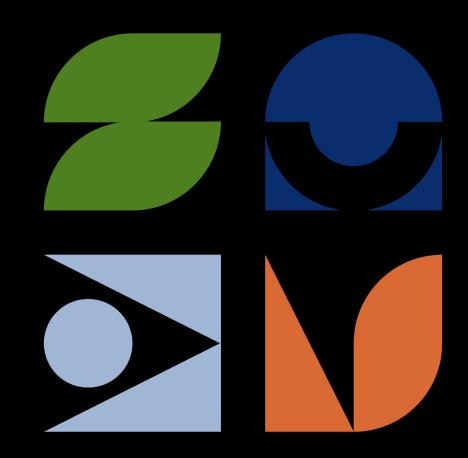
HELMHOLTZ

MODEL ZOO

Hans Werners, Jennifer Ahrens,

Engin Eren, Philipp Heuser







Helmholtz Model Zoo

A Model Zoo is a collection of (pre)trained deep learning models, ready to use for inference.

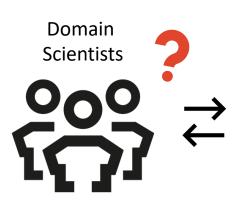
These can include:

- specifically trained and fine-tuned models for defined tasks and data,
- domain and data type agnostic generalised models
- foundation models
- Your models!









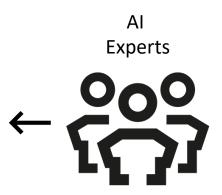
mamba slurm
dependencies

docker permission denied

pip cluster pull

versions install GPU

deploy



provide models in git



HELMHOLTZ MODEL ZOO

Domain Scientists



REST API

Webfrontend



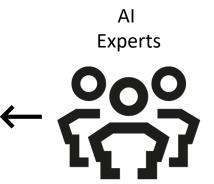


Trained
Deep Learning
Models
ready to use on
Haicore GPU nodes
@DESY



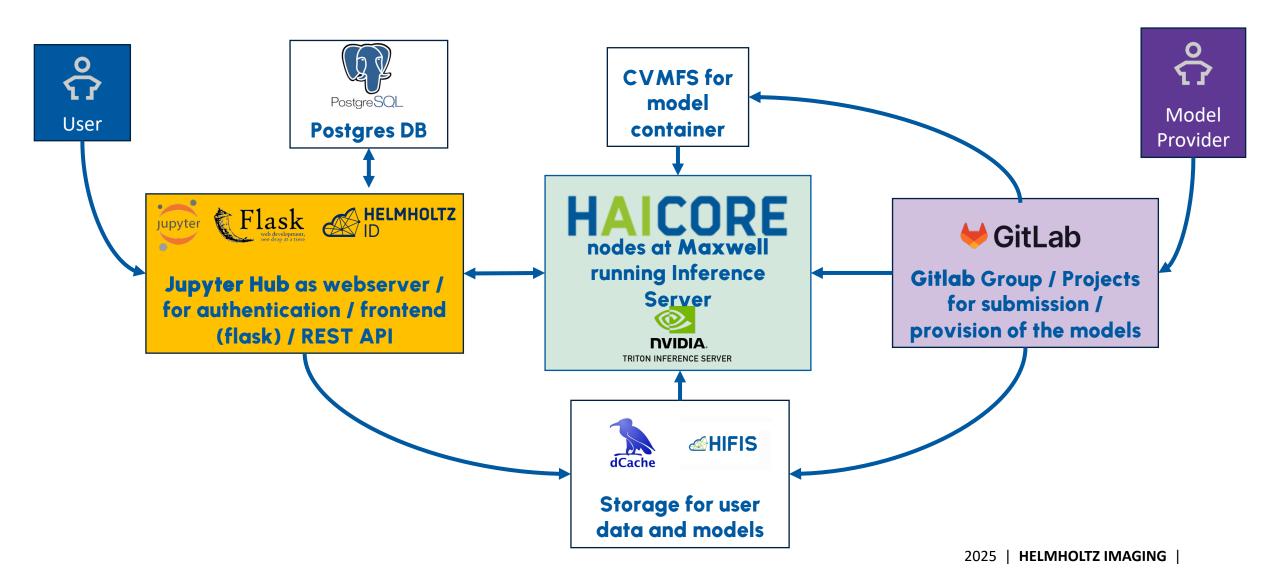


tests, generates and provides provide models and python code in gitlab project for IO, pre- and postprocessing



Helmholtz Model Zoo





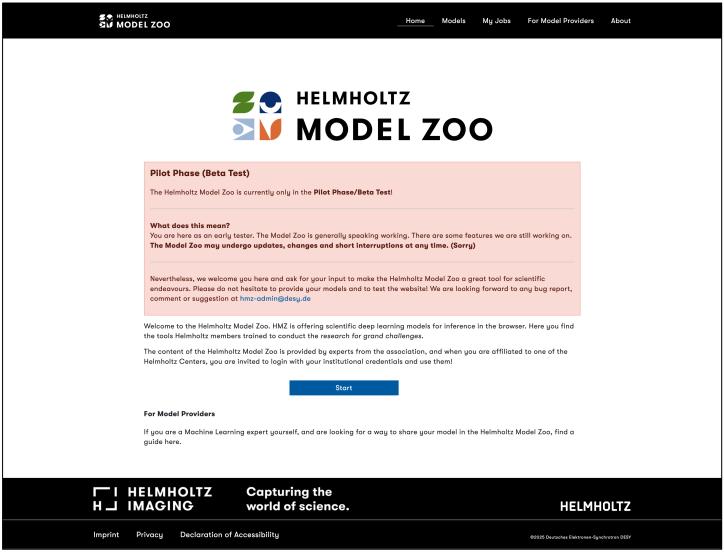
Why a "Helmholtz" Model Zoo?



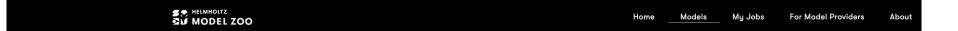
Can't you just use xxx?

- the data stays within Helmholtz
- flexible data-limit
- "private" models possible
- flexible usage (WebUI, REST-API, REST-API raw)
- free to use for Helmholtz
- support
- easy to use
- no installation hassle, no GPU hunting, no IT knowledge necessary



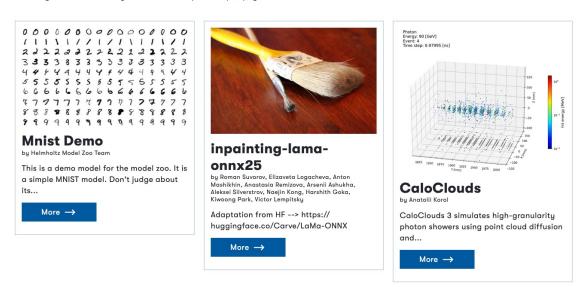






Model overview

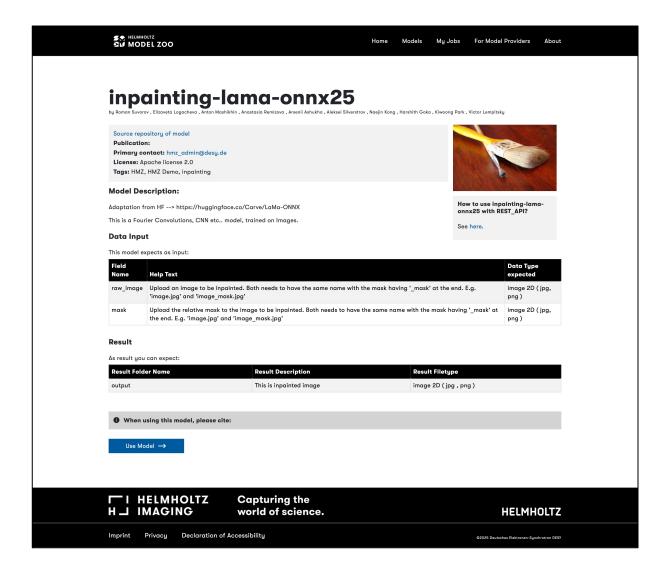
Selecting a model redirects you to the model-specific input-page.



☐ I HELMHOLTZ H → IMAGING Capturing the world of science.

HELMHOLTZ

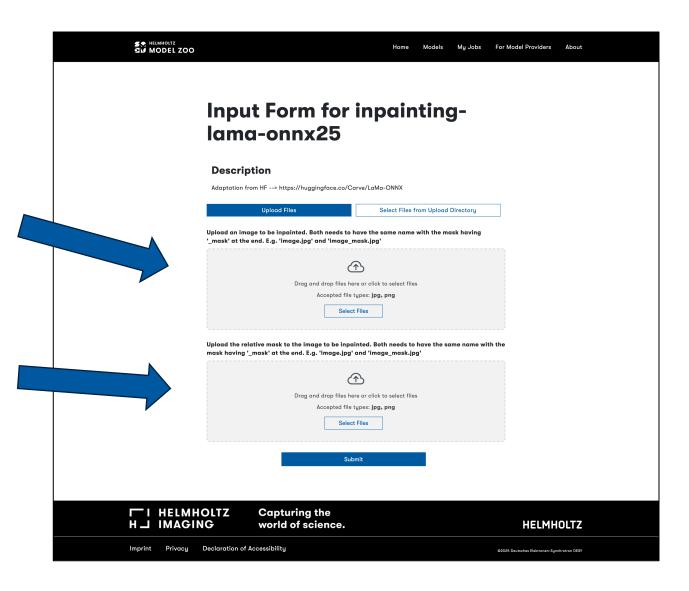




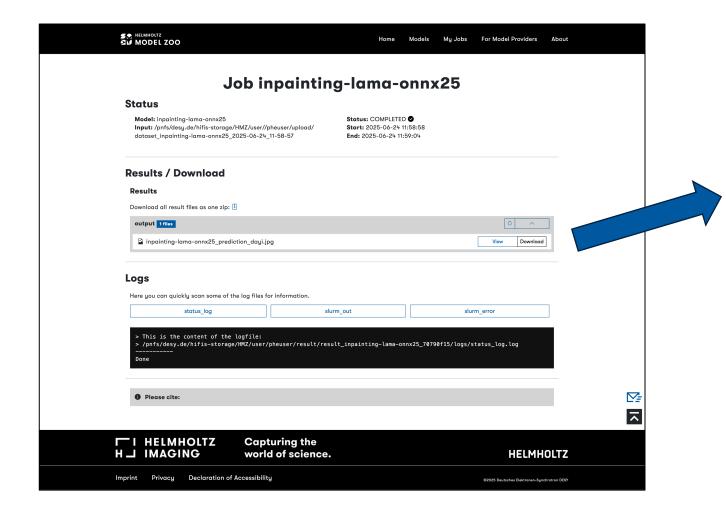


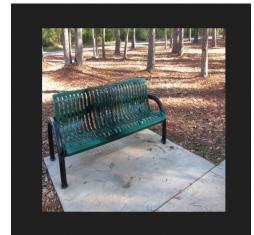




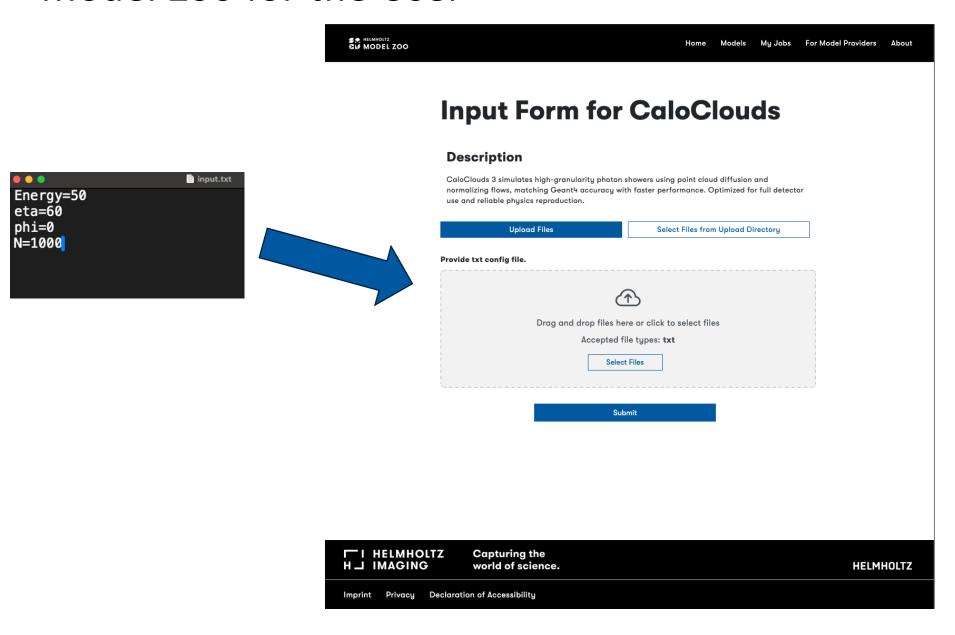




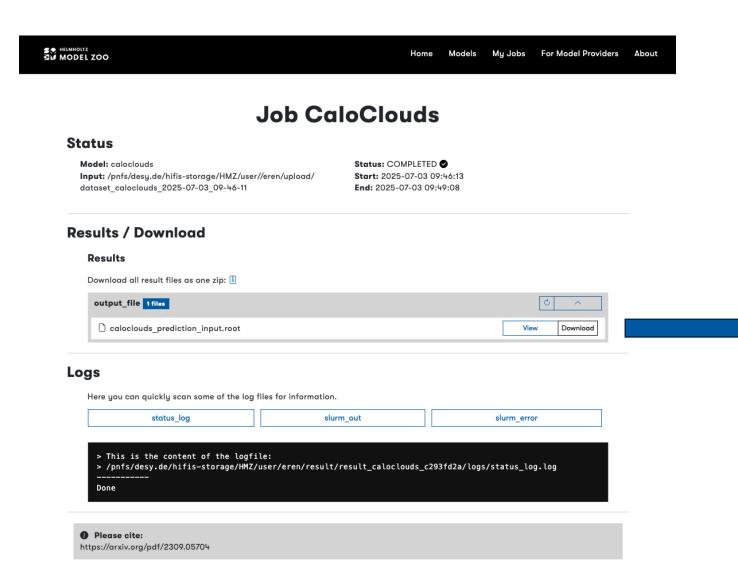






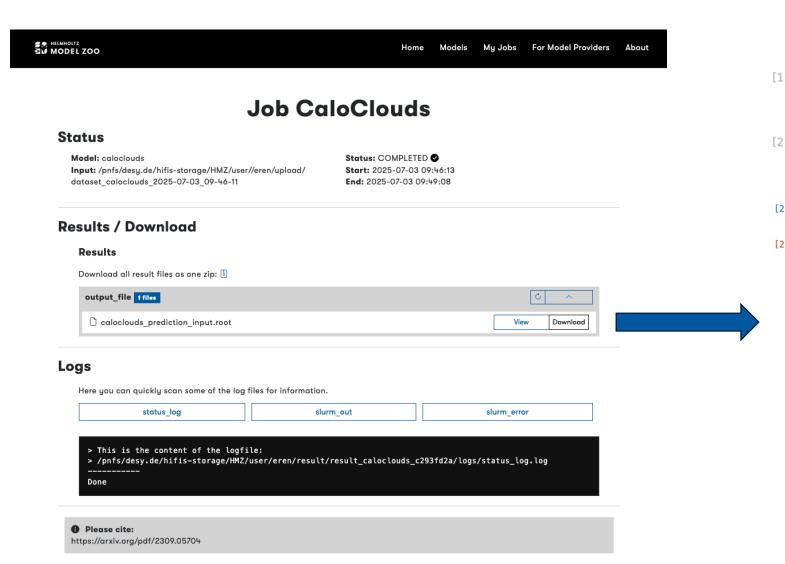






```
import uproot
     file = uproot.open("caloclouds_prediction_input.root")
[2]: print(file.keys())
     ['Events;1']
[13]: data[2:10]
[13]: [{noutput: 8330, output: [6, 13, ..., 9, 0.287]},
      {noutput: 8882, output: [9, 19, ..., 3, 0.166]},
      {noutput: 8426, output: [24, 48, ..., 11, 0.382]},
      {noutput: 8990, output: [8, 33, ..., 6, 0.253]},
      {noutput: 8394, output: [10, 12, ..., 7, 0.579]},
      {noutput: 8722, output: [11, 12, ..., 10, 0.343]},
      {noutput: 8470, output: [15, 31, ..., 9, 0.377]},
      {noutput: 8614, output: [7, 23, ..., 9, 0.0446]}]
     backend: cpu
     nbytes: 1.0 MB
     type: 8 * {
         noutput: int32,
         output: var * float64
```





```
import uproot
     file = uproot.open("caloclouds_prediction_input.root")
[2]: print(file.keys())
      ['Events;1']
[22]: import matplotlib.pyplot as plt
      plt.scatter(x,z,s=3)
[22]: <matplotlib.collections.PathCollection at 0x148de86817d0>
      25
      20
       15
       10
       5 -
        -50
                         50
                                100
                                        150
                                                200
                                                        250
                                                                300
```

FAQ user



Who can use the models?

Only colleagues from Helmholtz can use it (and members of a Helmholtz ID VO)

How can I run an inference?

- Use the webinterface.
- Use a REST API

What is about my data (user)?

- You upload the data to a folder owned by you in dcache via the webinterface.
- You own your data, and it does not leave Helmholtz.
- You can also use the documented ways to upload data to dcache via rclone, webdav, etc.
 (recommended for large datasets)
- You can upload a reasonable amount of data (probably not PB, but a few TB could be fine).
- We will delete all your data after 3 month, or so? (Not entirely defined yet)

For model providers

Who can provide models?

All scientists in Helmholtz can login via the Helmholtz ID and provide their models.

What kind of models? 💮 ONNX 👸 PyTorch 🔭 OpenVINO 🔁











Which hardware?

Our nodes are provided by HAICORE, run at DESY, each equipped with 4 Nvidia L40S.



Can I make my model "private"?

Yes, you will be able to give the usernames of people who are entitled to use your model? (Later via a Helmholtz ID VO)

What do I have to do?

- We provide you with a template in the (DESY)-Gitlab.
- We try to get as much as possible of your shoulders, keeping a maximum of flexibility
- You have to add your model and implement the necessary for pre- and postprocessing
- You have to fill in a form for the necessary meta data (some technical details, description, publication)

Models / config.pbtxt



Before a model can be served by Triton Inference Server, it must be converted into a backend-specific format that Triton can load and run. This step is called exporting the model.

```
import torch
import torch.nn as nn
class MyModel(nn.Module):
    def __init__(self):
        super().__init__()
       self.fc = nn.Linear(10, 5)
   def forward(self, x):
        return self.fc(x)
model = MyModel()
example_input = torch.randn(1, 10)
# Scripted (preferred for dynamic models)
scripted_model = torch.jit.script(model)
scripted_model.save("model/1/your_model_name_model.pt")
```

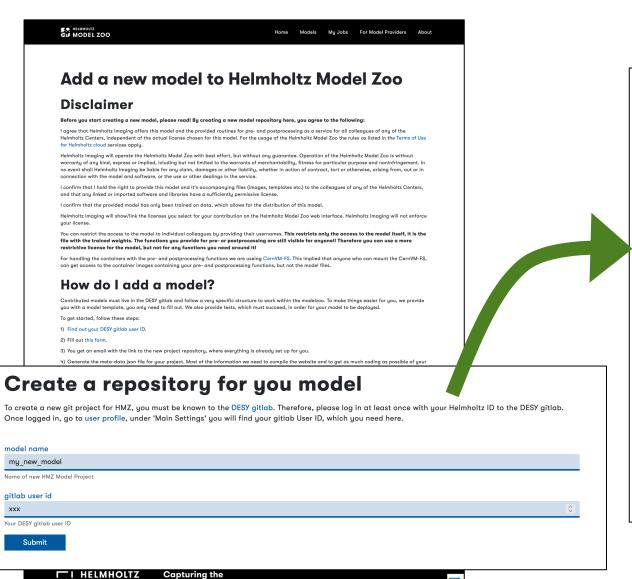
Models / config.pbtxt



```
name: "Hans_mnist"
platform: "pytorch_libtorch"
max_batch_size : 0
input [
   name: "input__0"
    data_type: TYPE_FP32
    dims: [ 1, 28, 28 ]
   reshape { shape: [ 1, 1, 28, 28 ] }
output [
   name: "output__0"
    data_type: TYPE_FP32
    dims: [ 1, 10]
   reshape { shape: [ 1, 10 ] }
```

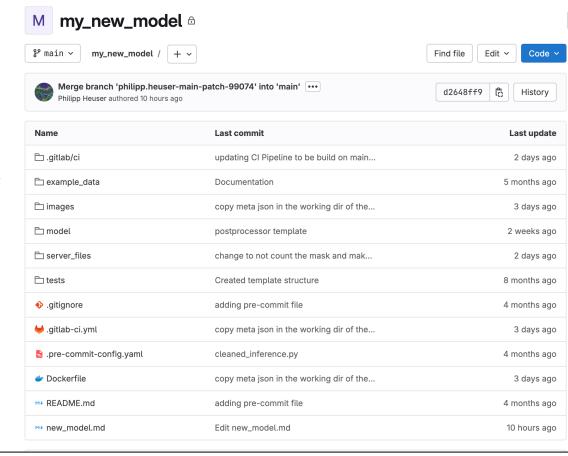
* a draft for the config.pbtxt





world of science.

Imprint Privacy Declaration of Accessibility



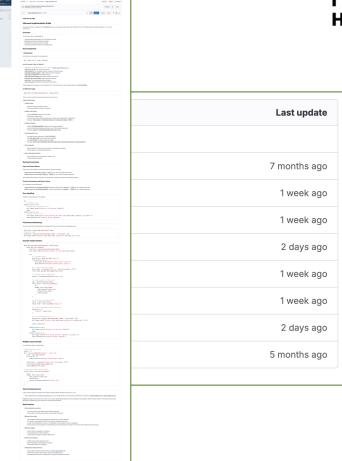
Next time...

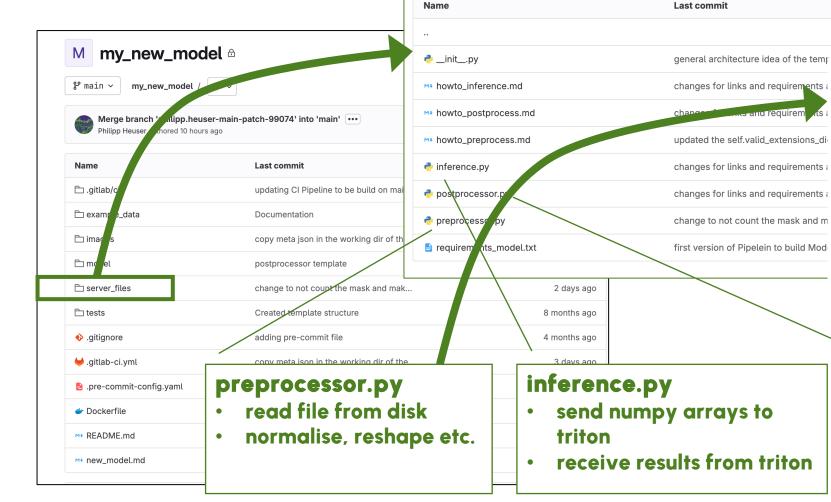


More to show!!!









postprocess.py

- transform model output to human readable
- save files to disk.

HMZ!



Pilot Phase (Beta Test)

The Helmholtz Model Zoo is currently only in the **Pilot Phase/Beta Test!**

What does this mean?

You are here as an early tester. The Model Zoo is generally speaking working. There are some features we are still working on. The Model Zoo may undergo updates, changes and short interruptions at any time. (Sorry)

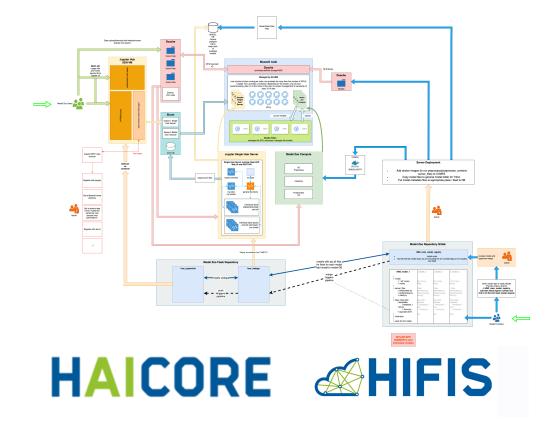
Nevertheless, we welcome you here and ask for your input to make the Helmholtz Model Zoo a great tool for scientific endeavours. Please do not hesitate to provide your models and to test the website! We are looking forward to any bug report, comment or suggestion at hmz-admin@desy.de

Preliminary URL: https://hmz-hub.desy.de

Get in touch hmz-admin@desy.de & talk to us here

Helmholtz Model Zoo





TI HELMHOLTZ
H J IMAGING

Thanks to:

Christian Voss Elena Gapon Frank Schlünzen **Johannes Reppin Kars Ohrenberg** Lusine Yakovleva **Moritz Hamminger Patrick Fuhrmann** Sven Sternberger **Thomas Hartmann** Tim Wetzel **Uwe Jandt** Yves Kemp **Alexander Trautsch** Stefan Bujack

Talk to us!



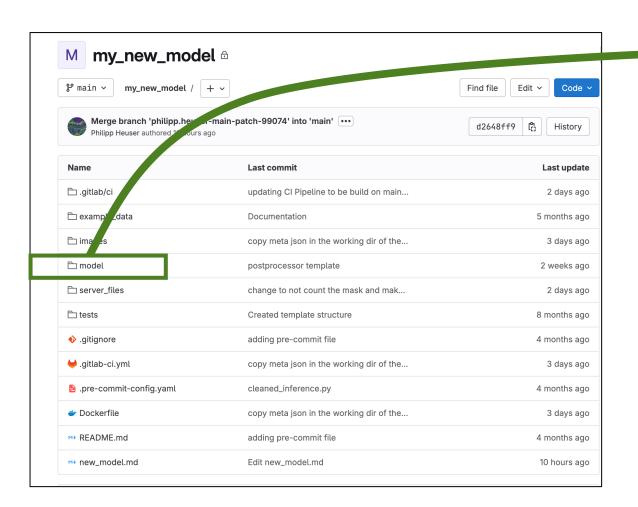






Г Н_

model folder

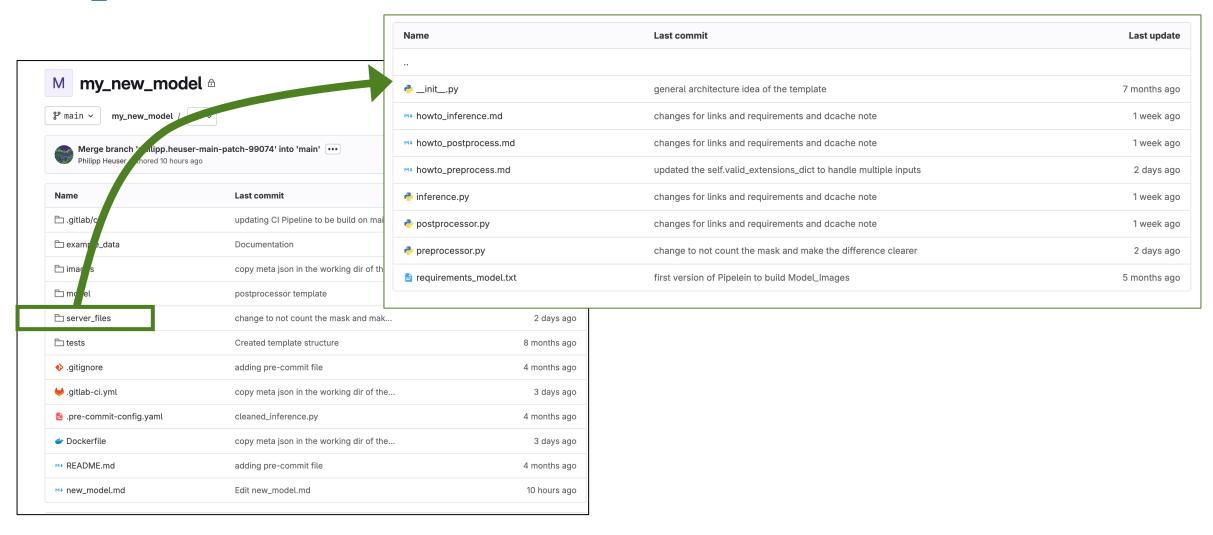


model.pt/onnx/....

config.pbtxt

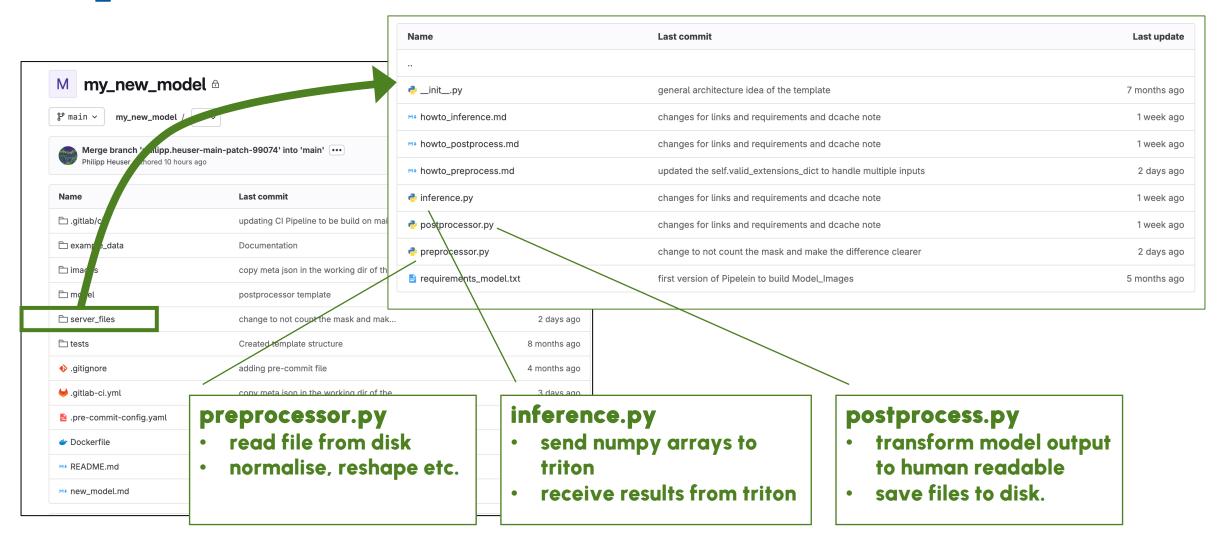


server_files folder

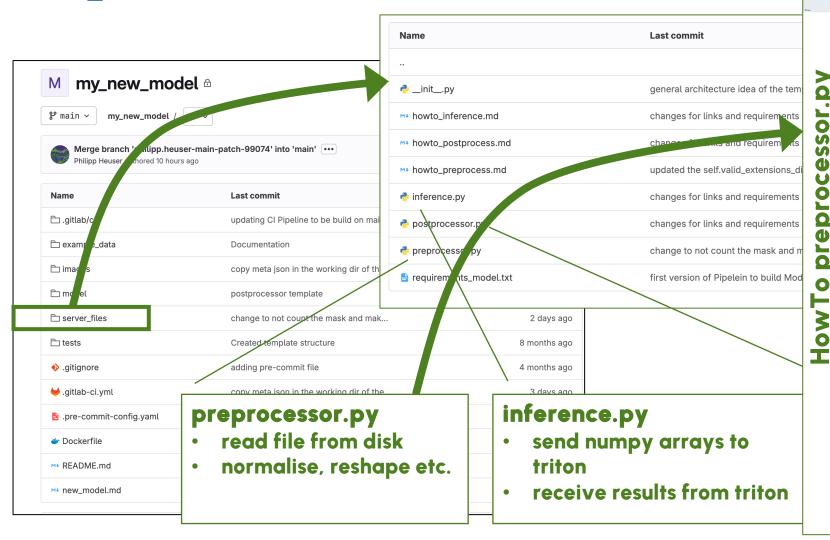


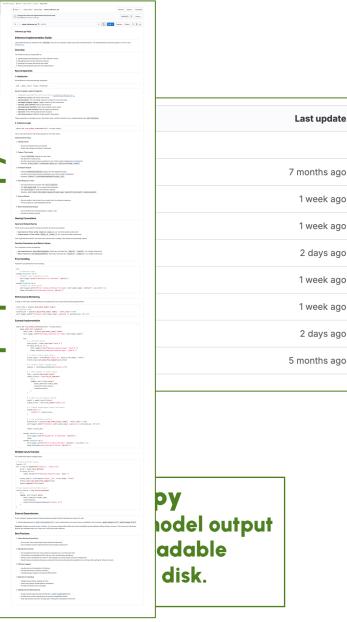


server_files folder



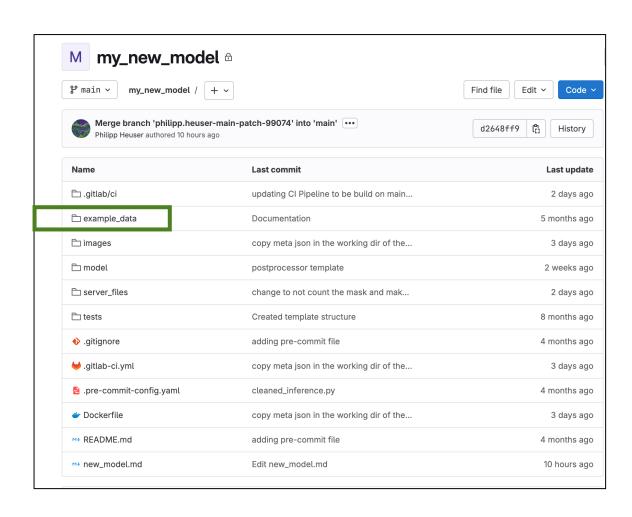
server_files folder





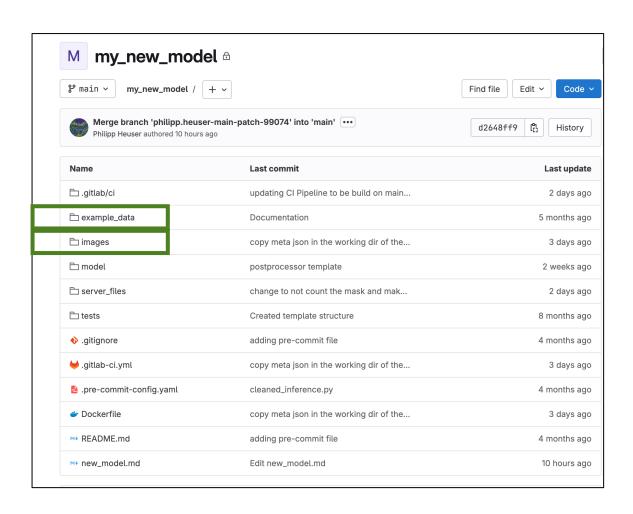






example data:
data to test the model

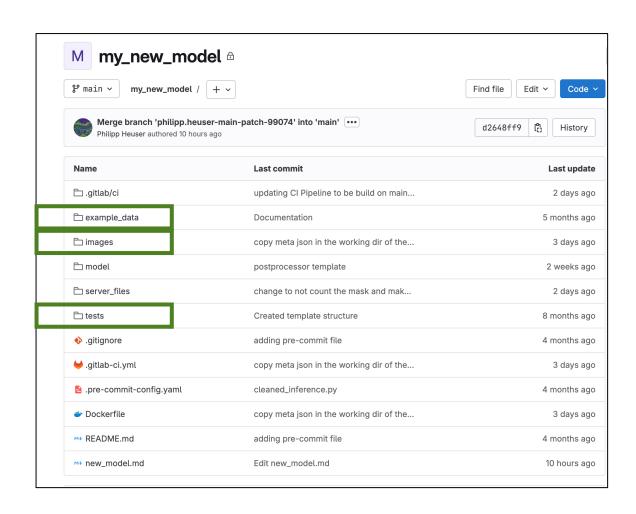




example data:
data to test the model

images:
image(s) for illustration of website





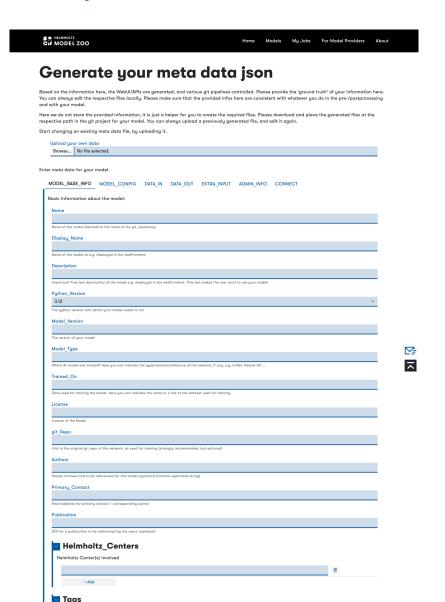
example data:
data to test the model

images:
image(s) for illustration of website

tests: test files to check your code.



meta.json



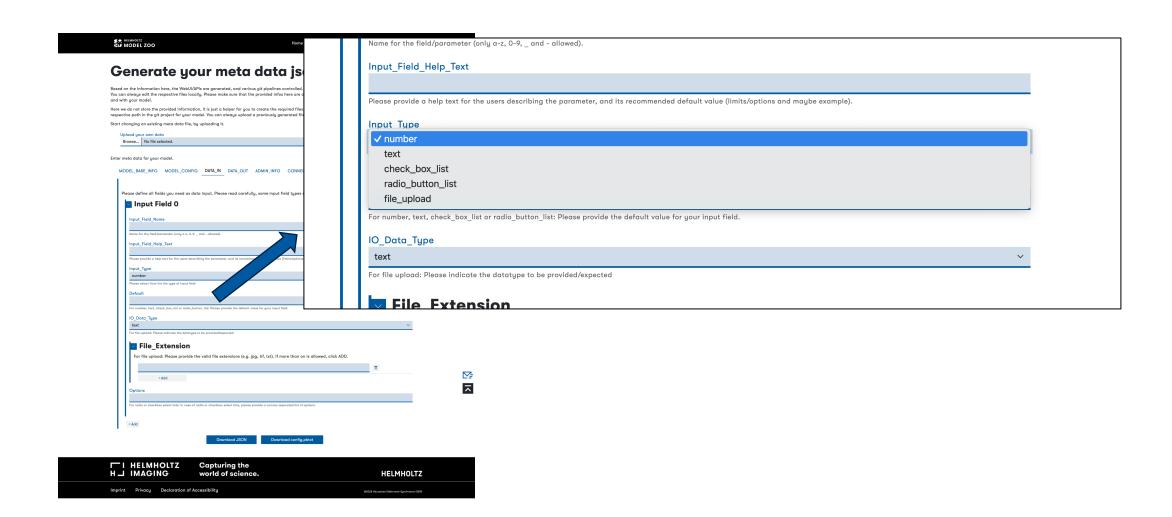
The meta.json is the heart of the HMZ,

- here we collect the information to render the website.
- to define the model submission.
- to create folders for IO.
- · etc.

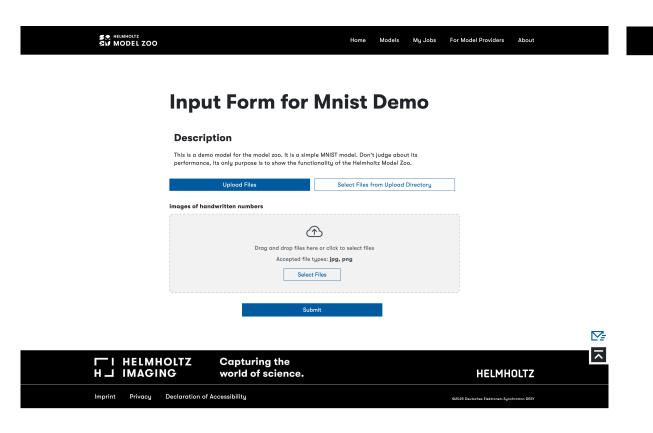
Please use the form on the model zoo website to create the *meta.json* and place it on the toplevel of your model repo.

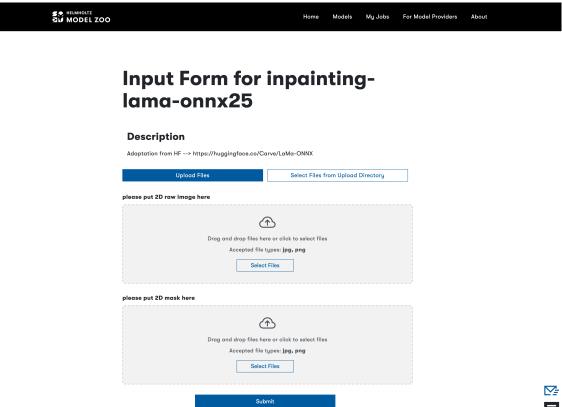
How to input web-interface













Models / config.pbtxt



Before a model can be served by Triton Inference Server, it must be converted into a backend-specific format that Triton can load and run. This step is called exporting the model.

```
import torch
import torch.nn as nn
class MyModel(nn.Module):
    def __init__(self):
        super().__init__()
       self.fc = nn.Linear(10, 5)
   def forward(self, x):
        return self.fc(x)
model = MyModel()
example_input = torch.randn(1, 10)
# Scripted (preferred for dynamic models)
scripted_model = torch.jit.script(model)
scripted_model.save("model/1/your_model_name_model.pt")
```

Models / config.pbtxt



```
name: "Hans_mnist"
platform: "pytorch_libtorch"
max_batch_size : 0
input [
   name: "input__0"
    data_type: TYPE_FP32
    dims: [ 1, 28, 28 ]
   reshape { shape: [ 1, 1, 28, 28 ] }
output [
   name: "output__0"
    data_type: TYPE_FP32
    dims: [ 1, 10]
    reshape { shape: [ 1, 10 ] }
```

* a draft for the config.pbtxt can be downloaded from the meta.json form as well



Data 10

Preprocessing (CPU)

normalisation, slicing, reshaping...



Triton

model inference



Postprocessing (CPU)

reconstruction of image probabilities to human readables

•••



Data 10

Preprocessing (CPU)

normalisation, slicing, reshaping...



Triton model inference



Postprocessing (CPU)

reconstruction of image probabilities to human readables

•••

Data 10

Data 10



Preprocessing (GPU)

normalisation, slicing, reshaping...

Triton

model inference

Postprocessing (GPU)

reconstruction of image probabilities to human readables

•••





Data 10

Preprocessing (CPU)

normalisation, slicing, reshaping...



Triton model inference



Postprocessing (CPU)

reconstruction of image probabilities to human readables

•••

Data 10

Data 10



Preprocessing (GPU)

normalisation, slicing, reshaping...

Triton

model inference

Postprocessing (GPU)

reconstruction of image probabilities to human readables

•••



Data 10

Data 10



Preprocessing (GPU)





Triton model 1

Triton model 2





Postprocessing (GPU)





Data 10

Preprocessing (CPU)

normalisation, slicing, reshaping...



Triton

model inference



Postprocessing (CPU)

reconstruction of image probabilities to human readables

•••

Data 10

Data 10



Preprocessing (GPU)

normalisation, slicing, reshaping...

Triton

model inference

Postprocessing (GPU)

reconstruction of image probabilities to human readables

•••



Data 10

Data 10



Preprocessing (GPU)



Triton model 1



Triton





Postprocessing (GPU)



ensemble config.pbtxt

```
name: "ensemble_model"
platform: "ensemble"
input [
    name: "input_volume"
    data_type: TYPE_FP32
    dims: [ 3, -1, -1 ]
output [
    name: "output_volume"
    data_type: TYPE_FP32
    dims: [ 3, -1, -1 ]
ensemble_scheduling {
  step [
      model_name: "preprocess"
      model_version: 1
      input_map {
        key: "input_volume"
        value: "input_volume"
      output_map {
        key: "output_slices"
        value: "output_slices"
      model_name: "detect_model"
```

```
model_name: "preprocess"
model_version: 1
input_map {
  key: "input_volume"
  value: "input_volume"
output_map {
  key: "output_slices"
  value: "output_slices"
model_name: "detect_model"
model_version: 1
input_map {
  key: "input_slice"
  value: "output_slices"
output_map {
  key: "output_slice"
  value: "processed_slices"
model_name: "postprocess"
model_version: 1
input_map {
  key: "input_slices"
  value: "processed_slices"
output_map {
  key: "output_volume"
  value: "output_volume"
```



The Model Zoo and you...



- What is needed to serve your model?
- What are your use-cases?
- Are there models which you urgently need?
- Any features you would like to see?

Let's get your first model in the model zoo today?