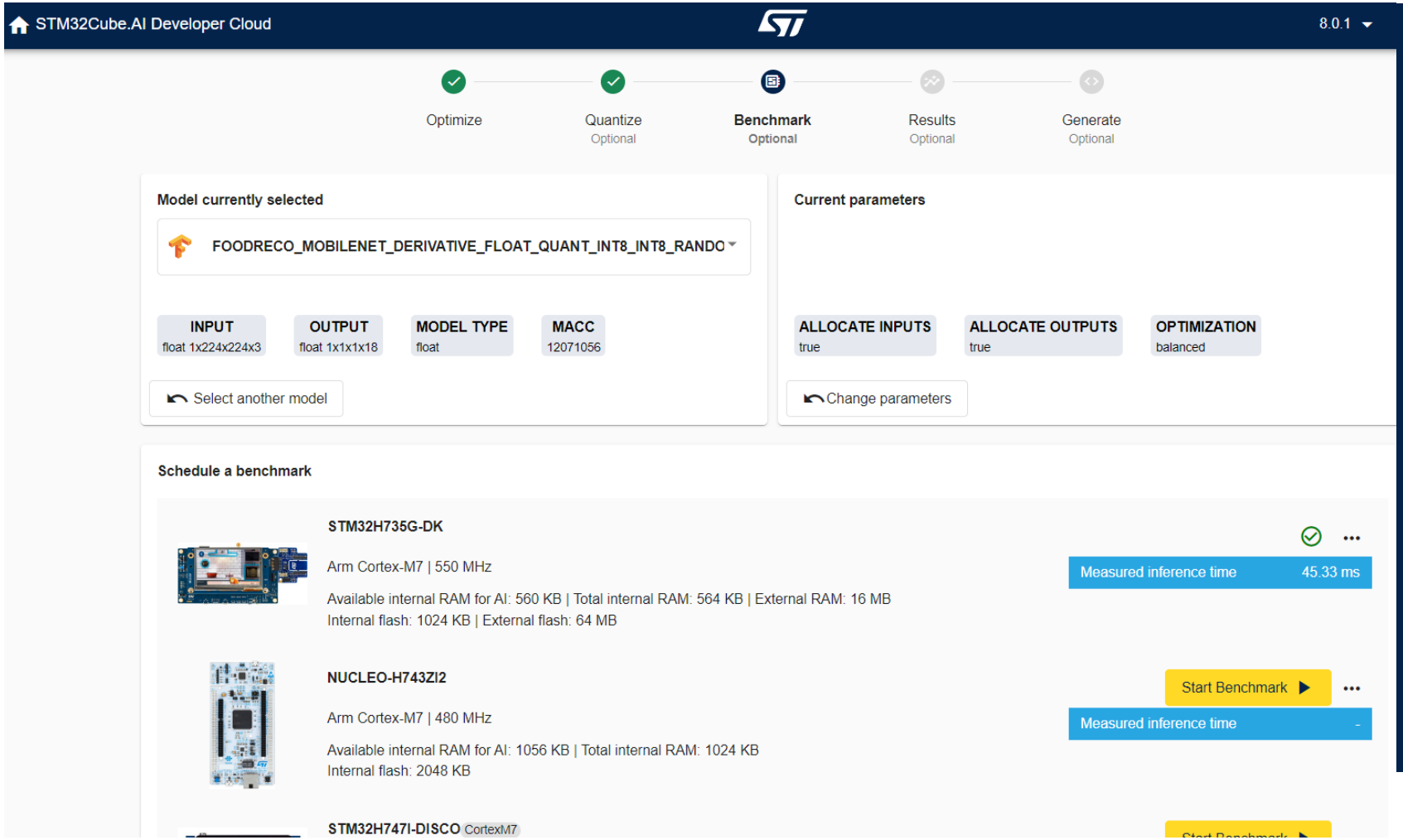




life.augmented

STM32 Cube.AI Developer Cloud

Benchmark (optional)



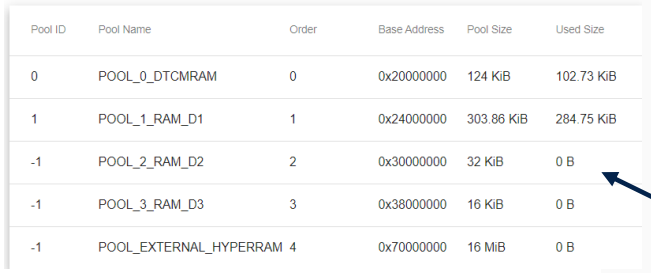
- Access the online benchmark service to remotely evaluate your AI performance on various STM32 boards.
- 9 boards available for checking inference time.
  - STM32H735G-DK
  - NUCLEO-H743ZI2
  - STM32H747I-DISCO
  - STM32H7B3I-DK
  - STM32F769I-DISCO
  - B-U585I-IOT02A
  - NUCLEO-G474RE
  - STM32F469I-DISCO
  - STM32L4R9I-DISCO

life.augmented

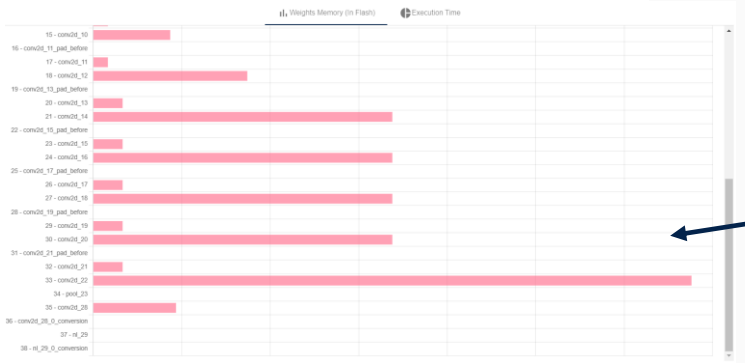
STM32 Cube.AI Developer Cloud

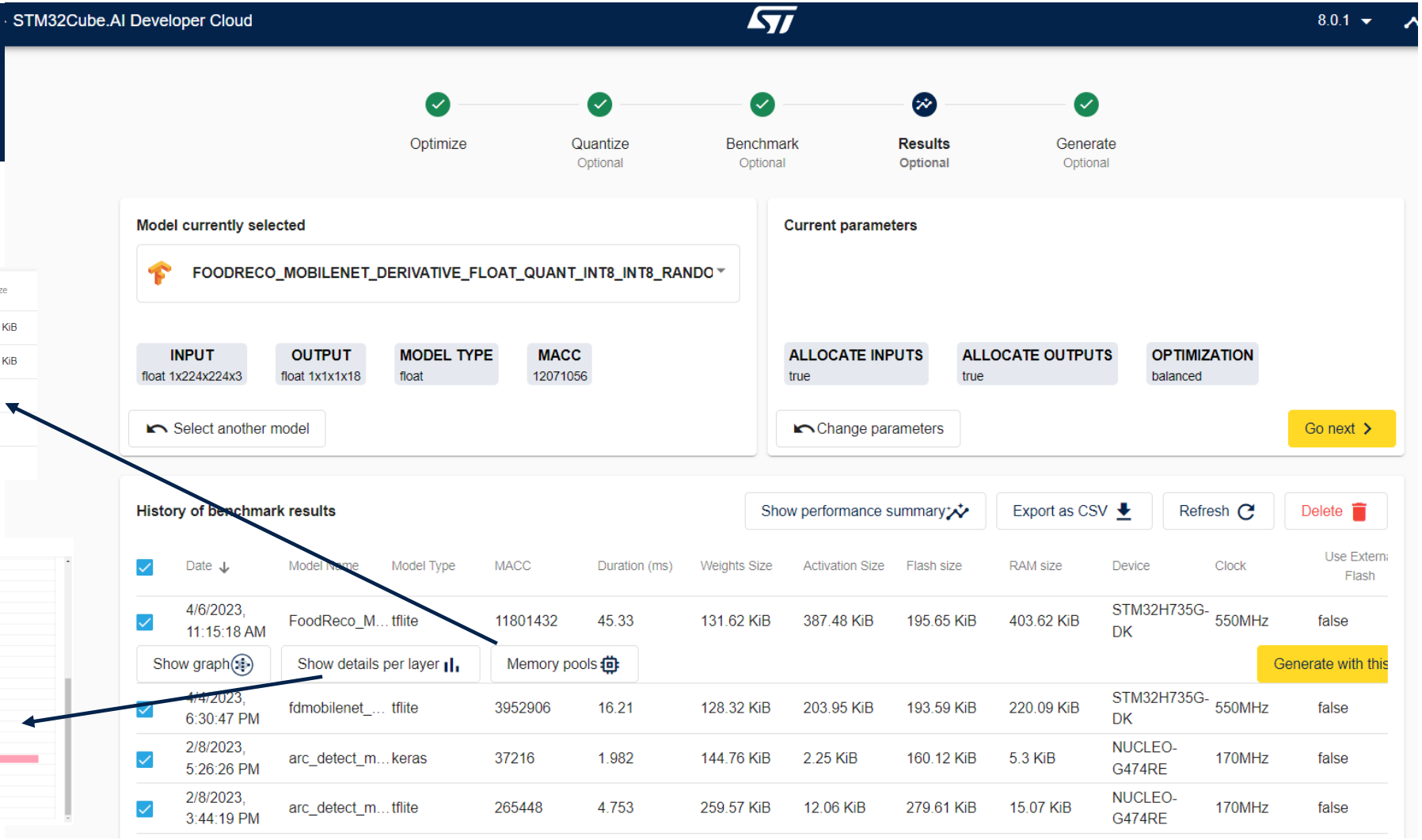
Result (optional)

- Visualize and optionally export the performance benchmark report



Pool ID	Pool Name	Order	Base Address	Pool Size	Used Size
0	POOL_0_DTCMRAM	0	0x20000000	124 KiB	102.73 KiB
1	POOL_1_RAM_D1	1	0x24000000	303.86 KiB	284.75 KiB
-1	POOL_2_RAM_D2	2	0x30000000	32 KiB	0 B
-1	POOL_3_RAM_D3	3	0x38000000	16 KiB	0 B
-1	POOL_EXTERNAL_HYPERRAM	4	0x70000000	16 MiB	0 B





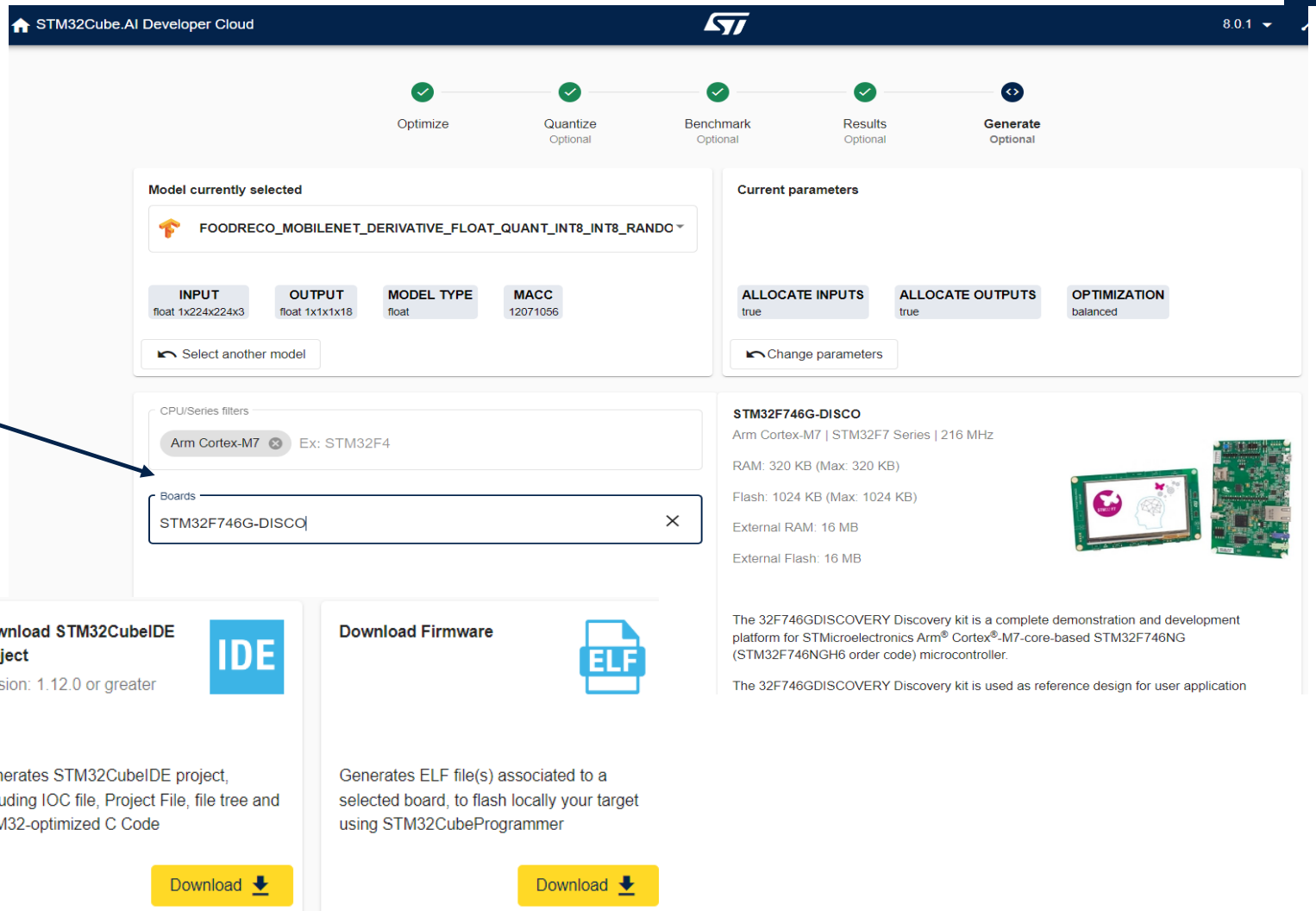
life.augmented

STM32 Cube.AI Developer Cloud

Generate (optional)

- Generate your optimized AI code in C-code or in a format compatible with the STM32CubeMX software suites.

Select CPU and boards you are using



life.augmented

Don't go alone

We have created a network of companies to support you

Trust our **authorized partners** to ensure the success of your project. Learn more at **st.com/stm32ai**

Wish to discuss a co-development partnership for ML/AI projects? Contact us at edge.ai@st.com

life.augmented

life.augmented

Please Scan the QR Codes and Stay Tuned with Us.

PDSA Wechat Subscription

Power & SPIN Microsite

Our technology starts with You

Find out more at www.st.com

life.augmented