Physics-informed and Unsupervised Riemannian Domain Adaptation for Machine Learning on Heterogeneous EEG Datasets

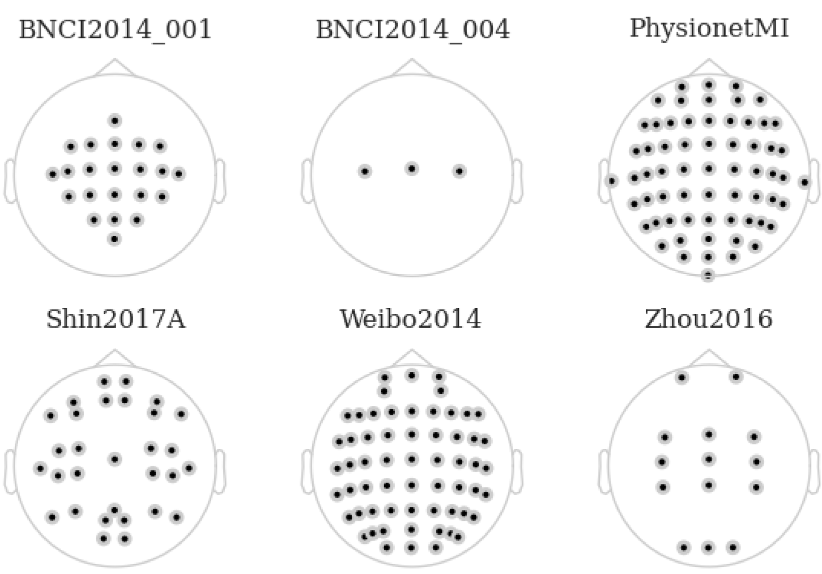
Apolline Mellot[1], Antoine Collas[1], Sylvain Chevallier[2], Denis Engemann[3], Alexandre Gramfort[1]
[1] University Paris-Saclay, Inria, CEA, Palaiseau, France. [2] TAU Inria, LISN-CNRS, University Paris-Saclay, France. [3] Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd., Basel, Switzerland. email: apolline.mellot@inria.fr

How to combine EEG datasets with different electrode configurations?

EEG signals are multivariate time series $\boldsymbol{X} \in \mathbb{R}^{P\times T}$ recorded with P sensors at T time steps. In this work we represent the EEG signal by its spatial covariance matrix $\boldsymbol{C} \in \mathbb{R}^{P\times P}$.

Causes of variability in EEG data:

- Subject and population differences: size of the head, age, body posture, individual brain anatomy...
- Recording devices: electrode type, number and location on the scalp, amplifier conditions...
- Experimental protocol: task performed during recording, eyes closed or open...



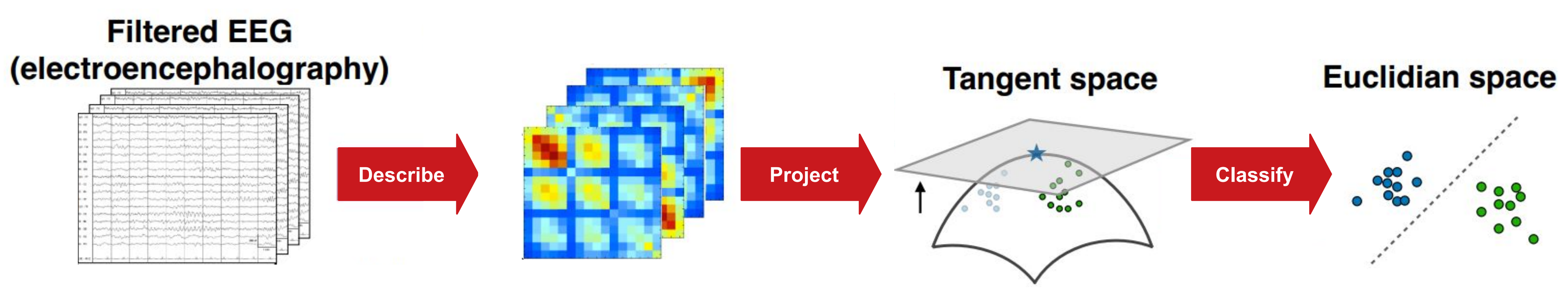
All these differences can lead to shifts in the data distributions, referred to as dataset shift [2]. Thus, a machine learning model trained on a dataset is not directly generalizable to a new datasets recorded in a different context.

→ **Focus:** Find a way to combine EEG datasets with different number of electrodes and varying positions, specifically when the number of common channels across datasets is insufficient.

Domain adaptation framework: We consider M datasets with different numbers of sensors P_j with $j = 1, \dots, M$. We aim to train a model on $M-1$ datasets, called the source datasets, and test it on the left-out dataset, called the target dataset.

Processing pipeline

Covariance-based classification pipeline [1]:



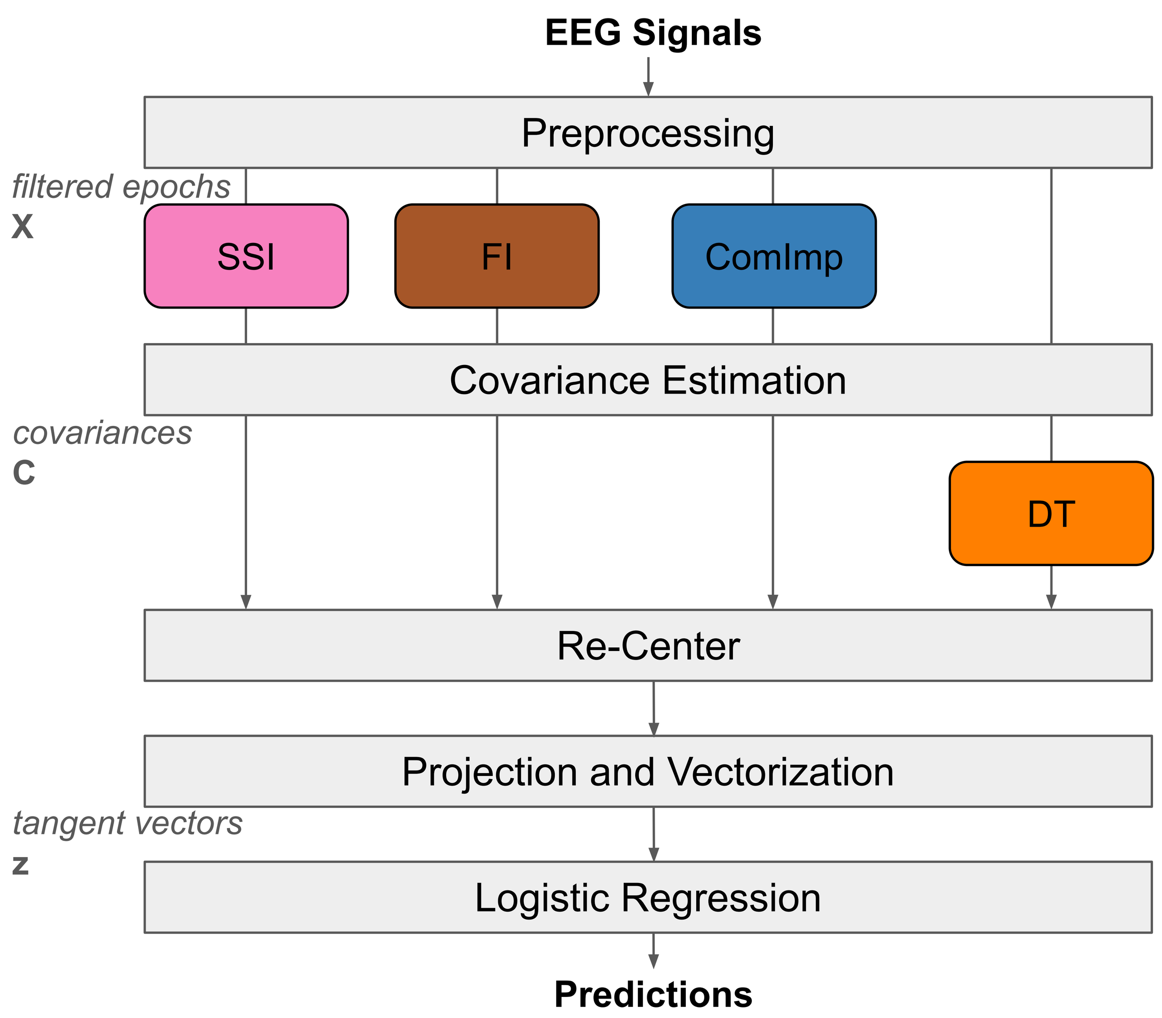
Re-center: In covariance-based BCI classification, the preferred transfer learning technique is to re-center every dataset to a common reference point on the manifold [7]:

$$\boldsymbol{C}^{(\text{rct})} = \overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}\overline{\boldsymbol{C}}^{-1/2} \quad (1)$$

Baseline methods:

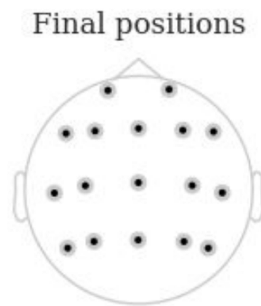
- Common channel selection
- Dimensionality Transcending (DT) [6]: geometry-based imputation
- ComImp [4]: signal-based imputation

Overall pipeline:



Proposed approach: field interpolation

Common usage: to reconstruct the signal of malfunctioning or too noisy channels.
Our idea: to use interpolation to map EEG signals from different electrode configurations to a fixed template of positions.
Principle of interpolation: A linear operator $\boldsymbol{A} \in \mathbb{R}^{P\times P_j}$ is constructed to map the P_j existing EEG channels to the P positions of the final template:



$$\hat{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{X} \quad (2)$$

with $\boldsymbol{X} \in \mathbb{R}^{P_j\times T}$ the recorded EEG signals and $\hat{\boldsymbol{X}} \in \mathbb{R}^{P\times T}$ the reconstructed signals.

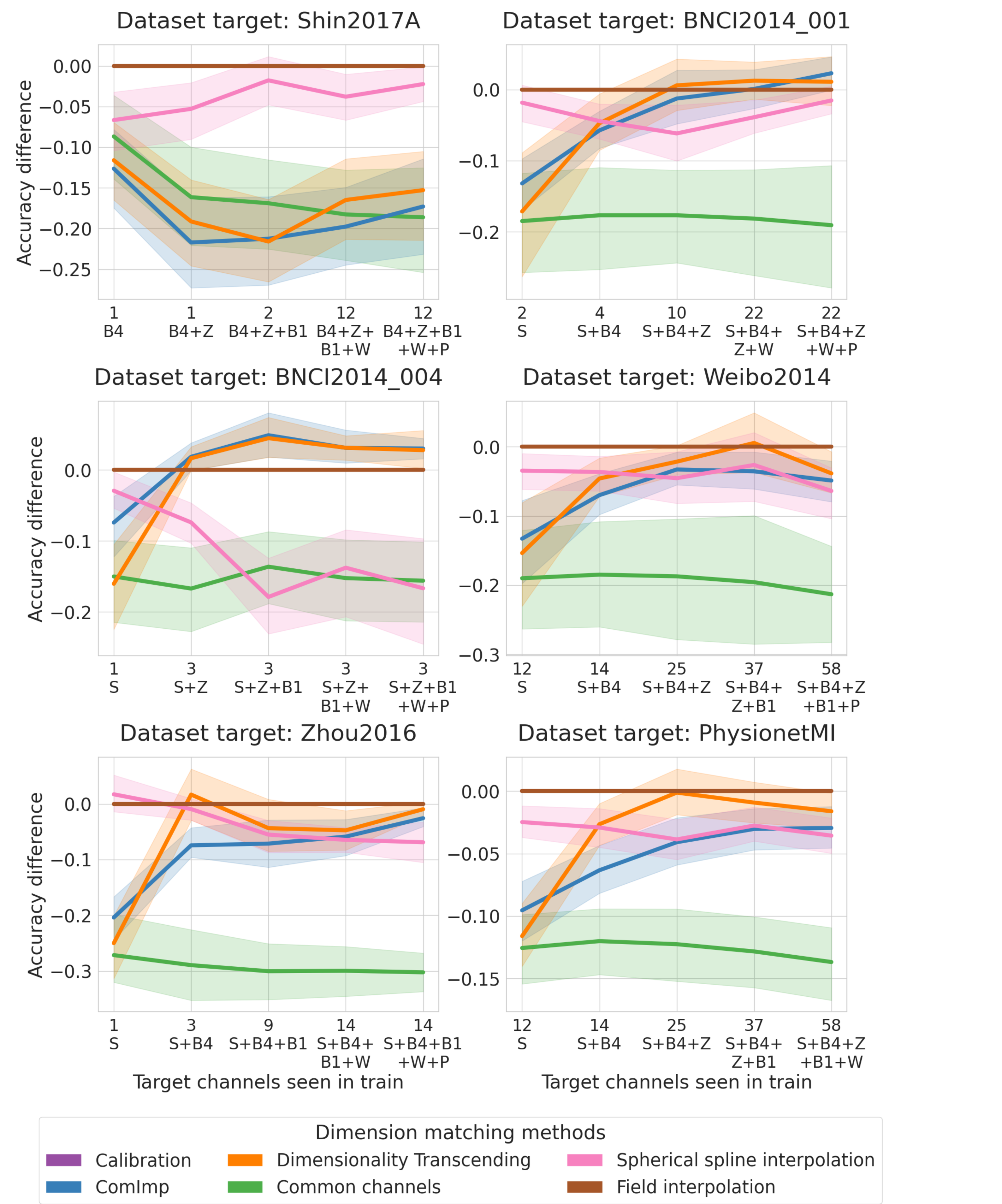
Two techniques:

- Spherical spline interpolation (SSI): projects positions onto a unit sphere and uses smooth functions to interpolate the starting positions to the final ones [5].
- Field interpolation (FI): estimates the generators activity in the brain and maps them to the final positions using a forward model based on Maxwell's equations [3].

Empirical benchmark

Datasets: 6 public BCI datasets, right hand/left hand classification.

Leave-one-dataset-out validation: Each plot corresponds to one target left-out dataset. The other datasets were successively combined to form the training set in order to have an increasing number of target channels seen in train.



Conclusion

- FI outperformed other approaches when few common channels are shared between source and target.
- FI performed similarly to other methods when a large variety of data is available.
- Interpolation can be applied to raw data before feature extraction.

[1] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.

[2] Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 2021.

[3] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroinformatics*, 7:267, 2013.

[4] Thu Nguyen, Rabindra Khadka, Nhan Phan, Anis Yazidi, Pål Halvorsen, and Michael A Riegler. Combining datasets to increase the number of samples and improve model fitting. *arXiv preprint arXiv:2210.05165*, 2022.

[5] François Perrin, Jacques Pernier, Olivier Bertrand, and Jean Francois Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.

[6] Pedro LC Rodrigues, Marco Congedo, and Christian Jutten. Dimensionality transcending: a method for merging BCI datasets with different dimensionalities. *IEEE Transactions on Biomedical Engineering*, 68(2):673–684, 2020.

[7] Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2017.