CANDELA – ITHACA s.r.l.

standard **CAN**dle-based **D**istance **E**stimation with **L**earning **A**lgorithms

Andrea Lessio, Virginia Ajani, Martina Giovalli , Paolo Viviani, Vanina Fissore
Beatrice Bucciarelli, Sibilla Perina, Deborah Busonero

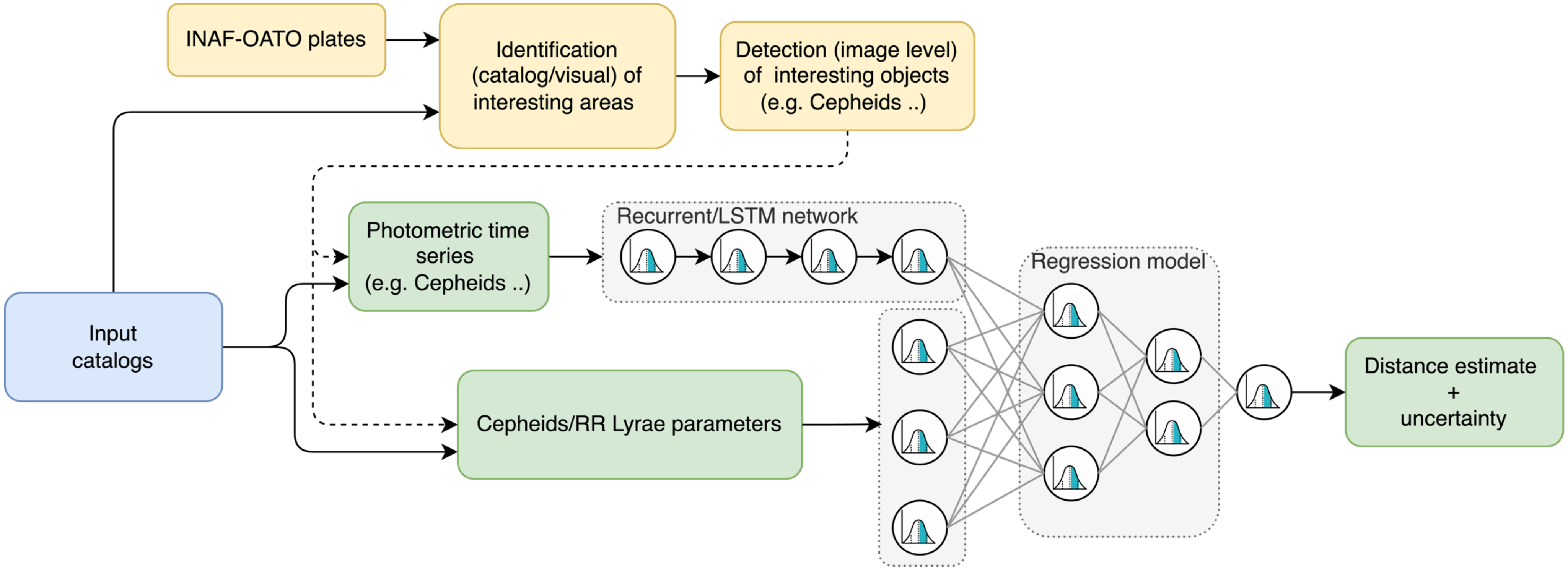
Spoke 3 III Technical Workshop, Perugia 26-29 Maggio, 2025

Scientific Rationale

- ESA Satellite **Gaia** has delivered a massive amount of data (**DR3** ~ 10 TB)
- Other data sources e.g. OGLE, INAF OATO plates archive, rich of information
- Leverage advantages of **machine learning/deep learning techniques** to extract useful information encoded in the data
- **Goal**: development of algorithms and models using learning techniques for **estimating astronomical parameters** (e.g. parallax, distance) for the analysis of data from the Gaia space satellite for different types of distance indicators (**RR Lyrae, Cepheids**) and data (**catalogs**, **photometric series, astronomical plates**)
- **ITHACA s.r.l** has expertise in big data processing, image processing and machine learning techniques

Technical Objectives, Methodologies and Solutions

- **Input**: Gaia DR3, OGLE catalog, astronomical plates from INAF-OATO



- **Output**: generalized distance estimation with learning algorithms

Current status and next steps

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
WP1 - Model with Cepheids catalog	x	x	x	x										
WP2 - Study of uncertainties propagation				x	x	x	x	x	x	x	x	x		
WP3 - Model with RR Lyrae catalog							x	x	x	x	x	x		
WP4 - Identify INAF-OATO plates of interest				x	x	x								
WP5 - Object detection on plates, enrich input								x	x	x	x	x	x	x
				MS1		MS2						MS3, MS4		MS5

We are here 29 of May 2025

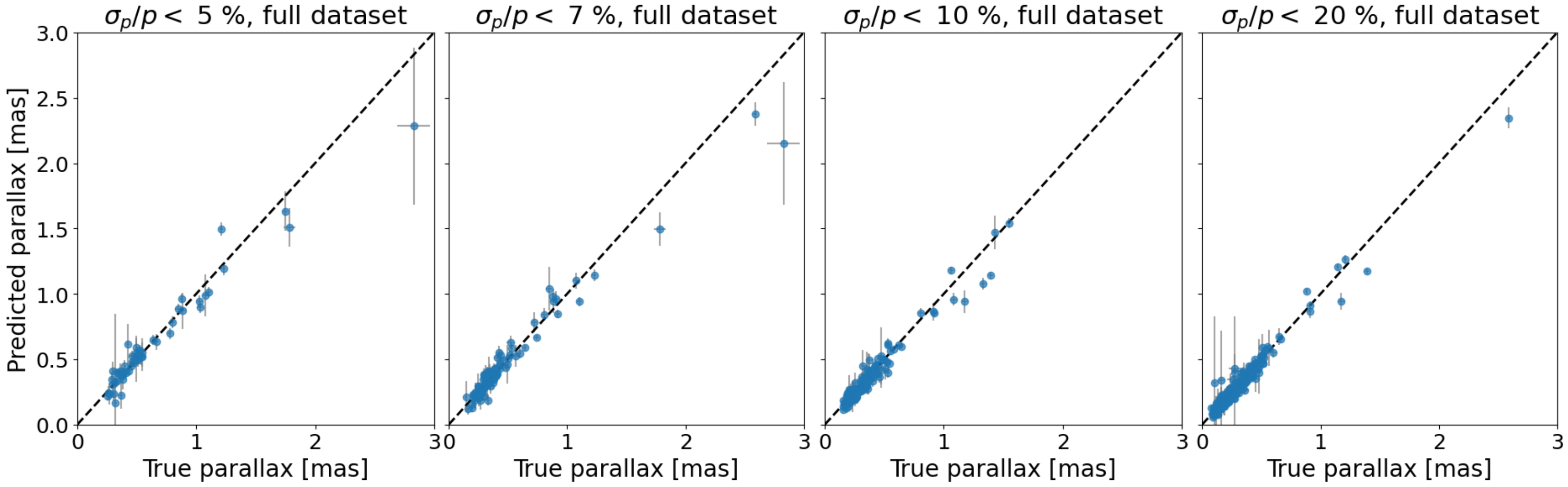
Ask for **two months extension**, project end shifted from end August 2025 → end October 2025

- **MS1** → Identification of ML/DL for Cepheids catalog ✓
- **MS2** → List of interesting AOIs within INAF-OATO plates ✓
- **MS3** → Methodology on uncertainty propagation on the inputs and on intrinsic of the model **(65%)**
- **MS4** → Extension of the ML/DL model to RR Lyrae **(40%)**
- **MS5** → Enrichment of input dataset with detected object in interesting plates from INAF-OATO, generalised model **(40%)**

Technical Objectives, Methodologies and Solutions, WP1

WP1: Identification and development of a ML/DL model, including analysis of photometric time series and stellar parameters, for inference of distance of Cepheid-type standard candles. Validation on a reference dataset provided by INAF-OATO.

- Input: GAIA DR3 Cepheids catalog | comparison among standard ML models, **Gaussian Process Regressor (GPR)** - first implementation to identify model on parallax

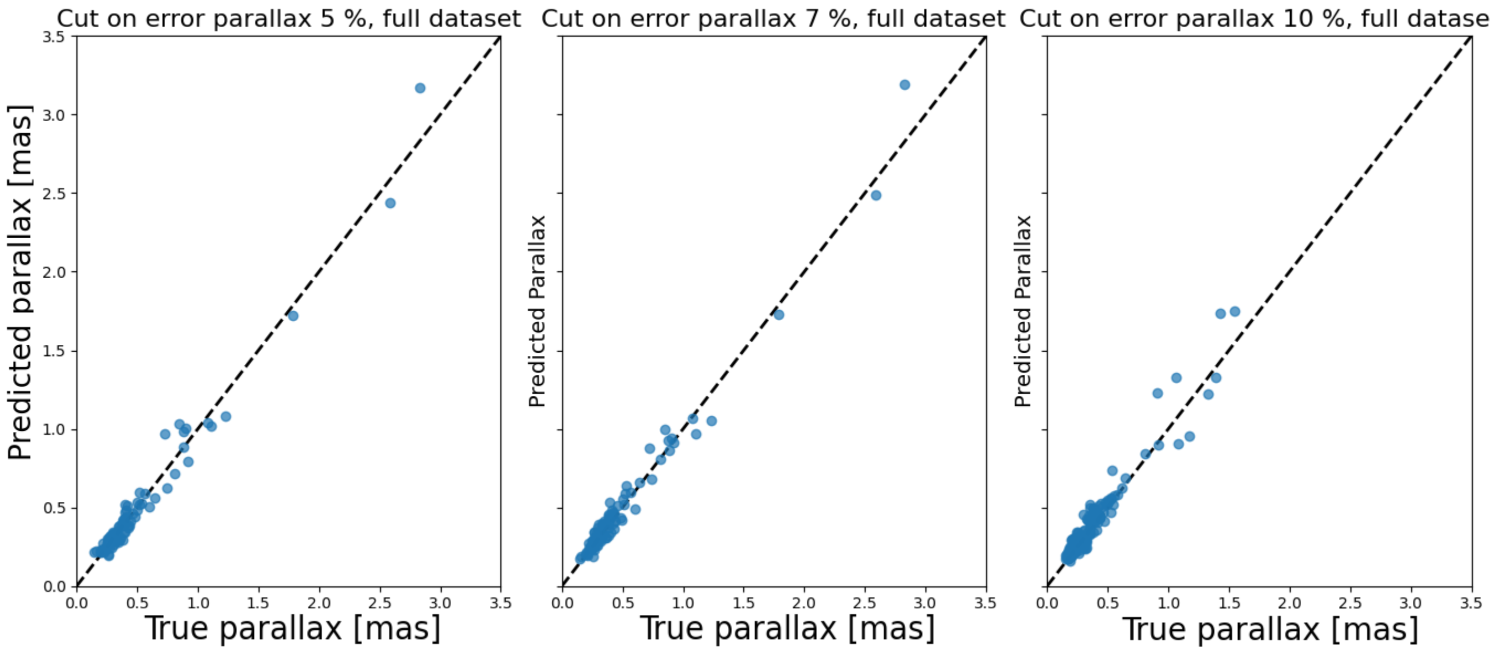


Full dataset	MSE	R^2
$\frac{\sigma_\pi}{\pi} < 5\%$	0.01215	0.94085
$\frac{\sigma_\pi}{\pi} < 7\%$	0.00961	0.94673
$\frac{\sigma_\pi}{\pi} < 10\%$	0.00318	0.95160
$\frac{\sigma_\pi}{\pi} < 20\%$	0.01132	0.89983

→ GPR implementation provides better perforance metrics, and allows to include error propagation and correlation among features

Technical Objectives, Methodologies and Solutions, WP1

- Input: GAIA DR3 Cepheids catalog, **MLP neural network**

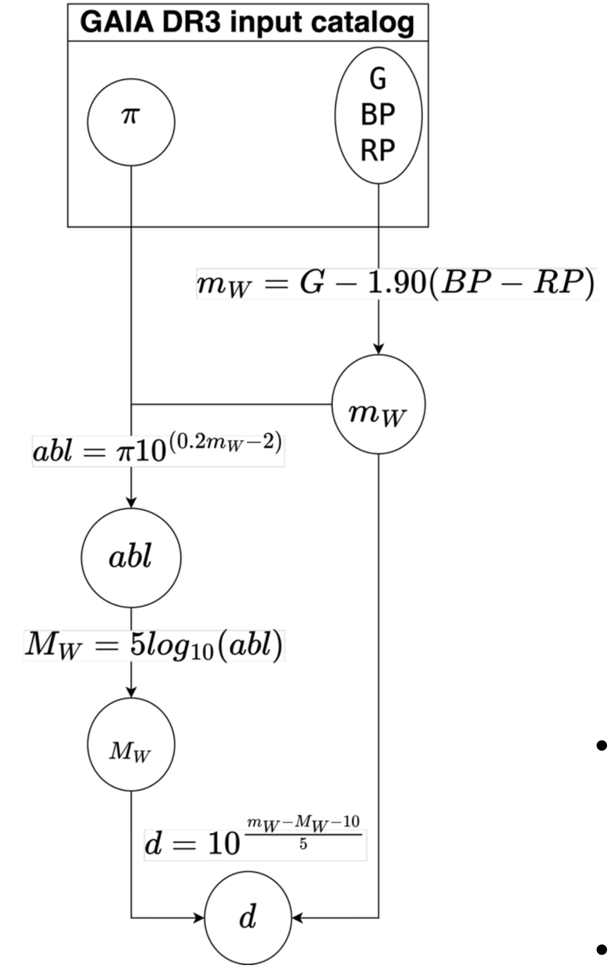


Model	Full dataset	MSE	R^2
Gaussian Processes	$\frac{\sigma_\pi}{\pi} < 5\%$	0.01215	0.94085
	$\frac{\sigma_\pi}{\pi} < 7\%$	0.00961	0.94673
	$\frac{\sigma_\pi}{\pi} < 10\%$	0.00318	0.95160
	$\frac{\sigma_\pi}{\pi} < 20\%$	0.01132	0.89983
MLP Neural Network	$\frac{\sigma_\pi}{\pi} < 5\%$	0.00527	0.97078
	$\frac{\sigma_\pi}{\pi} < 7\%$	0.00467	0.97408
	$\frac{\sigma_\pi}{\pi} < 10\%$	0.006004	0.90858
	$\frac{\sigma_\pi}{\pi} < 20\%$	0.00568	0.9496984

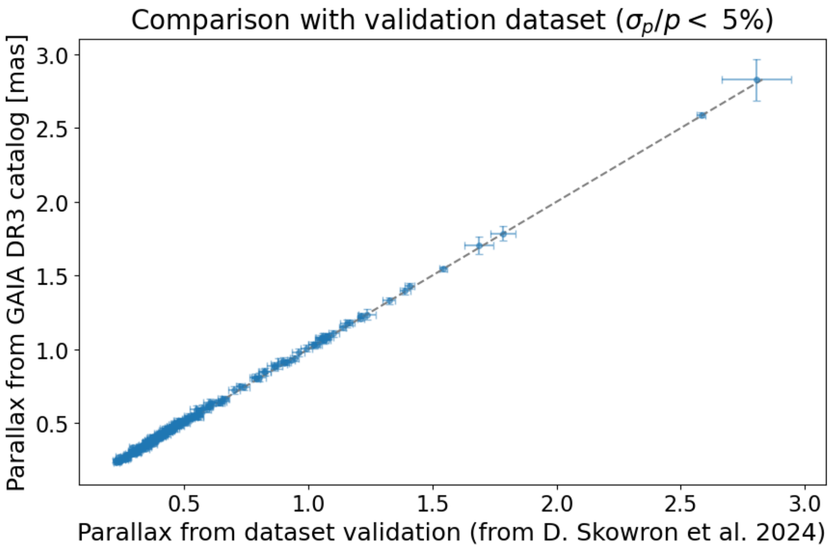
- MLP provides the best combination of performance metrics (still need to include the error treatment)
- What we really want as output is not the parallax, but the distance of a given Cepheids →

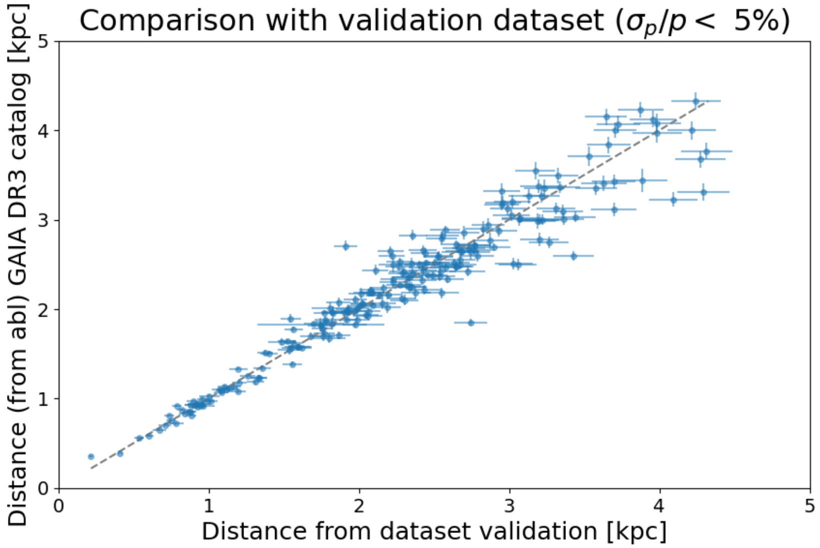
$$\boxed{\pi[\mathrm{mas}] \rightarrow d[\mathrm{kpc}]}$$

Technical Objectives, Methodologies and Solutions, WP1



Madore B. 1982 + Ripepi et al. 2019





- Validation of the **training set** on a reference independent mid-IR photometric dataset from WISE provided by INAF-OATO, from D. Skowron et al. 2024 (https://arxiv.org/abs/2406.09113)

- Next step: validate **output distances**

Technical Objectives, Methodologies and Solutions, WP2

WP2: Study of the propagation of uncertainties for the class of models (i.e., deep neural networks, recurrent neural networks) of interest, with the aim of providing an accurate estimate of the uncertainty on the predicted distance

- Error propagation from the construction of the target to the GPR model:

GAIA DR3 catalog

$$\sigma_\pi \qquad \mathcal{F} = \frac{\bar{\Phi}}{\sigma_\Phi}$$

→

Derived

$$\sigma_{G_{mag}} = \left|\frac{dG_{mag}}{d\Phi}\right| \sigma_\Phi$$

$$\sigma^2_{mW} = \sigma^2_{G_{mag}} + (1.90)^2\,\sigma^2_{BP_{mag}} + (1.90)^2\,\sigma^2_{RP_{mag}}$$

$$\sigma_{abl} = abl\sqrt{\left(\frac{\sigma_\pi}{\pi}\right)^2 + (0.2)^2\,\sigma^2_{mW}}$$

$$\sigma_{M_W} = 5\frac{1}{abl\ln(10)}\sigma_{abl}$$

→

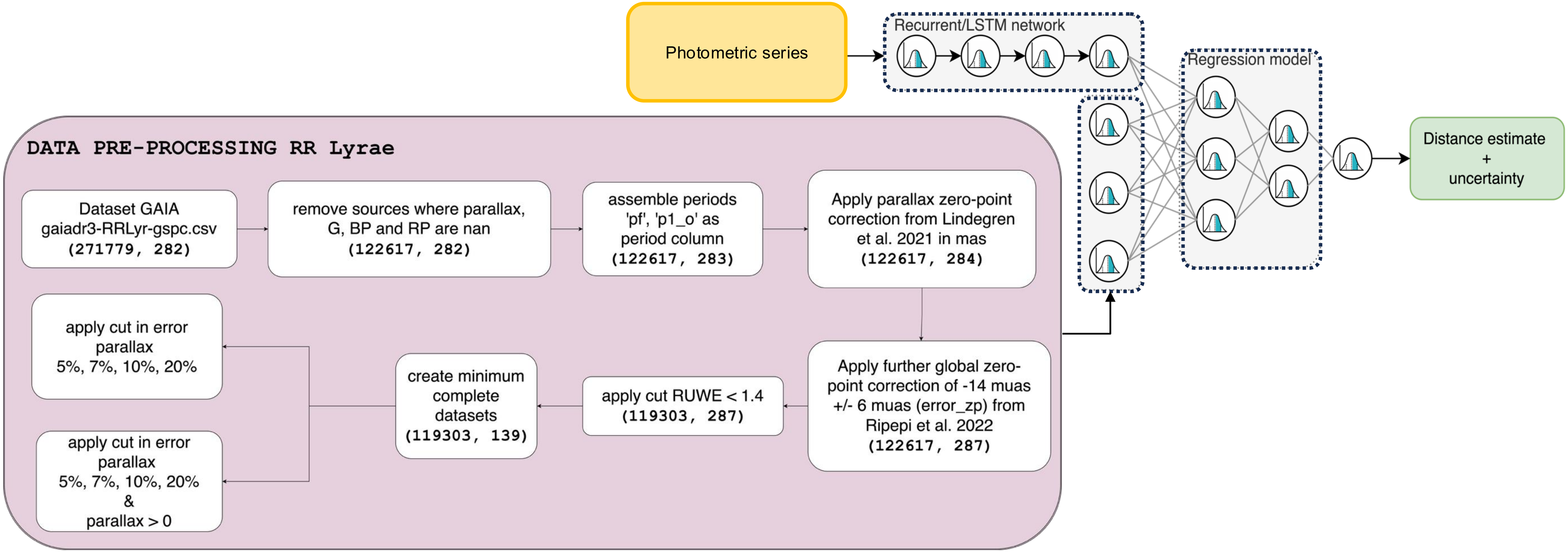
$$\sigma_{\mathcal{D}} = \mathcal{D} \times \frac{1}{5}\sqrt{\sigma^2_{mW} + \sigma^2_{M_W}}$$

- GPR → allows to take into account that the amount of noise (variance) in the data varies across different input values. In practice: add σ_π or σ_{abl} to the diagonal of the kernel matrix during fitting

- Neural Network, MLP → does not take into account uncertainty out-of-the-box. Monte-carlo sampling is being evaluated to propagate the uncertainty on the input values, then it will be complemented with other techniques (e.g. Monte-Carlo dropout, to estimate the intrinsic model uncertainty)

Technical Objectives, Methodologies and Solutions, WP3

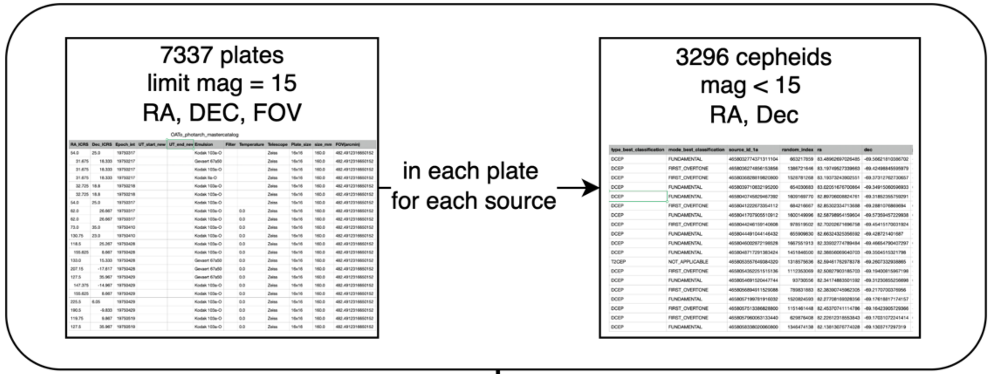
WP3: Extension, fit of the model developed in WP1 to standard RR Lyrae type candles. Validation with a reference dataset provided by INAF



Technical Objectives, Methodologies and Solutions

WP4: Identification of areas of interest using existing catalogs and visual inspection of astronomical plates and .fits images provided by INAF-OATO.

Positional cross-match between OATO astronomical plates catalog and GAIA cepheids

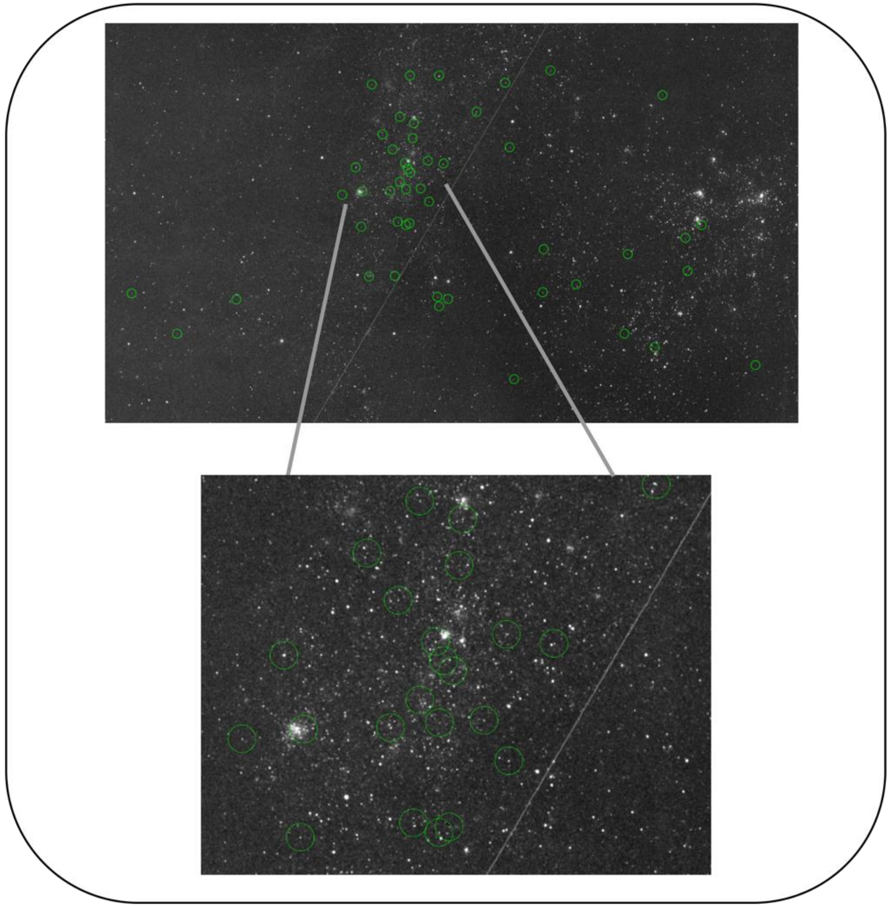


List of plate IDs and number of matches

plate_id	num_of_src
8966	93
557	73
743	60
742	60
630	58
816	57
366	57
815	52
7189	50
586	48
585	48
7190	45

→ **WP5**

Plate 7189: visual inspection



Technical Objectives, Methodologies and Solutions, WP5

WP5: Identification of interesting objects (e.g. Cepheids, RR Lyrae) in the INAF-OATO plates to further enrich the standard candle catalogs with complementary information and generalize the developed algorithm to a different input dataset.

Plate 7189 `.fits` from INAF OATO

ASTROMETRIC CALIBRATION

Inversion + SExtractor
X,Y MAG_AUTO detected objects

	MAG_AUTO	X_IMAGE	Y_IMAGE
2407	-0.2339	9645.8379	11549.3975
2408	0.4855	11228.0342	11547.9668
2677	10.0097	4947.0161	11412.2061

...

Meta-data with Astrometry.net
wcs.fits
(header with RA, DEC)

MATCH FROM MAGNITUDE VALUES

$b_{jkc_{plate}} = A * x_{sol}$

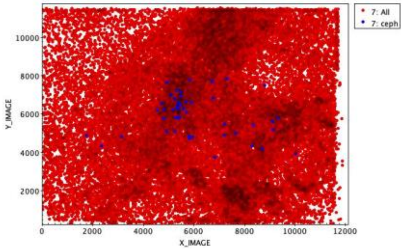
Magnitude match
$b_{jkc_{plate}}$ - $b_{jkc_{cat}}$

List of interesting Cepheids ouptut of WP4

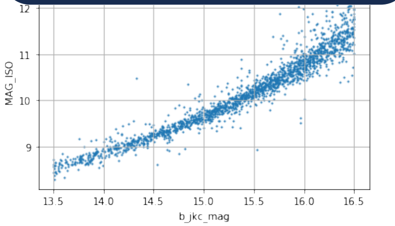
plate_id	num_of_src
8966	93
557	73
743	60
742	60
630	58
816	57
366	57
815	52
7189	50
586	48
585	48
7190	45

PHOTOMETRIC CALIBRATION = characteristic curve **b_jkc_mag** vs **MAG_ISO**

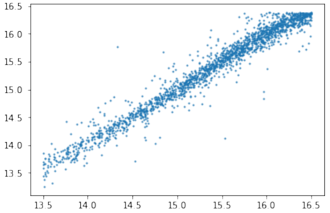
Detection via SExtractor
RA, DEC, MAG_ISO



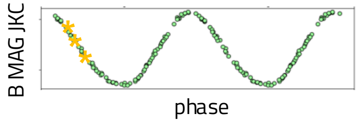
Select deblended Cepheids
$\varepsilon < 0.15$,
$0.3 < B - V < 1.4$,
$13.5 < b_{jkc} < 16.5$



Least square problem:
Minimize $\Sigma(Ax - m)^2$
A = Chebychev polynoms

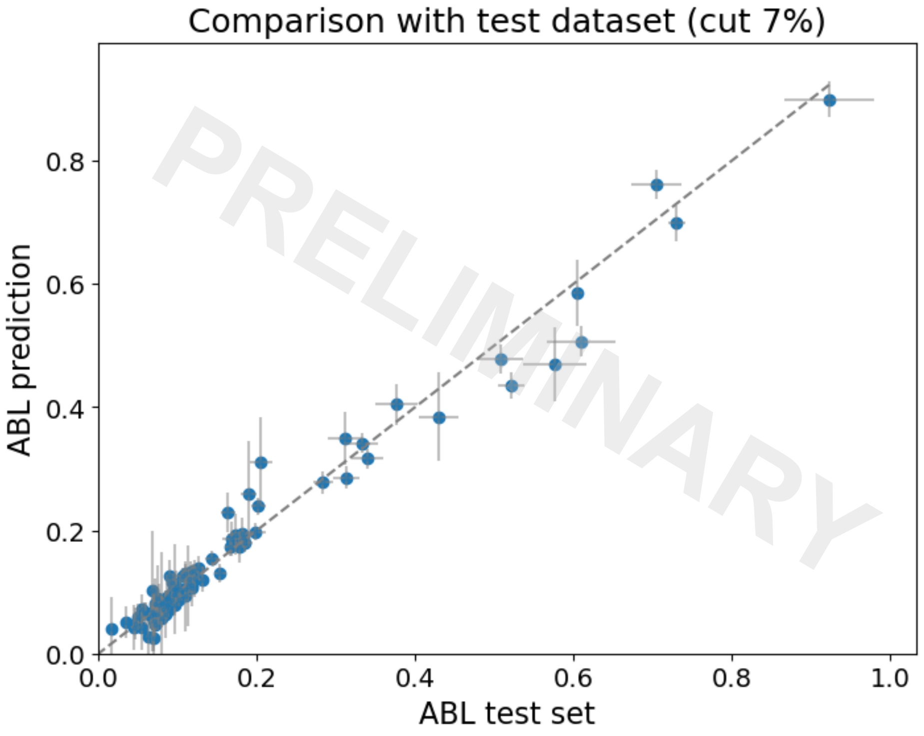


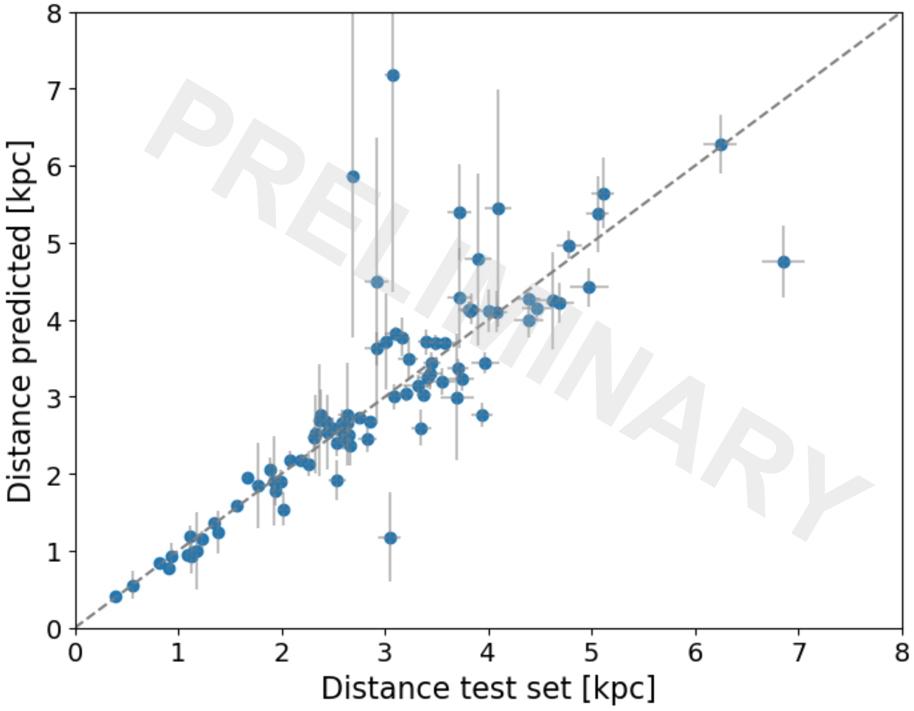
Translate GAIA epoch photometry to b_{jkc}, extend light curve



Main results

Model GPR	5%		7%		10%	
	MSE	R^2	MSE	R^2	MSE	R^2
w/ G, BP, RP $\pi > 0$	0.00104	0.9662	0.00122	0.9517	0.00203	0.9239
w/o G, BP, RP $\pi > 0$	0.00128	0.9619	0.00138	0.9524	0.00250	0.9191
w/ G, BP, RP $\pi > 0$ with error	0.00105	0.96878	0.00111	0.9618	0.001456	0.9530
w/o G, BP, RP $\pi > 0$ with error	0.00115	0.96578	0.00117	0.9596	0.001486	0.9521





Model 1 - ML → **GPR** for Cepheids with error propagation without G, BP, RP magnitudes in training

Next step → MLP with error propagation

Main results

- List of interesting plates with matches:
  - **1373** plates with Cepheids matches
  - **763** plates with more than one Cepheid match

- 49 detected interesting object with astrometric calibration in WP5 for plate 7189, field with LMC

Plate 7189

Object: LMC (campo A)
Telescope: GPO-ESO
Epoch: 31-05-1992
Site: La Silla
Optical design: Refractor

Statistics of residuals	
count	1.907000e+03
mean	9.382927e-15
std	1.541988e-01
min	-1.129944e+00
25%	-8.124337e-02
50%	2.666663e-04
75%	6.833642e-02
max	1.421636e+00

Final Steps

WP1: first candidate model identified, GPR with error propagation without G, BP and RP magnitude in training ✓

WP2: extend uncertainty propagation to MLP neural network + incorporate photometric series with LSTM + obtain results with MLP for distance

WP3: extend model to RR Lyrae catalog → apply Model 1 and MLP to RR-Lyrae reduced datasets

WP4: list of AOIs plates matching Gaia catalog provided to INAF OATO ✓

WP5: interesting objects in plates found in WP4, calibration performed for one plate → extend for other plates + convert plate epoch into GAIA epoch and extend light curve

Thank you for your attention!

ADDITIONAL SLIDES

Gaussian Processes

$$f\left(\vartheta\right) \sim GP[\mu\left(\vartheta\right), K\left(\vartheta, \vartheta'\right)],$$

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K\left(\vartheta, \vartheta\right) + \sigma_n^2 \mathbb{I} & K\left(\vartheta, \vartheta_*\right) \\ K\left(\vartheta_*, \vartheta\right) & K\left(\vartheta_*, \vartheta_*\right) \end{bmatrix}\right),$$

$$\begin{aligned} \mu_\star &= \mu\left(\vartheta_*\right) + K\left(\vartheta_*, \vartheta\right)\left[K\left(\vartheta, \vartheta\right) + \sigma_n^2 \mathbb{I}\right]^{-1}\left(f - \mu\left(\vartheta\right)\right), \\ \Sigma_* &= K\left(\vartheta_*, \vartheta_*\right) - K\left(\vartheta_*, \vartheta\right)\left[K\left(\vartheta, \vartheta\right) + \sigma_n^2 \mathbb{I}\right]^{-1} K\left(\vartheta, \vartheta_*\right). \end{aligned}$$

With the kernel in our case defned as $K\left(x_i, x_j\right) = \exp^{-\frac{d\left(x_i, x_j\right)^2}{2\ell^2}}$

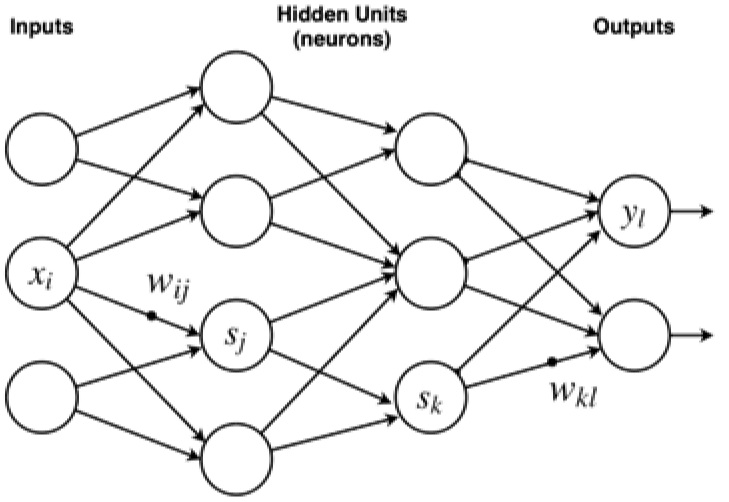
Neural network - MLP

MLPs are a kind of Artificial (Deep) Neural networks that use all-to-all connectivity between the neurons of hidden layers.
They can perform both classification and regression tasks (like in this case), and they work by iteratively minimising the value of a loss function, that is typically a metric of the distance between a "ground truth" value and a "predicted" value.

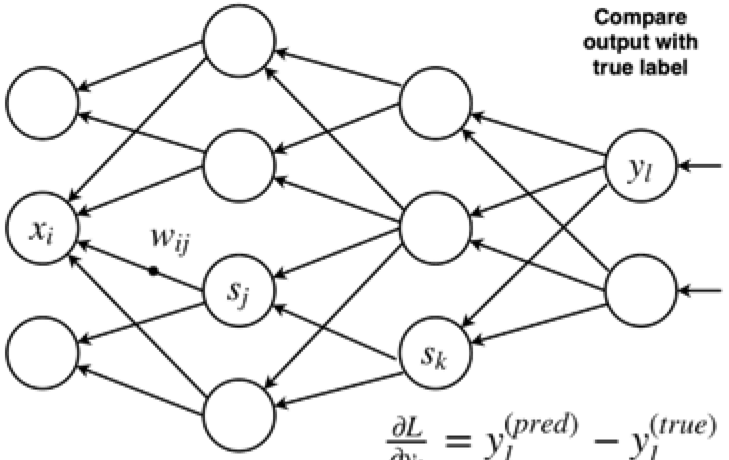
Although they do not provide built-in uncertainty estimation, several techniques have been explored to achieve it with DNNs. For instance

Propagation of input uncertainty: Monte-carlo sampling augments the dataset by sampling input data from a normal distribution with avg and stddev. This provides a distribution of the output values that represents the propagation of the error.

Model intrinsic uncertanty quantification: Monte-carlo dropout (https://proceedings.mlr.press/v48/gal16.html) randomly "turns off" some neurons, providing a distribution of the output that is correlated to the uncertainty of the model itself.



$$z_j = \sum_i x_i w_{ij} \qquad s_j = f(z_j)$$



$$\frac{\partial L}{\partial y_l} = y_l^{(pred)} - y_l^{(true)}$$

$$\frac{\partial L}{\partial s_j} = \sum_k w_{jk} \frac{\partial L}{\partial z_k} \qquad \frac{\partial L}{\partial y_l} = \frac{\partial L}{\partial y_l} \frac{\partial y_l}{\partial z_l}$$