Regression with Dense Deural Network using TensorFlow

Simone Albanesi
(Dated: January 27, 2022)

I. INTRODUCTION

Working notes on the regression part with a dense Neural Network (NN) using `Keras` with `TensorFlow` in backend. We use this framework since allows us to exploit the flexibility of NNs (while `Scikit-learn` doesn't).

The task is to apply the NN to the quantities recovered from `GstLAL` x and to predict values $y^{\text{pred}} = f(x)$ that are a better approximation of the injected values y^{true}.

Some pros of NNs are:

- virtually infinite number of architectures, many aspects to tune on our problem
- they are used everywhere nowadays, so it is easy to find material online
- is possible to put constraints on the output
- easy to customize

On the other hand, to fully exploit the NN we need a big dataset and the high number of tunable aspects makes the cross validation more difficult. In our case we will use at most 2 hidden layers. In principle more complex architectures could be used but in our case I don't think it would be worth since we do not have a huge training sample. For regression tasks, the activation function in the output layer generally is a linear function.

Steps of the regression:

1. **Normalization of the data**. The usual normalization (`StandardScaler`) trasforms the data so that we have zero mean and standard deviation equal to 1. Nonetheless I prefer to use the linear map `MinMaxScaler` that maps the data in the interval $[-1, 1]$. This choice is linked to the constraints on the output, see Sec. II.
2. **Initialization of the model**. We decide the architecture and all the other aspects. A typical configuration is:
   - layers: two layers with 100 neurons, but e.g. on the `v0c0` dataset a more complex architecture could be slightly better (see Sec. IV for more info on this);
   - activation function in hidden layers: `ReLU`. Other options like the sigmoid or the hyperbolic tangent require longer training and provide worse results.;
   - activation function in output layer: `linear`;
   - optimizer: `Adam` (optimized version of the stochastic gradient descent);
   - loss function: `MeanSquaredError()` (see Sec. VI for more details and other options);
   - learning rate: 0.001;
   - epochs and batch-size: a good combination is 250 and 32, respectively. Sometimes I use a bigger batch-size to speed-up the training (e.g. I used 128 for the cross-validation on the `v0c0` dataset).
3. **Training**. Using the dataset `v0c0`, $N_{\text{epochs}} = 250$ and $N_{\text{batch}} = 128$ the training takes from 60 to 90 seconds, depending on the architecture. During the training we also use the 20% of the training dataset for the validation so that we can monitor the loss and the R^2 coefficient (see definition below) at each epoch.
4. **Evaluation of the accuracy**. To evaluate the goodness of the model we use the coefficient of determination $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}$, where $SS_{\text{res}} = \sum (y_i^{\text{true}} - f(x_i))^2$, $SS_{\text{tot}} = \sum (y_i^{\text{true}} - \bar{y}^{\text{true}})^2$, and $\bar{y}^{\text{true}}$ is the mean. This is also the `score` of `Scikit-learn`. It is better to evaluate the R^2 coefficient for each feature instead of a global R^2.

II. CONSTRAINTS ON THE OUTPUT

We need some constraints on the output since we don't want to predict negative masses or naked singularities. A solution is to modify the activation function σ in the output layer. Another option could be to enforce the constraints in the loss function, but I have not tried this (and in any case this would not guarantee physical predictions). The idea is to make σ saturate when the prediction is out of the physical range. The first step is to normalize the data in the interval $[-1, 1]$ using the `MinMaxScaler`. Then for σ we use the following prescription:

$$\hat{\sigma}(x) = \begin{cases} -1 & x \leq -1 \\ x & -1 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases} \tag{1}$$

Using a sigmoid ansatz is possible to find a smooth approximants

$$\sigma^N(x) = \frac{2}{e^{-2f^N(x)} + 1} - 1, \tag{2}$$

where

$$f^N(x) = \sum_{n=odd}^{N} \frac{x^n}{n}. \tag{3}$$

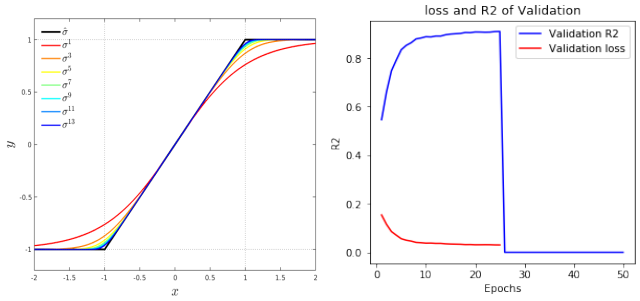

FIG. 1. **Left panel**: activation functions to use in the output layer so that the output is constrained in the interval $[-1, 1]$. **Right panel**: history of a NN with two hidden layers with 20 neurons and $\sigma^5(x)$ as output function. The drop in the accuracy at the 18th epoch is due to the presence of NaN. That's why we use $\hat{\sigma}$ and not a smooth approximant σ^N.

The function $f^N(x)$ is simply found requiring that $\sigma^N(x) = x + O(x^{N+2})$ for $x \to 0$. These activation functions are shown in the left panel of Fig. 1. However, for $N \geq 3$ is possible that some NaN appear during the training, see e.g. the right panel of Fig. 1 where we show a case with $\sigma^5(x)$. I remember that once also $\sigma^3(x)$ failed, but I cannot find this configuration again (is a rare event). In any case to be safe I will use the piecewise function $\hat{\sigma}$ ($\sigma^1(x)$ is not steep enough). Clearly it is also possible to saturate only the upper of the lower boundary, in particular we for the masses I will often use

$$\hat{\sigma}_{\rm LB}(x) = \begin{cases} -1 & x \leq -1 \\ x & -1 \leq x. \end{cases} \tag{4}$$

However I am not sure that removing the upper boundary makes sense, maybe it would be better to use the upper constraint also for the masses, TBD.

III. SOME RESULTS

The regression with NN works pretty well on the `v0` datasets, while on the dataset `v1` the R^2 coefficients are smaller since there is more degeneracy. We use 2 hidden layers with 100 neurons (i.e. 12,411 trainable parameters, while $N_{\rm sample} = 2 \cdot 10^4$) and we train on 250 epochs using $N_{\rm batch} = 64$. Different architectures are discussed in Sec. IV but the results are more or less the same for complex-enough NNs. For the spin components we use $\hat{\sigma}$ as output activation function, for the masses and the mass ratio we use $\hat{\sigma}_{\rm LB}$ and a linear function for the angle. In Fig. 2, Fig. 3 and Table I there are the results for this model. In the first plots we plot the predicted value against the injected values, while in the seconds we plot the prediction (orange) on the recovered (blue). Note that the recovered here are not physically consistent but it is ok since the main goal at this point is to see if a neural network is able to recover the original quantities. The relevant thing is that we are able to predict values that have physical meaning.

TABLE I. Loss (Mean Squared Error) and R^2 coefficients for the two datasets `v0c0`, `v1c0`. We used 2 hidden layers with 100 neurons, $N_{\rm batch} = 64$, 250 epochs, constrained output. See Sec. III for discussion.

	`v0c0`	`v1c0`
loss (MSE)	0.0079	0.0533
R^2 mean	0.9652	0.7533
$R^2[m_1]$	0.9962	0.8938
$R^2[m_2]$	0.9860	0.9454
$R^2[s_x^1]$	0.9582	0.6134
$R^2[s_y^1]$	0.9529	0.6344
$R^2[s_z^1]$	0.9549	0.6436
$R^2[s_x^2]$	0.9539	0.6853
$R^2[s_y^2]$	0.9570	0.6409
$R^2[s_z^2]$	0.9568	0.6468
$R^2[\theta]$	0.9073	0.6648
$R^2[q]$	0.9957	0.9244
$R^2[M_c]$	0.9978	0.9938

IV. CROSS VALIDATION ON LAYERS

In order to decide the architecture to use I did a cross validation on the `v0c0` dataset (the reason why I have not cross-validated on the `GstLAL` dataset is explained in Sec. V). I trained the models on 250 epochs using $N_{\rm batch} = 128$ for different architectures. The results are shown in Fig. 4. The plots show that NNs with 2 layers with ~ 100 neurons each provide good enough scores and more complex NNs do not provide much better results. After a certain number of trainable parameters the R^2 coefficient reaches a plateaux and doesn't increase any more. The only exceptions are: (i) M_c: the best score is obtained with single-layer NNs (the recovered quantity is obtained only adding random Gaussian noise); (ii) θ: this is the quantity with lowest score and the R^2 coefficient continues to (slightly) increase without reaching a plateaux even for ~ 300 neurons in each layer.

Finally note that many NNs have more parameters than training samples, $N_{\rm param} > N_{\rm train}$, since $N_{\rm train} = 2 \cdot 10^4$ for the `v0v0` dataset. This shouldn't be a problem for NN (and indeed the score is good also on the test-set that has $N_{\rm test} = 1.5 \cdot 10^4$), however we could also decide to use smaller NNs since the accuracy in more complex NNs is not drastically better than in simpler models. For example in Fig. 2 and Fig. 3 I used two hidden layers with 100 neurons so that $N_{\rm param} = 12411 < N_{\rm sample} = 2 \cdot 10^4$.

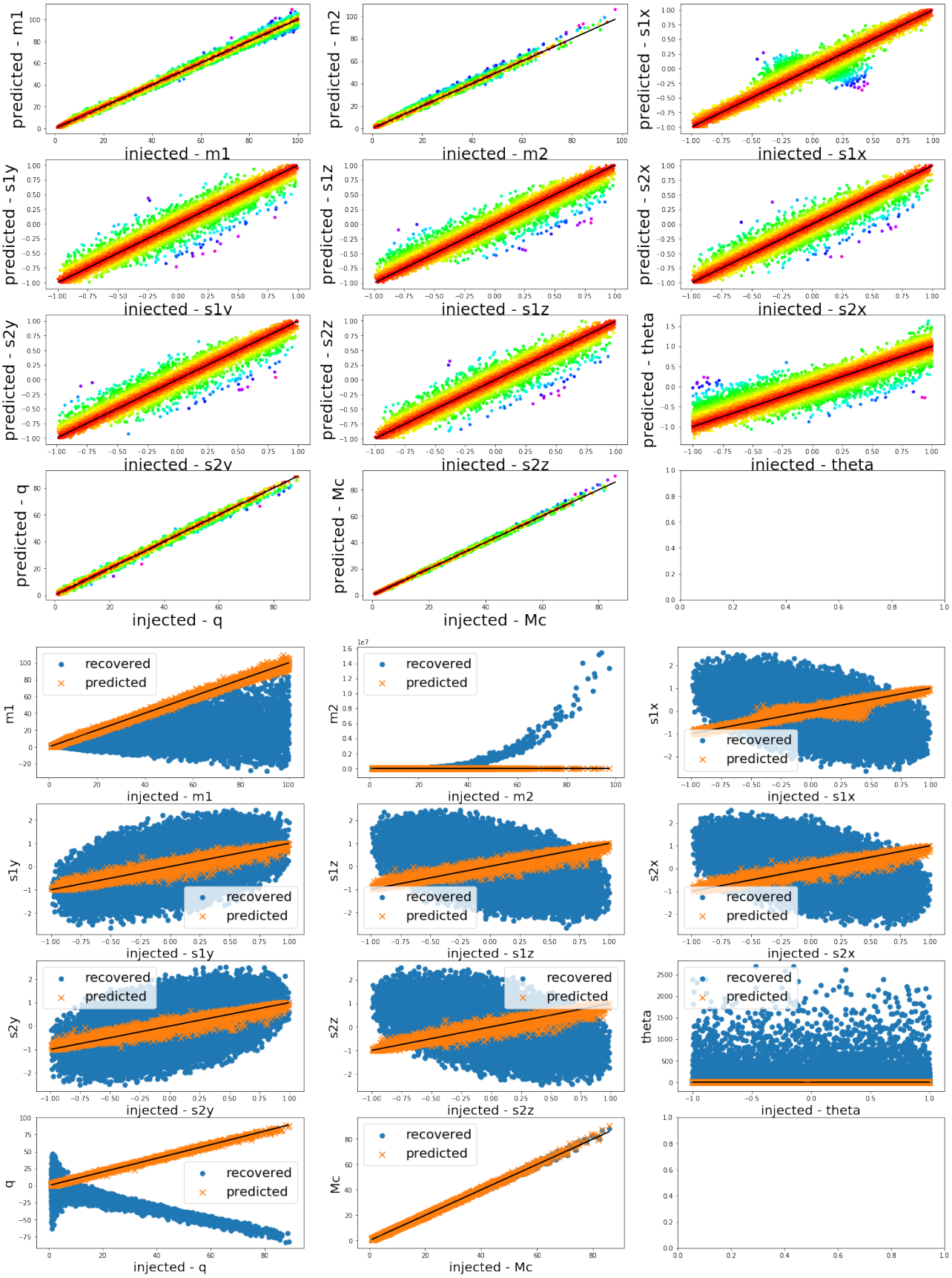

FIG. 2. Results of the regression on the `v0c0` dataset. We used 2 hidden layers with 100 neurons, $N_{\text{batch}} = 64$, 250 epochs, constrained output. See Sec. III for discussion. The colors in the first plots is related to the absolute difference between predicted and injected. The black line is the bisector.

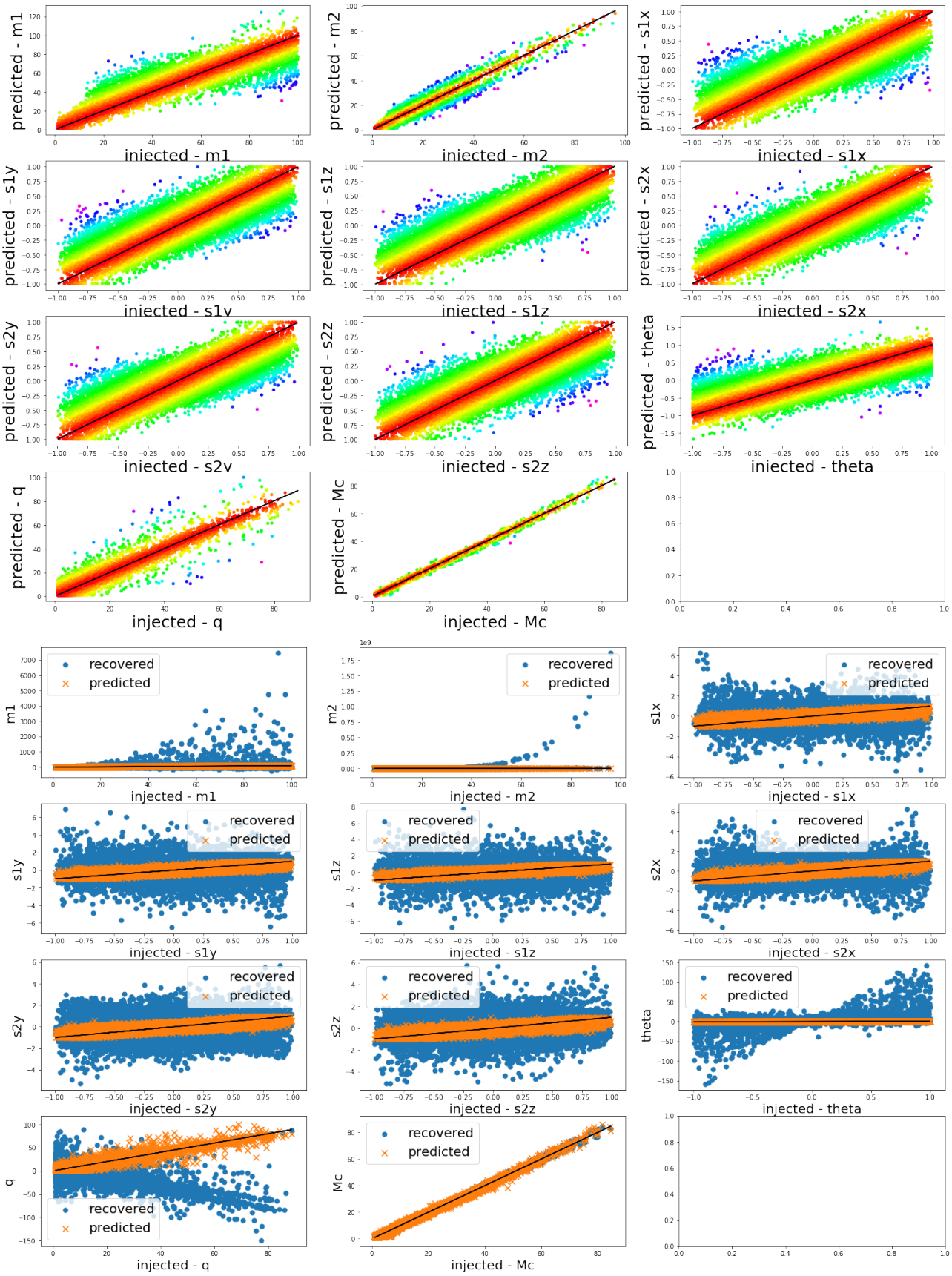

FIG. 3. Results of the regression on the `v1c0` dataset. We used 2 hidden layers with 100 neurons, $N_{\text{batch}} = 64$, 250 epochs, constrained output. See Sec. III for discussion. The colors in the first plots is related to the absolute difference between predicted and injected. The black line is the bisector.

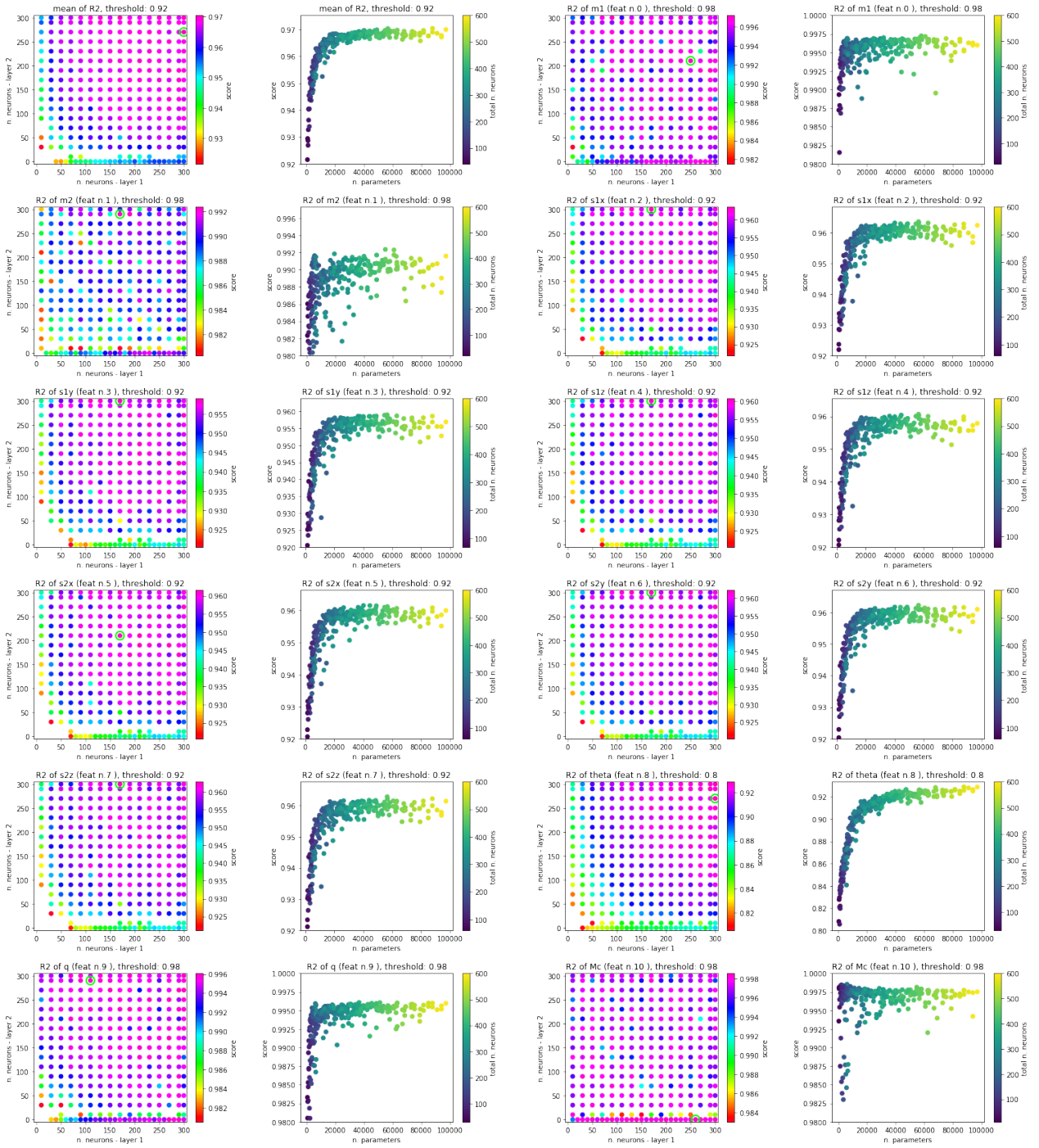

FIG. 4. R^2 coefficients for different architectures. We show only the models that have an R^2 coefficient above the threshold indicated in the title of each panel. We show the results for the mean of R^2 and for the R^2 of each feature. In the rainbow-scatter plot the green circle marks the highest R^2 coefficient.

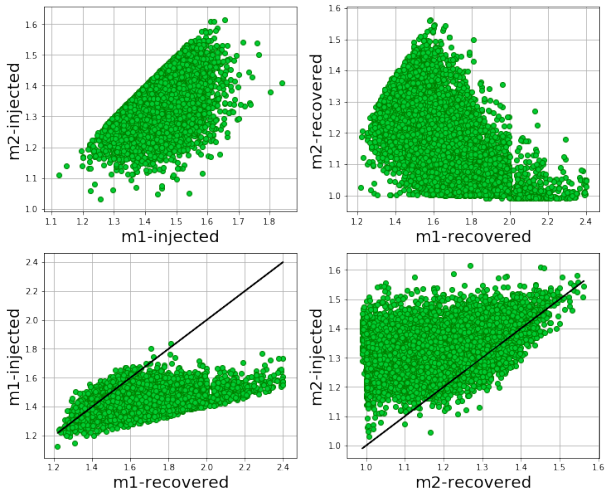

FIG. 5. Test-dataset, compare with Fig.1 of Chatterjee-1911.00116. Do the systematics in the waveform model matter?

V. GSTLAL BNS DATASET

We also have a dataset obtained with injected parameters of nonspinning BNSs recovered with `GstLAL`. In this case we have only three features: m_1, m_2, M_c. The regression on this dataset is more problematic:

- The output of the regression for the two masses m_i seems almost independent on the architecture and is not improved with more complex NNs. The result for M_c instead is more sensible on the architecture (but the regression seems useless for M_c)
- The predictions of m_i have a sharp edge in correspondence of the region with most degeneracy, see Fig. 6
- The model stops to learn after the second epoch. Training for more epochs only improves the prediction of the lowest and highest chirp masses, but does not improve the prediction of m_i.
- The mass ratio computed with the predicted masses is in a very narrow range while the injection have a much wider interval, see the last panel of Fig. 6.

Therefore in this case the regression does not seem to work properly. I have also tried to do the regression only on $m_{1,2}$ and M_c and recovering the other mass analytically but the results are the same. Even SVR gives the same results (qualitatively).

VI. LOSS FUNCTION

In order to fix the problem with q in the `GstLAL` dataset, I tried to modify the loss function to minimize. The default function for regression tasks is the Mean Squared Error

$$J_{\rm mse} = \text{mean} \sum_i \left(y_i^{\rm true} - f(x_i)\right)^2 . \tag{5}$$

I tried to modify it including a q-penalty and a M_c-penalty:

$$\begin{aligned} J = \text{mean} \Bigg(& \sum_i \left(y_i^{\rm true} - f(x_i)\right)^2 \\ & + \lambda_{\rm q} \sum_i \left(q_i^{\rm true} - q_i^{\rm pred}\right)^2 \\ & + \lambda_{\rm M_c} \sum_i \left(M_{c,i}^{\rm true} - M_{c,i}^{\rm pred}\right)^2 \Bigg) \end{aligned} \tag{6}$$

where (not surprisingly)

$$q_i^{\rm pred} = \frac{m_{2,i}^{\rm pred}}{m_{1,i}^{\rm pred}}, \tag{7}$$

$$M_{c,i}^{\rm pred} = \frac{\left(m_{1,i}^{\rm pred} m_{2,i}^{\rm pred}\right)^{3/5}}{\left(m_{1,i}^{\rm pred} + m_{2,i}^{\rm pred}\right)^{1/5}}, \tag{8}$$

and $q^{\rm true}$, $M_c^{\rm true}$ are analogous. Note the M_c-term is not the same in MSE since here the chirp mass is computed using m_i while in the MSE part M_c is the third feature. In order to have the predicted masses in the loss function we need to rescale x but since the loss function is defined using the Keras backend and for the moment this is implemented only for `MinMaxScaler`. We cannot use `numpy` in the Keras backend, so the loss function is not fully vectorized (but this is not a big issue). I tried to fit the data with different values for the hyperparameters λ_q and $\lambda_M c$ but in every case I didn't find a solution to the issues of Sec. V since the effect of these penalty is marginal, even if we use high values for λ_i. Maybe this is not the correct way to include penalties.

VII. GSTLAL BNS DATASET - A DIFFERENT APPROACH

I also tried to do something different on the `GstLAL` dataset. So the pipeline recovers M_c perfectly (I am not even sure that make a regression on M_c makes sense), then I suppose that it recovers also one of the two masses (I don't know which one) and compute the second one analytically. Therefore the degeneracy problem in one mass should be "specular" in the other masse, so maybe to remove the degeneracy problem in the regression we can consider suitable combinations of the two masses m_i instead of considering them separately. The using this quantity $g(m_1, m_2)$ and M_c we can obtain two masses m_i. It turns out that symmetric quantities are better,

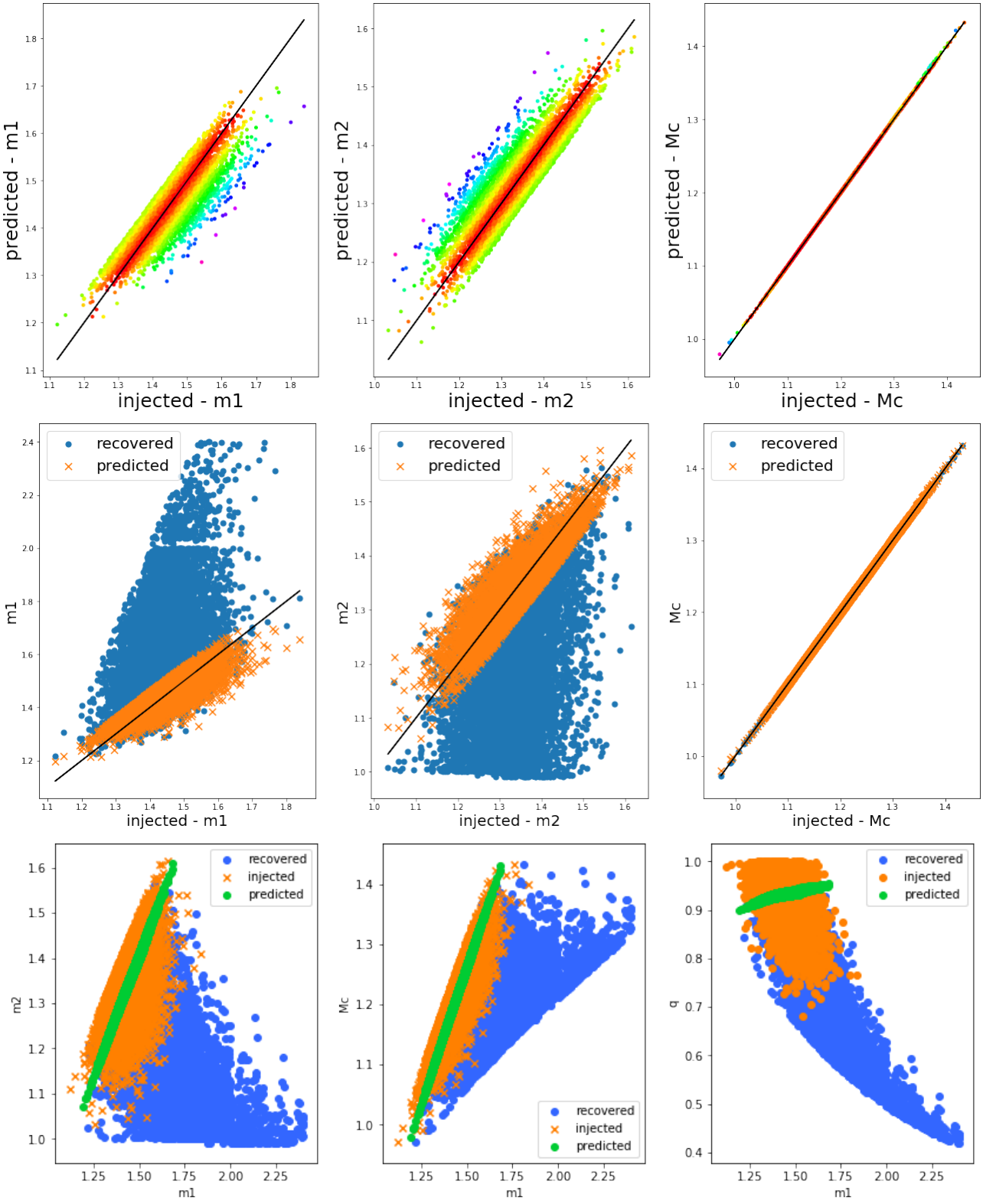

FIG. 6. Results of the regression on the `GstLAL` dataset. We used one hidden layer with 100 neurons, $N_{\text{batch}} = 64$, 50 epochs, unconstrained output. The colors in the first plots is related to the absolute difference between predicted and injected. The black line is the bisector. The final R^2 coefficients for m_1, m_2 and M_c are 0.7363, 0.7819, and 0.9999, respectively. In the last three panels we plot m_2, M_c and q over m_1 for the injected, recovered and predicted quantities.

in particular $p_k = (m_1 m_2)^k$ can be regressed very well if the exponent k is not too high, obtaining R^2 coefficients such that $1 - R^2 \leq 10^{-4}$. Then the two masses can be obtained analytically from

$$m_{1,2} = \frac{p^3}{2M_c^5}\left(1 \pm \sqrt{1-4\nu}\right), \tag{9}$$

where $p = m_1 m_2$ (i.e. we omit the index $k = 1$) and ν is the symmetric mass ratio

$$\nu \equiv \frac{M_c^{10}}{p^5} = \frac{m_1 m_2}{(m_1+m_2)^2} \in (0, \frac{1}{4}]. \tag{10}$$

However, note that the condition

$$\nu = M_c^{10}/p^5 \leq \frac{1}{4} \tag{11}$$

is not enforced during the training (i.e. the predictions for p_k and M_c are not constrained in this sense), then it is possible that (11) will be slightly violated, making the two masses complex. For this reason we recover the two masses with a modified version of Eq. (9)

$$m_{1,2}^{\rm pred} = \frac{p_{\rm pred}^3}{2M_{c,\rm pred}^5}\left(1 \pm \sqrt{1 - 4\min\left(\nu_{\rm pred}, \frac{1}{4}\right)}\right). \tag{12}$$

Note that the condition (11) can be violated pretty often if we have many equal-mass binaries in the dataset (and this is precisely our case). However note that with this approach the predictions for m_i and M_c can be not consistent, i.e. $M_c^{\rm pred} \neq (m_1^{\rm pred} m_2^{\rm pred})^{3/5}/(m_1^{\rm pred}+m_2^{\rm pred})^{1/5}$. To enforce this condition we then compute m_2^* requiring

$$M_c^{\rm pred} \equiv \frac{(m_i^{\rm pred} m_j^*)^{3/5}}{(m_i^{\rm pred} + m_j^*)^{1/5}}, \tag{13}$$

where $i \neq j = 1, 2$, explicitly (omitting the superscript 'pred' in the RHSs):

$$m_j^* = \frac{M_c^{5/3}\left(3^{1/3} 2 M_c^{5/3} + 2^{1/3} S_i^2\right)}{6^{2/3} m_i^{3/2} S_i}, \tag{14}$$

$$S_i = \left(9 m_i^{5/2} + \sqrt{81 m_i^5 - 12 M_c^5}\right)^{1/3}. \tag{15}$$

The results for this kind of regression using the features $p = m_1 m_2$ and M_c are shown in Fig. 7. We use a NN with one hidden layer with 100 neurons and ReLU, linear activation function in the output layer (i.e. no constraints), MSE loss function and we train for 100 epochs using $N_{\rm batch} = 128$. While the regression on p and M_c is almost perfect (R^2 coefficients of 0.99993 and 0.99997, respectively), the other recovered quantities show some strange behavior and, most importantly, they are strongly dependent on the initial state on the NN, i.e. training the NN different times using always the same options leads to pretty different results for m_i and q. Moreover, the mean predicted/injected errors for m_1, m_2, M_c, and q are slightly worse since here we get 2.8%, 2.7%, 0.02%, and 5.4% respectively, while with the standard regression of Sec. V we got 2.3%, 2.2%, 0.05% and 4.6%.

At this point I am starting to be pessimistic on the regression part of this project (at least with NN, maybe GPR will save us), even if it is possible that using the final dataset will solve this problem since we will have more features that could help to overcome the problem of the masses (or maybe will worsen it).

Another attempt could be to do the regression on ν and M_c and then compute m_i from these two quantities. In this case we could apply a constrained output function to ν in order to have a more consistent output (and in this case we should write a new class for the linear-scaler since the `MinMaxScaler` that we are using now in the constrained output wouldn't be the best choice for constraining a quantity like ν). However I don't think that this will solve all the problems that we have with this dataset.

VIII. THINGS TO DISCUSS

- The regression with NNs does not work properly on the GstLAL BNS dataset. Also Yanyan had some problems on this dataset. Probably this is linked to the recovery of the masses, but we need to (i) fully understand why this happens, (ii) fix it. However it's possible that this problem won't be there with a dataset with masses in a much broader range and with other features (or maybe will be there anyway, who knows).

- Does the waveform model matter a lot?

- NN and SVR have the same problem on the GstLAL BNS dataset and the only similarity between the two algorithms is that they both minimize a loss function. To my understanding GPR has a completely different approach: how does it work on this dataset? Does it have the same issues?

- Maybe we could try to simulate the `GstLAL`'s noise in a more realistic way for `v0c0` and `v1c0`, but I am not sure that this would be useful since we already have a GstLAL dataset (even if only with BNS).

- Since we are starting to use more complex NNs (at least for `NewRealistic` datasets, i.e. `v0c0` and `v1c0`), should we try to use some dropout? [not a first order priority in any case]

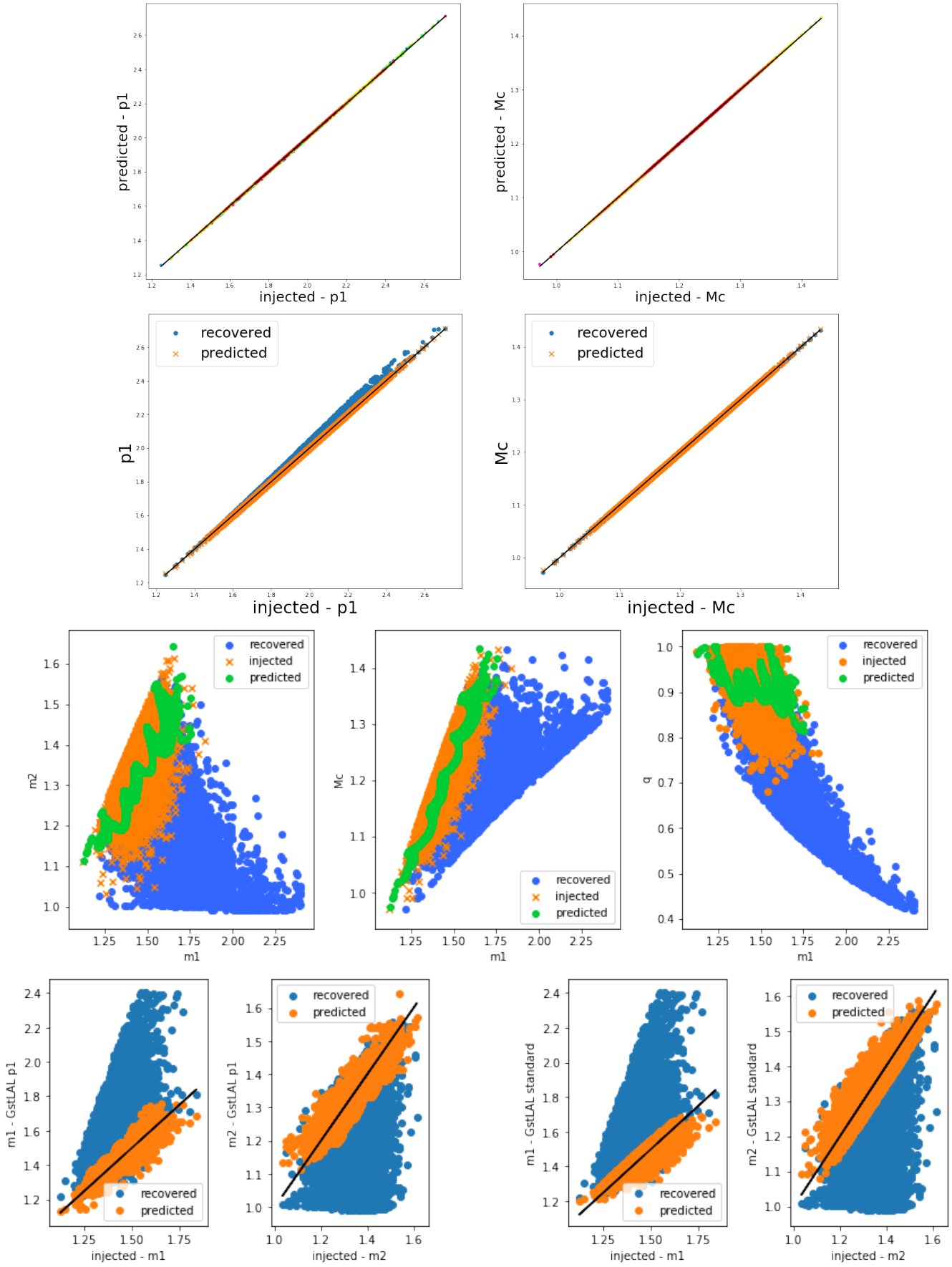

FIG. 7. Results of the regression on the `GstLAL` dataset using the approach of Sec. VII. We used one hidden layer with 100 neurons, $N_{\rm batch} = 128$, 100 epochs, unconstrained output. The colors in the first plots is related to the absolute difference between predicted and injected. The black line is the bisector. The final R^2 coefficients for $p = m_1 m_2$ and M_c are 0.99993 and 0.99997, respectively. In the three panels in the middle we plot m_2, M_c and q over m_1 for the injected, recovered and predicted quantities. Finally, in the first two panels of the last row we show the (indirectly) predicted masses with the (p, M_c) regression versus the injected masses, while in the last two panels of the last row we show the predicted masses obtained with the standard regression of Sec. V versus the injected ones.