Deep Active Learning for Detection of Mercury's Bow Shock and Magnetopause Crossings

Sahib Julka[1] (✉), Nikolas Kirschstein[1], Michael Granitzer[1], Alexander Lavrukhin[2], and Ute Amerstorfer[3]

[1] University of Passau, Germany
{sahib.julka,michael.granitzer}@uni-passau.de,
nikolas.kirschstein@gmail.com
[2] Skobeltsyn Institute of Nuclear Physics, LMSU, Moscow, Russian Federation
lavrukhin@physics.msu.ru
[3] Space Research Institute, Austrian Academy of Sciences, Graz, Austria
ute.amerstorfer@oeaw.ac.at

Abstract. Accurate and timely detection of bow shock and magnetopause crossings is essential for understanding the dynamics of a planet's magnetosphere. However, for Mercury, due to the variable nature of its magnetosphere, this remains a challenging task. Existing approaches based on geometric equations only provide average boundary shapes, and can be hard to generalise to environments with variable conditions. On the other hand, data-driven methods require large amounts of annotated data to account for variations, which can scale up the costs quickly. We propose to solve this problem with machine learning. To this end, we introduce a suitable dataset, prepared by processing raw measurements from NASA's MESSENGER [4] mission and design a five-class supervised learning problem. We perform an architectural search to find a suitable model, and report our best model, a Convolutional Recurrent Neural Network (CRNN), achieves a macro F1 score of 0.82 with accuracies of approximately 80 % and 88 % on the bow shock and magnetopause crossings, respectively. Further, we introduce an approach based on active learning that includes only the most informative orbits from the MESSENGER dataset measured by Shannon entropy. We observe that by employing this technique, the model is able to obtain near maximal information gain by training on just two Mercury years worth of data, which is about 10 % of the entire dataset. This has the potential to significantly reduce the need for manual labeling. This work sets the ground for future machine learning endeavors in this direction and may be highly relevant to future missions such as BepiColombo, which is expected to enter orbit around Mercury in December 2025.

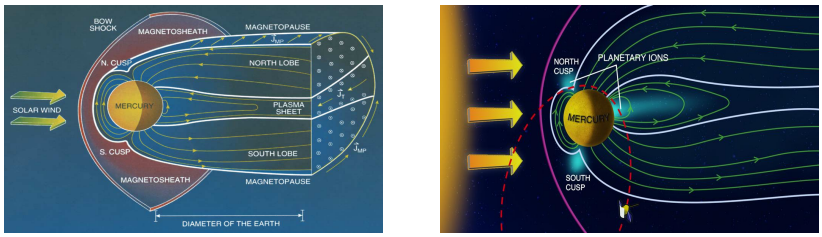
Keywords: · Active Learning · Neural Networks · Magnetosphere

[4] **ME**rcury **S**urface, **S**pace **E**nvironment, **GE**ochemistry, and **R**anging

1 Introduction

The *magnetosphere* of a planet is the region surrounding it where its magnetic field dominates over the magnetic field of the interplanetary space. The *magnetopause* marks the outer boundary of the magnetosphere. Above the magnetopause, lies the *magnetosheath*, which is the region between the magnetopause and the *bow shock* — a shock wave that slows down the approaching supersonic solar wind, and deflects it around the planet's magnetospheric cavity. Principally, the locations and characteristics of these regions around a planet are affected by the varying solar wind conditions [9]. This is particularly the case for Mercury (C.f. Figure 1 (a)), the innermost planet in our solar system. Adding to it, its weak magnetic field — only about 1 % of the Earth's [6], makes the magnetic conditions around the planet even more dynamic, and thus interesting to study. Studying such magnetospheres can yield valuable insights into understanding more complex magnetospheres, such as that of our planet Earth.

It has long been of scientific interest in the planetary science community to study Mercury's bow shock and magnetopause signatures. To this end, NASA launched a space-probe called MESSENGER orbiting Mercury for a long-term empirical study. The relatively small size of Mercury's magnetosphere, an order of magnitude less than the Earth's, allowed the collection of large amounts of data in a significantly shorter time. During the four years of its voyage from 2011 to 2015, the spacecraft completed over 4000 orbits around the planet. As sketched in Figure 1(b), it passed through all the magnetic regions, yielding more than 8000 incidences of bow shock and magnetopause crossings.



(a) (b)

Fig. 1: (a) Schematic view of Mercury's magnetic conditions [22].The bow shock slows down the approaching solar wind to subsonic speeds. The magnetopause further acts as an obstacle. (b) A typical MESSENGER orbit path: the spacecraft passed from the *interplanetary magnetic field* (IMF) through bow shock, magnetosheath, magnetopause and magnetosphere regions of Mercury and then through the same sequence in reverse [26].

Based on the data from the MESSENGER magnetometer, several studies proposed geometric models of Mercury's magnetosphere [12,24,25,17]. However, due to their global and static nature, they could only provide an average shape

of the bow shock and magnetopause boundaries. The respective authors found that the models struggle to capture the many fluctuations and nuances necessary to generalise to all events. This issue may successfully be tackled by employing data-driven statistical machine learning techniques. Given sufficient data, deep neural networks have shown increasing promise in approximately modelling any distribution, and have successfully been applied to complex tasks relating to event detection, including but not limited to rare event detection in audio signals [5,4] and images [13]. The problem of detecting boundary crossings in a continuous stream of magnetic flux data could be viewed similarly.

The planetary science community recognises the importance of this paradigm shift [16]. We follow suit and propose to solve this problem, as a first step, in a supervised deep learning setting. However, supervised learning requires a suitable dataset and expert annotations. As this effort can get very costly given the usually large amounts of unlabelled data in planetary sciences, it is prudent to only annotate the most useful samples. Active learning can facilitate efficient manual labeling by taking classifier specifics into account. However, it may not necessarily be useful in all domain contexts. In planetary science, however, the problem domain and data gathering context could provide an important frame for devising a domain-specific active learning strategy.

In particular, it is reasonable to assume that different orbits may exhibit similarities in their magnetic field structure, yet at the same time at least one entire Mercury year would be necessary to capture all seasonal nuances. It remains, however, unknown how the inter-orbital year distributions vary, and thus questions such as what is the lower bound on number of orbits required to obtain a near maximum informational gain remain open. In this regard, we examine how the model performance scales with available data on orbit-level. Further, we consider it necessary not only to accurately classify and localise the crossing timestamps, but also to classify ahead in time, which would be highly beneficial for tasks such as instrument parameter adjustment, during real time use. More precisely, our contributions can be summarised as follows:

1. We introduce a dataset suited to machine learning tasks and make it available open source: `https://github.com/epn-ml/messenger-prep`
2. We conduct an architectural study to investigate the applicability of data-driven neural networks by using just magnetometer data, without the solar wind conditions, and to identify some best practices.
3. We devise a domain-specific active learning strategy and investigate how many Mercury years' worth of data are required for a sufficiently representative model.
4. We provide a high-quality codebase that may be used as a framework for further studies on neural detection of bow shock and magnetopause crossings. It is publicly available at: `https://github.com/epn-ml/Freddie`

2 Related Work

The task of modelling the boundary crossings is not new. Naturally, Earth has had the lion's share of related work as evidenced by [21,20,14,23]. This enabled subsequent studies investigating various structural and statistical properties of the magnetopause [10]. The empirical and statistical studies require that a consistent catalogue of boundary crossings is available from the in situ data. This process has been recognised to be time-consuming, ambiguous and poorly reproducible, and one that would significantly benefit from automation.

To this end, [11] proposed a threshold-based method. However that turned out to be hard to generalise given the different scales and distributions from different missions [15]. In another line of work, models using paraboloids of revolution with variable flaring angles are explored for Mercury [2], Earth [1], Jupiter [7], and Saturn [3]. These models were obtained by parabolic parameterisation of the magnetopause and bow shock crossing shapes. The averaged boundary shapes can be used as initial parameter values for magnetospheric magnetic field modeling. In this vein, [12] attempted to model Mercury's boundary crossings using such a model. This was followed by [24], where the authors explored the applicability of hyperboloids and a figure similar to the Earth's magnetopause shape, and also [25] whose authors modelled it as a three dimensional non-axially symmetric shape. Philpott et. al [17] extended the aforementioned studies using a combination of an axisymmetric shape and a three-dimensional shape with indentations in the cusp regions and a magnetotail that is wider in the north-south versus east-west direction.

All these approaches share the drawback of applying static models that cannot capture variable conditions in the environment, since they propose a fixed geometric shape cemented for all times. We utilise the boundary crossing catalogue provided by Philpott et. al as approximate guides for supervised deep learning. This is particularly useful to test our active learning strategy so in the future works this is suited to a semi-supervised setup, where only the most necessary samples are required to be annotated by the domain expert.

3 Dataset

As the Fast Imaging Plasma Spectrometer (FIPS) on MESSENGER was not equipped to capture the solar wind data, there are no in situ estimates of solar wind parameters controlling the solar wind dynamic pressure which in part determines the position and the flaring angle of the bow shock and magnetopause boundaries. Thus, we are limited to features based on magnetic field measurements only. We chiefly use Reduced Data Record (RDR) data products of the MESSENGER MAG magnetometer instrument, obtained from the NASA PDS PPI repository, and process it in the follwing manner: First, we remove the calibration signals, in order to not be biased by them. Next, we enrich the dataset with Mercury position information. Then, to prepare the data indexed on orbit

boundaries, we split them based on UTC-based day boundaries, and MESSENGER orbit apoapsis [5] points as markers to separate individual orbits. To simplify subsequent analysis, we also include the estimated planetary dipole magnetic field contribution for each point, planetocentric distance of the spacecraft, and recalculate position and magnetic field data in the aberrated MSO coordinate system which accounts for the non-negligible orbital velocity of Mercury relative to the speed of the solar wind. For more details and links to original sources, please refer to the dataset repository.

Consequently, we obtain a prepared dataset comprising of 4049 orbits. Additionally, we perform a few more removal steps, specific to our pipeline in this work: (a) Missing values: Some of the orbits lack individual measurements or even entire time steps. We conveniently remove those orbits, instead of correcting or filling with interpolation. (b) Overhanging crossings: Some orbits have crossings that extend into neighboring orbits or vice versa. After the cleaning step, there remain 2776 orbits. We randomly split these orbits into *training*, *validation* and *test* sets with a 70-20-10 percentage split, and normalise using Z-score standardisation.

Finally, we leverage the crossing annotations by Philpott et al. [17] visualised in Figure 2, to assign each time step a magnetic region. This yields the class distribution shown in table 1, which exhibits a significant imbalance that we address later.

Table 1: Class labels with their abbreviations and frequency of occurence. The boundary classes are highly underrepresented.

label	**magnetic region**	**share**
0	interplanetary magnetic field (IMF)	64.8 %
1	bow shock crossing (SK)	3.7 %
2	magnetosheath (MSh)	14.8 %
3	magnetopause crossing (MP)	2.3 %
4	magnetosphere (MSp)	14.4 %

4 Methodology

4.1 Problem Formulation

To obtain aggregates, and have an augmented set of fixed shaped input vectors, we use a sliding window. It has a stride of one, which ensures each time step of the original series is contained in multiple windows, such that the crossings can be presented to the model in all possible arrangements, to account for translation-equivariance [6]. Hence, the model's input is a window of $w \in \mathbb{N}$ successive time

[5] The apoapsis of an elliptic orbit is the point farthest away from the planet.
[6] The position of the event in the window should not matter.

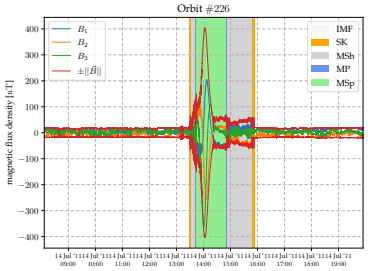


Fig. 2: Example annotation for orbit #226 (best viewed in colour). Annotations mark the start and end of a magnetic region. We label the entire region inside as belonging to the respective crossing. Each crossing appears twice in an orbit.

steps. Each of these time steps consists of $d \in \mathbb{N}$ scalar features. We abstractly represent the input window as follows:

$$\boldsymbol{X} \coloneqq \left[\boldsymbol{x}^{(1)} \; \boldsymbol{x}^{(2)} \; \cdots \; \boldsymbol{x}^{(w)} \right] \in \mathbb{R}^{d \times w}$$

Consider the last time step $\boldsymbol{x}^{(w)}$ in a window as representing the 'present'. Instead of merely classifying the 'past' time steps within a window, we also seek to compute predictions on the magnetic region for $f \in \mathbb{N}$ future time steps. Therefore, we expect the output per time step as *one-hot* vectors. As we expect the model to predict a class per time step, we pack multiple of these one-hot vectors next to each other into a matrix. Thus, the target output matrix of one-hot vectors is $\boldsymbol{Y} \in \mathbb{R}^{5 \times (w+f)}$.

Formally, the task can be framed as a multi-dimensional multi-class classification with a future component: Given the window $\boldsymbol{X}$, we predict a sequence of magnetic region probabilities, where each column sums up to one:

$$\hat{\boldsymbol{Y}} \coloneqq \left[\begin{array}{ccc|ccc} p_{1,1} & \cdots & p_{1,w} & p_{1,w+1} & \cdots & p_{1,w+f} \\ \vdots & & \vdots & \vdots & & \vdots \\ p_{5,1} & \cdots & p_{5,w} & p_{5,w+1} & \cdots & p_{5,w+f} \end{array}\right] \in [0,1]^{5 \times (w+f)}$$

With the setup ready, the normalised vectors are then passed through the neural networks, and the final activations can be represented as: $\boldsymbol{\gamma}_\theta : p_{ij} = \boldsymbol{\gamma}_\theta(\boldsymbol{X})$ where $\boldsymbol{\gamma}$ is a chosen model, with θ as its parameters. We experimentally find a window size of two minutes, i.e., $w = 120$, to be both practical and computationally kind, and fix the future size to $f = 20$ seconds. The selected features include the 3 three-dimensional features, namely MSO position, flux density and measurement errors, chosen via manual tuning on the evaluation split, resulting in dimensionality $d = 9$.

To measure the error between the prediction $\hat{\boldsymbol{Y}}$ and the ground truth $\boldsymbol{Y}$, we employ the standard categorical cross-entropy loss. Counteracting the considerable class imbalance inherent in the dataset, we weight each class inversely proportional to its frequency $f_c \in \mathbb{N}$ in the dataset by virtue of $w_c \coloneqq (\sum_{i=1}^{5} f_i)/f_c$.

The resulting weighted loss for a single time step j in a window is then

$$\mathcal{L}_j(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) \coloneqq -\sum_{i=1}^{5} w_i \boldsymbol{Y}_{ij} \log(\hat{\boldsymbol{Y}}_{ij})$$

Note that the sum is only a formal construct, since exactly one of the $\boldsymbol{Y}_{ij}$ for fixed j is non-zero. By averaging across all time steps in a window, we straightforwardly obtain the window loss:

$$\mathcal{L}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) \coloneqq \frac{1}{w+f} \sum_{j=1}^{w+f} \mathcal{L}_j(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = -\frac{1}{w+f} \sum_{i=1}^{5} w_i \sum_{j=1}^{w+f} \boldsymbol{Y}_{ij} \log(\hat{\boldsymbol{Y}}_{ij})$$

Finally, the average loss over all windows extracted from the training set forms the overall optimisation target.

4.2 Model Architectures

As a first step in a feasibility study for model selection, we consider a total of six architecture categories, namely: Multi Layer Perceptron (MLP), Convolutional Neural Network (CNN), Fully Convolutional Neural Network (FCNN), Recurrent Neural Network (RNN), Convolutional Recurrent Neural Network (CRNN), and Convolutional Attentional Neural Network (CANN). The reader is encouraged to refer to the code repository for specific implementation details. While our architecture search space is biased towards shallower models, we are only concerned with their relative performance, and by interpreting Table 2 find the CRNN to be a suitable candidate for further experimentation.

4.3 Active Learning

For our active learning experiment, we exploit the domain specific data gathering properties, particularly that bow shock characteristics differ between orbits. Consequently, we ask the question whether an orbit-level informativeness measure can be constructed to reduce the amount of manual labelling. To evaluate the impact of this orbit-level informativeness measure, we compare the model performance when adding orbits to the training process.

We use an instance of *pool-based* active learning [19]: Initially untrained, the model repeatedly selects samples from a pool of yet unlabeled samples, obtains the labels, and trains incrementally on them. To address our performance scaling question, we increment the training set not by individual windows but on the level of entire orbits. In order to choose the next orbit(s) to add, it is needed that we rank all yet unused orbits according to an *informativeness* measure. Although solely relying on the top uncertain samples could sometimes lead to overfitting [18], since we always add an entire orbit covering all classes, we find this to be non-issue in our study, and for convenience, resort to it.

As our model has a series of Multinoulli distributions for output, we may measure uncertainty as a function of the output probabilities. *Shannon entropy*

is a mathematically well-funded measure of uncertainty in a probability distribution that we utilise as the basis of our active learning strategy: Consider the training set $\mathcal{D} \subseteq \mathbb{R}^{d\times w} \times \{\mathrm{IMF}, \mathrm{SK}, \mathrm{MSh}, \mathrm{MP}, \mathrm{MSp}\}^{w+f}$ with number of features $d \in \mathbb{N}$, window size $w \in \mathbb{N}$ and future size $f \in \mathbb{N}$. Given a model prediction $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}^{(1)}, \ldots, \hat{\boldsymbol{y}}^{(w+f)}] \in [0,1]^{5\times(w+f)}$, we define its uncertainty as

$$\mathfrak{u}(\hat{\boldsymbol{Y}}) \coloneqq \max_j H(\hat{\boldsymbol{y}}^{(j)}) = -\min_j \sum_{i=1}^{5} y_i^{(j)} \log(y_i^{(j)}), \mathfrak{u}(\hat{\boldsymbol{Y}}) \coloneqq \max_j H(\hat{\boldsymbol{y}}^{(j)}) = -\min_j \sum_{i=1}^{5} y_i^{(j)} \log(y_i^{(j)}),$$

where $H : \triangle^4 \to \mathbb{R}$ is the Shannon entropy on the standard 4-simplex [7].

To achieve this on the orbit level, we must reduce the individual window uncertainties to a single orbit score. As we are only interested in the crossings, we can argue that the most uncertain windows of an orbit will usually overlap with a crossing region, and thus for simplicity, only consider the uncertainty of such windows for the overall orbit uncertainty. Let hence $\mathcal{D}_o \subseteq \mathcal{D}$ be the windows belonging to the orbit $o \in \mathbb{N}$ and

$$\widetilde{\mathcal{D}}_o \coloneqq \{(\boldsymbol{X}, \boldsymbol{y}) \in \mathcal{D}_o \mid \boldsymbol{y} \cap \{\mathrm{SK}, \mathrm{MP}\} \neq \emptyset\}$$

be only those samples that overlap with a bow shock or magnetopause boundary region. The average uncertainty over these windows then defines the integrated orbit uncertainty of a model $\hat{f}_{\boldsymbol{\theta}} : \mathbb{R}^{d\times w} \to \mathbb{R}^{d\times(w+f)}$ for our task:

$$\mathfrak{U}_{\hat{f}_{\boldsymbol{\theta}}}(\widetilde{\mathcal{D}}_o) \coloneqq \frac{1}{|\widetilde{\mathcal{D}}_o|} \sum_{(\boldsymbol{X},\boldsymbol{y})\in\widetilde{\mathcal{D}}_o} \mathfrak{u}(\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{X}))$$

Using this uncertainty measure, we formulate our active learning procedure in Algorithm 1. Instead of strictly adding orbits one-by-one, we more generally allow for an *increment function* $\partial : \mathbb{N}_0 \to \mathbb{N}$ that dictates the number of most uncertain orbits to add, depending on the number of already seen orbits.

5 Experiments

5.1 Model Evaluation

We compare the six models listed in section 4.2 as to their classification performance on the test set. To this end, we employ the following metrics: Macro F1, overall accuracy, and the class-wise accuracy for the critical bow shock and magnetopause classes, respectively. Table 2 illustrates results for all the models, with their respective number of trainable parameters as an indicator of their size. We see a clear improvement between the variants with and without the recurrent component. The combination of convolutional and recurrent does noticeably better than either alone, however the contribution is marginal compared to that of RNN alone. The CRNN however achieves the highest overall scores

[7] $\triangle^{n-1} \coloneqq \{(p_1, p_2, \ldots, p_n) \in \mathbb{R}^n \mid \forall i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1\} \subseteq [0,1]^n$

```
/* actively trains the given model on an incrementally
   growing subset of the training data                    */
active_learning(f̂_θ : ℝ^{d×w} → ℝ^{5×(w+f)} : model,
                Ω ⊆ 𝒫(𝒟) : set of all training orbits,
                ∂ : ℕ_0 → ℕ : increment function):
    𝒯 ≔ ∅  // current training orbits
    while |𝒯| < |Ω| :
        U ≔ hash_table()  // empty hash map
        for 𝒟_o ∈ Ω \ 𝒯 do
            U[𝒟_o] ≔ 𝔘_{f̂_θ}(𝒟̃_o)  // determine orbit uncertainty
        𝒯 ≔ 𝒯 ∪ top_k(U, ∂(|𝒯|))  // add ∂(|𝒯|) most uncertain
         orbits
        f̂_θ ≔ train(f̂_θ, 𝒯)  // retrain model on updated set
    return f̂_θ
```

Algorithm 1: Active learning scheme for incrementally adding orbits to the training procedure in a flexible manner.

and the highest magnetopause accuracy. Our experimental CANN, with an attention mechanism, accomplishes almost the same magnetopause performance but lags slightly behind on the overall metrics. Although the CANN achieves a higher bow shock accuracy than the CRNN, we continue our experiments with the latter for its best overall performance.

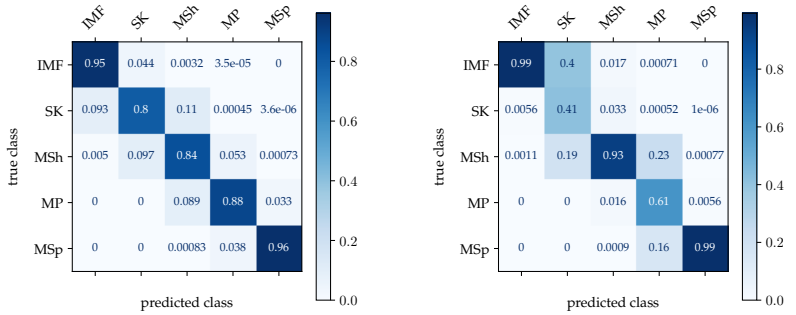
Table 2: Comparison of the model architectures.

model	macro F1	accuracy	SK accur.	MP accur.	# params
MLP	74.73 %	86.60 %	73.87 %	84.05 %	245180
CNN	77.80 %	89.29 %	74.75 %	84.62 %	1413372
FCNN	78.97 %	90.88 %	78.83 %	89.08 %	1444796
RNN	79.93 %	92.03 %	**81.50 %**	91.75 %	237701
CRNN	**81.21 %**	**93.04 %**	79.22 %	**92.22 %**	267333
CANN	80.20%	92.46 %	81.30 %	**92.23 %**	246469

Further, upon evaluating the best model on the test set, we see no real evidence of overfitting (C.f. Table 3), which is a good sign. The model performs better on the relatively easier classes of IMF, magnetosheath and magnetosphere. The confusion matrices in Figure 3 show that the model consistently does better on recall over precision.

Table 3: CRNN performance on the test versus evaluation set.

set	macro F1	accuracy	SK accur.	MP accur.
eval	81.21 %	93.04 %	79.22 %	92.22 %
test	81.95 %	93.13 %	79.93 %	87.51 %



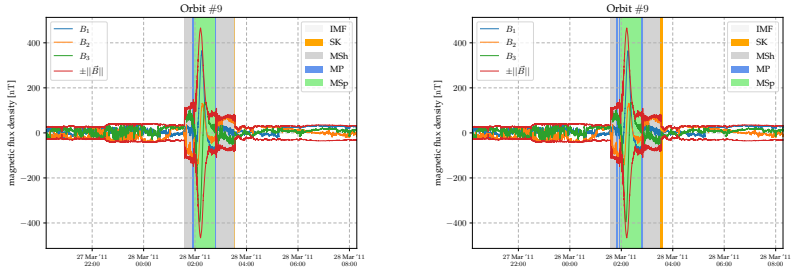
(a) Row-wise normalisation. Diagonal corresponds to recall.

(b) Column-wise normalisation. Diagonal corresponds to precision.

Fig. 3: Normalised confusion matrices for the CRNN. The results indicate applicability for real-time predictions.

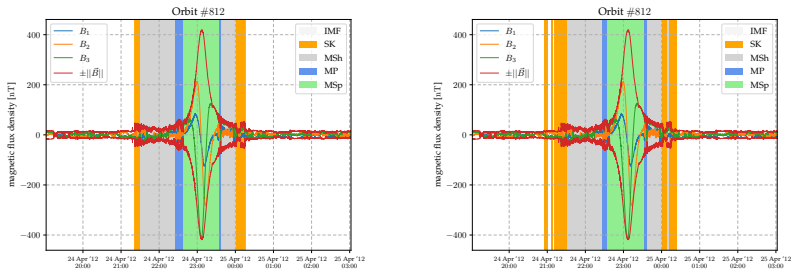
Qualitative Evaluation To confirm our findings, we evaluate the CRNN qualitatively, and utilise its past-only classifications in a window to infer predictions for an entire orbit. Each time step receives distinct predictions from all sliding windows it is contained in, which we integrate by averaging to obtain an overall class probability distribution, and arg max yields a class prediction for the time step. We do this for all orbits in the test set and plot their magnetic flux density along with the predictions.[8] Upon visually inspecting all orbits in the test set, we can confirm that the model overall predicts contiguous magnetic regions. Further, we notice that in some cases, the crossings, albeit exaggerated w.r.t existing ground truth, correctly predicts the boundaries (Figure 4), indicating that it might be learning associations not available explicitly in labels. Although more work needs to be done in this regard, this is very promising as it might lead to explanations that could benefit the physical understanding of certain phenomena. Nevertheless, we also identify some major qualitative issues that still remain, such as scattered predictions, and boundary exaggerations (C.f. Figure 5), some of which may possibly be tackled by solutions that we identify in Section 6.

[8] All plots for the entire test set are made available in the code repository linked in Section 1.



(a) Ground truth (b) Model prediction

Fig. 4: An example prediction where boundaries are slightly exaggerated. The network tries to compensate for the conservative annotation, while yielding a better prediction on the duration of the crossings (best viewed in colour).



(a) Ground truth (b) Model prediction

Fig. 5: An example prediction with significantly exaggerated and scattered bow shock crossing.

5.2 Active Learning

For our active learning experiments, we run Algorithm 1 with two different choices for the increment function: one leading to a constantly growing training set and one leading to an linearly growing training set. In this manner, we explore how the classification performance scales with available data and determine the order of magnitude of orbits required for a sufficiently informed model.

Constant Increment A straightforward choice of increment function would be to add orbits one by one. Due to computational concerns and observed overfitting on single orbits, we found $\partial(n) := 10$ to be more suitable.

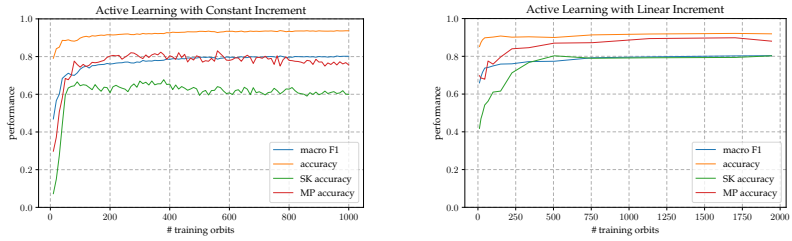
Linear Increment Due to some problems we identified with a constant increment, we conduct another active learning experiment with the linear increment function $\partial(n) := \max\{\lfloor n/2, 10 \rfloor\}$. This choice ensures a constant proportion of 'new' vs 'old' training orbits while preventing overfitting to a single orbit.

Figure 6 plots the evaluation metrics discussed previously over the number of already included orbits for both increment function choices. The learning curve

for the constant increment shows only the first 1000 orbits, as the experiment could not run until completion, but the evolution is clearly evident. In both cases, we observe a rapid increase of all metrics in the beginning, followed by a period of flattening. After no more than 500 orbits, the performance metrics are comparable to those of the passively trained model.

In the constant case, the class accuracies for bow shock and magnetopause later decrease, while the overall metrics continue to rise. This divergence implies that the model focusses more on the majority classes and increasingly ignores the two boundary classes we are concerned with. We suspect the constant increment causes this mediocre development. Since the number of orbits added in each iteration does not depend on the number of already seen orbits, their relative proportion becomes increasingly skewed towards the known orbits. As a result, the marginal returns diminish while learning from new orbits but continues to optimise over the familiar ones repeatedly.

These observations explain our choice of the linear increment function. Indeed, it leads to a much better development while at the same time requiring substantially less iterations and hence computational cost. Due to the latter reason, the improvement is slower in the beginning but reaches far higher scores in the long run. However, they do not surpass the performance of the passively trained model. This indicates that the lower bound on number of orbits required is not too high, further emphasising the need for clever data sampling approaches.

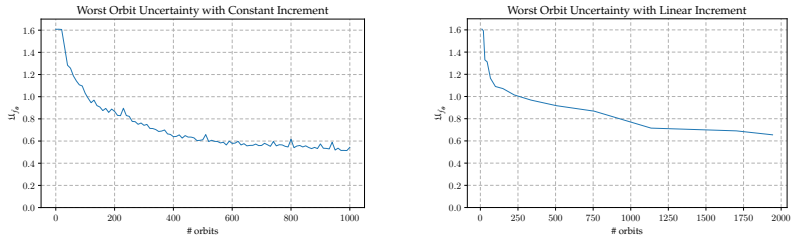


(a) Constant increment (b) Linear increment

Fig. 6: Performance metric development during active learning.

Besides the performance metrics, we also evaluate the development of our uncertainty measure during the active learning process. After all, it is the very measure by which orbits are selected for training in the active learning scheme and indicates the model's confidence about its decisions. We are interested in the point from where the uncertainty does not significantly decrease anymore, implying that the model has nearly saturated its learning capabilities. In a sense, the model has 'seen enough' until that point. Figure 7 plots the worst occurring orbit uncertainty at each iteration as a function of the number of orbits included in the process. Both start in the beginning with a value of just under

$\log(5) \approx 1.609$, which is the entropy of a uniform distribution on the five outcomes. This is not a coincidence but rather results from the definition of our orbit uncertainty measure in Section 4.3 and the model's random parameter initialisation. Then, the orbit uncertainty decreases rapidly during the subsequent iterations. Analogously to the performance metrics in Figure 6, the uncertainty eventually flattens out and seems to almost asymptotically approach values of 0.5 and 0.6 respectively. Again, the maximum marginal improvement appears during the first half, until about 500 orbits are included.



(a) Constant increment (b) Linear increment

Fig. 7: Orbit uncertainty development during active learning.

Taking all insights together, we conclude that the model's learning capacity saturates after 450 to 500 orbits. This constitutes an upper bound for the number of orbits required for a representative model. When summing the duration spanned by the concrete orbits chosen by the model, this equates to roughly two full Mercury years' worth of MESSENGER orbits. We may therefore claim that two Mercury years make for a sufficient set of observations for the model to learn from. On the other hand, revisiting Figure 6 and Figure 7 confirms our intuition that one complete Mercury year (around 230 orbits in this case) is at least required. It remains for future work, hence, to explore the range in between. With the improvements we propose in Section 6, it might even be possible to lower this bound to just one Mercury year.

6 Conclusion

In this work, we built a discriminative end-to-end deep learning model for detecting Mercury's bow shock and magnetopause crossing signatures based on raw measurements from NASA's MESSENGER mission. Additionally, we devised an active learning scheme to address the question of how many orbits worth of measurement data is required for a representative model. To this end, we prepared a dataset suited to machine learning tasks, which we make available publicly to facilitate future research in this direction. To inspect the applicability of machine learning to this problem we formulated a five class supervised machine learning task, that given a window of measurements, predicts the classes for each time

step in the current window, and at the same time predicts the classes the next specified number of time steps. We applied various architectures and configurations of neural networks to determine a suitable fit for the architecture, and observed that the CRNN performs relatively better, which is consistent with findings in other types of signals too, where both spatial and temporal features are of relevance. We also observe that the neural networks are capable of predicting ahead in time, which is a good indication that there might be presence of autoregressive characteristics in the signal. Our best CRNN model achieves a macro F1 of about 82 % and consistently predicts magnetopause crossings better than the bow shock crossings. This is no surprise since the magnetopause crossings are also better discernible to the human eye. Further, the recall scores of 78 % and 86 % on the bow shock and magnetopause crossings respectively, are significantly and consistently better than the precision scores of 39 % and 61 % respectively. There can be several explanations for this: first, the model clearly prefers not missing a boundary at the cost of false positives. Given the use case, it is more important that a boundary is not missed, over exaggerated crossings. Second, the annotations we used are clearly too conservative in many instances, so the network tries to compensate for those based on the learned statistical associations.

Based on the best model, we approached the central question underlying this work with an active learning scheme. It employs the uncertainty sampling strategy with a custom orbit-level measure based on Shannon entropy, by which we iteratively determine the next orbits to include in the training set. After a preliminary experiment with a constantly growing training set, we conducted our main experiment with a linear increment, and observed it to be significantly better. It likely ensures a constant portion of unseen orbits throughout all iterations. Although these strategies might suffer slightly from overfitting on the very first set of orbits, we were able to derive that at least one and at most two Mercury years' worth of measurement data may be sufficient for a representative model that performs reasonably. Finally we recognise while our work yields comprehensive insights into the structure of the MESSENGER magnetometer data and hence the magnetic dynamics around Mercury, it can only provide a starting point in machine learning endeavors.

As part of future work, it would be worthwhile to improve quantitative evaluation by employing metrics that are more sensitive to temporal onsets and offsets. It would also be interesting to investigate if it suffices to let the model predict only one class per window. For inference, this would result in one prediction for each time step instead of multiple votes. This may tackle some of the issues where some crossings are scattered. Likewise, the future classification output may be compressed to a single value. For instance, this could use a binary flag that indicates whether the class predicted for the present time step changes in the near future. It would be useful to explore how concept drift detection techniques help in this regard.

By and large, this work reveals two insights on a broader level: First, deep learning can be used to build sophisticated models of the bow shock and mag-

netopause, a favourable alternative to the existing geometric models that suffer the downside of being static, and often do not accurately predict the duration of the crossings. Second, active learning serves not only for enhancing labeling efficiency but also for addressing data representativeness questions. We strongly encourage future work to continue and improve our study, taking note of the suggestions made above. The outcomes might become relevant for the upcoming Mercury mission BepiColombo [8], which with its twin-aircraft probe will collect significantly more data.

Acknowledgements The authors acknowledge support from *Europlanet 2024 RI* that has received funding from the European Union's *Horizon 2020* research and innovation programme under grant agreement No. 871149.

References

1. Alexeev, I.I., Belenkaya, E.S., Bobrovnikov, S.Y., Kalegaev, V.V.: Modelling of the electromagnetic field in the interplanetary space and in the earth's magnetosphere. Space science reviews **107**(1), 7–26 (2003)
2. Alexeev, I.I., Belenkaya, E.S., Slavin, J.A., Korth, H., Anderson, B.J., Baker, D.N., Boardsen, S.A., Johnson, C.L., Purucker, M.E., Sarantos, M., et al.: Mercury's magnetospheric magnetic field after the first two messenger flybys. Icarus **209**(1), 23–39 (2010)
3. Alexeev, I., Kalegaev, V., Belenkaya, E., Bobrovnikov, S.Y., Bunce, E., Cowley, S., Nichols, J.: A global magnetic model of saturn's magnetosphere and a comparison with cassini soi data. Geophysical research letters **33**(8) (2006)
4. Amiriparian, S., Baird, A., Julka, S., Alcorn, A., Ottl, S., Petrović, S., Ainger, E., Cummins, N., Schuller, B.: Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks (2018)
5. Amiriparian, S., Cummins, N., Julka, S., Schuller, B.: Deep convolutional recurrent neural network for rare acoustic event detection. In: Proc. DAGA. pp. 1522–1525 (2018)
6. Anderson, B.J., Acuña, M.H., Korth, H., Slavin, J.A., Uno, H., Johnson, C.L., Purucker, M.E., Solomon, S.C., Raines, J.M., Zurbuchen, T.H., et al.: The magnetic field of mercury. Space science reviews **152**(1), 307–339 (2010)
7. Belenkaya, E., Bobrovnikov, S.Y., Alexeev, I., Kalegaev, V., Cowley, S.: A model of jupiter's magnetospheric magnetic field with variable magnetopause flaring. Planetary and Space Science **53**(9), 863–872 (2005)
8. Benkhoff, J., Van Casteren, J., Hayakawa, H., Fujimoto, M., Laakso, H., Novara, M., Ferri, P., Middleton, H.R., Ziethe, R.: Bepicolombo—comprehensive exploration of mercury: Mission overview and science goals. Planetary and Space Science **58**(1-2), 2–20 (2010)
9. Fairfield, D.H.: Average and unusual locations of the earth's magnetopause and bow shock. Journal of Geophysical Research **76**(28), 6700–6716 (1971)
10. Haaland, S., Paschmann, G., Øieroset, M., Phan, T., Hasegawa, H., Fuselier, S., Constantinescu, V., Eriksson, S., Trattner, K.J., Fadanelli, S., et al.: Characteristics of the flank magnetopause: Mms results. Journal of Geophysical Research: Space Physics **125**(3), e2019JA027623 (2020)

11. Jelínek, K., Němeček, Z., Šafránková, J.: A new approach to magnetopause and bow shock modeling based on automated region identification. Journal of Geophysical Research: Space Physics **117**(A5) (2012)
12. Johnson, C.L., Purucker, M.E., Korth, H., Anderson, B.J., Winslow, R.M., Al Asad, M.M., Slavin, J.A., Alexeev, I.I., Phillips, R.J., Zuber, M.T., et al.: Messenger observations of mercury's magnetic field structure. Journal of Geophysical Research: Planets **117**(E12) (2012)
13. Kraeft, S.K., Sutherland, R., Gravelin, L., Hu, G.H., Ferland, L.H., Richardson, P., Elias, A., Chen, L.B.: Detection and analysis of cancer cells in blood and bone marrow using a rare event imaging system. Clinical cancer research **6**(2), 434–442 (2000)
14. Lin, R., Zhang, X., Liu, S., Wang, Y., Gong, J.: A three-dimensional asymmetric magnetopause model. Journal of Geophysical Research: Space Physics **115**(A4) (2010)
15. Nguyen, G., Aunai, N., Michotte de Welle, B., Jeandet, A., Fontaine, D.: Automatic detection of the earth bow shock and magnetopause from in-situ data with machine learning. Annales Geophysicae Discussions pp. 1–22 (2019)
16. Nikolaou, N., Waldmann, I.P., Tsiaras, A., Morvan, M., Edwards, B., Yip, K.H., Tinetti, G., Sarkar, S., Dawson, J.M., Borisov, V., et al.: Lessons learned from the 1st ariel machine learning challenge: Correcting transiting exoplanet light curves for stellar spots. arXiv preprint arXiv:2010.15996 (2020)
17. Philpott, L.C., Johnson, C.L., Anderson, B.J., Winslow, R.M.: The shape of mercury's magnetopause: The picture from messenger magnetometer observations and future prospects for bepicolombo. Journal of Geophysical Research: Space Physics **125**(5), e2019JA027544 (2020)
18. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM Computing Surveys (CSUR) **54**(9), 1–40 (2021)
19. Settles, B.: Active learning. Synthesis lectures on artificial intelligence and machine learning **6**(1), 1–114 (2012)
20. Shue, J.H., Chao, J., Fu, H., Russell, C., Song, P., Khurana, K., Singer, H.: A new functional form to study the solar wind control of the magnetopause size and shape. Journal of Geophysical Research: Space Physics **102**(A5), 9497–9511 (1997)
21. Sibeck, D.G., Lopez, R., Roelof, E.C.: Solar wind control of the magnetopause shape, location, and motion. Journal of Geophysical Research: Space Physics **96**(A4), 5489–5495 (1991)
22. Slavin, J.A.: Mercury's magnetosphere. Advances in Space Research **33**(11), 1859–1874 (2004)
23. Wang, Y., Sibeck, D., Merka, J., Boardsen, S., Karimabadi, H., Sipes, T., Šafránková, J., Jelínek, K., Lin, R.: A new three-dimensional magnetopause model with a support vector regression machine and a large database of multiple spacecraft observations. Journal of Geophysical Research: Space Physics **118**(5), 2173–2184 (2013)
24. Winslow, R.M., Anderson, B.J., Johnson, C.L., Slavin, J.A., Korth, H., Purucker, M.E., Baker, D.N., Solomon, S.C.: Mercury's magnetopause and bow shock from messenger magnetometer observations. Journal of Geophysical Research: Space Physics **118**(5), 2213–2227 (2013)
25. Zhong, J., Wan, W., Slavin, J., Wei, Y., Lin, R., Chai, L., Raines, J., Rong, Z., Han, X.: Mercury's three-dimensional asymmetric magnetopause. Journal of Geophysical Research: Space Physics **120**(9), 7658–7671 (2015)

26. Zurbuchen, T.H., Raines, J.M., Slavin, J.A., Gershman, D.J., Gilbert, J.A., Gloeckler, G., Anderson, B.J., Baker, D.N., Korth, H., Krimigis, S.M., et al.: Messenger observations of the spatial distribution of planetary ions near mercury. Science **333**(6051), 1862–1865 (2011)