Evaluating Tangent Spaces, Distances, and Deep Learning Models to Develop Classifiers for Brain Connectivity Data

Michael Wang

Master's Thesis Defense - August 27, 2020

CONNplexity Lab - Purdue University

Committee Members: Dr. Joaquín Goñi, Dr. Mario Ventresca, Dr. Juan Wachs

Presentation Outline

1. Introduction and Background
2. Methods
    - Dataset
    - Post-processing Methods
    - Classifiers
3. Results
    - Task Identification
    - Twin and Subject Identification
4. Discussion

Introduction and Background

Introduction to fMRI

- **F**unctional **M**agnetic **R**esonance **I**maging
- Non-invasive neuroimaging technique
- Very high spatial resolution[6]

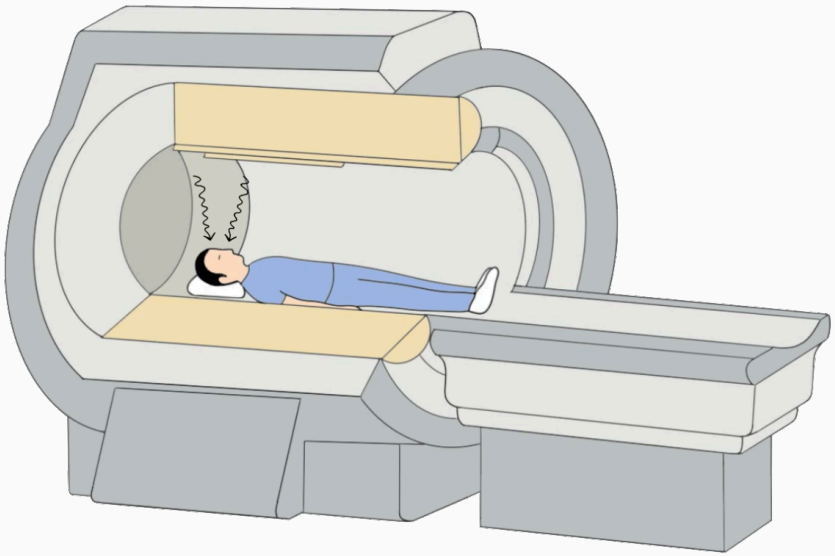


Figure 1: fMRI Scanner

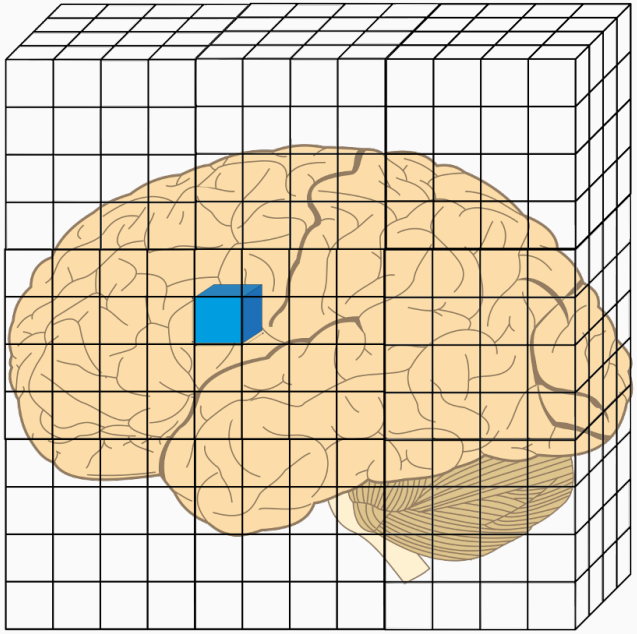
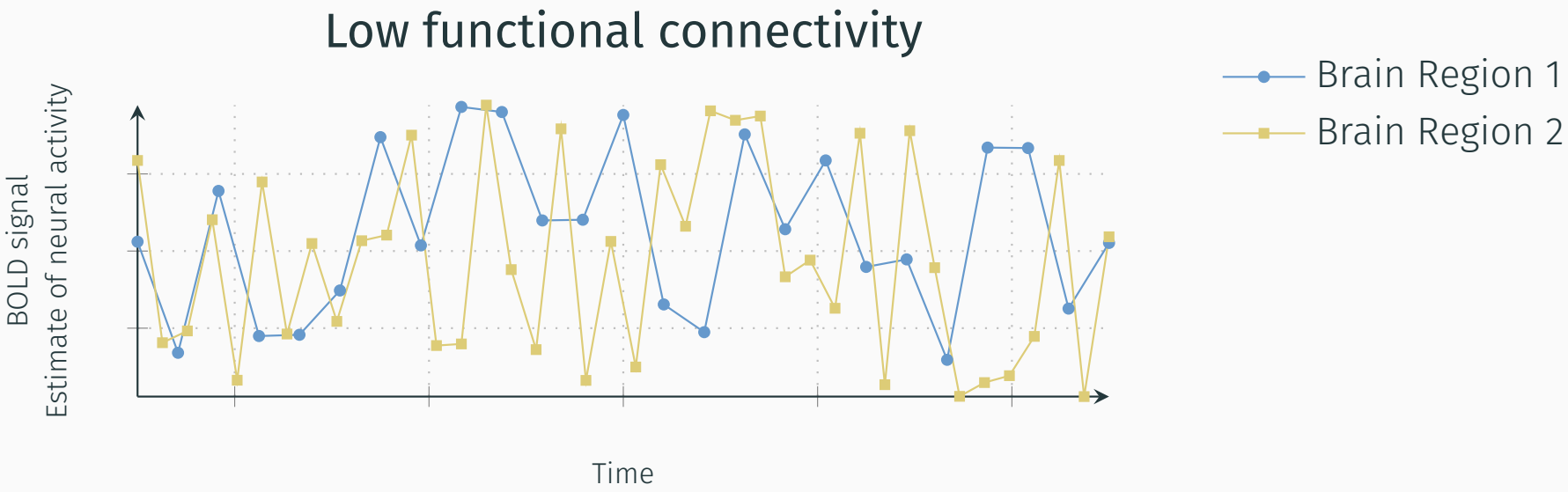
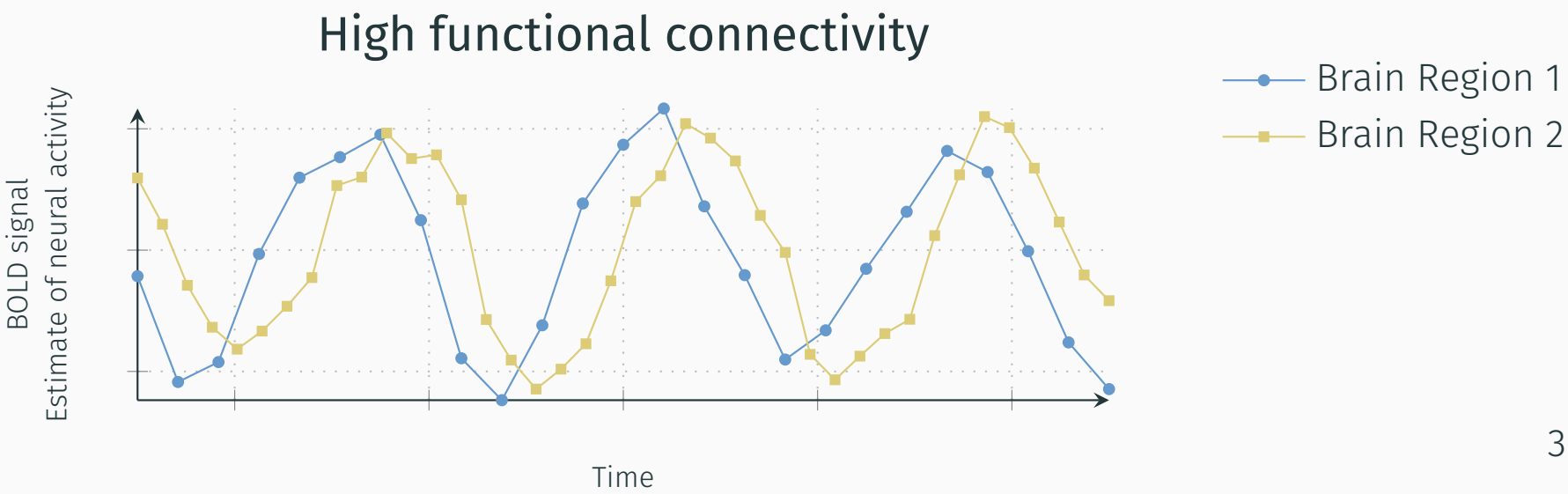


Figure 2: 1mm^3 Voxels

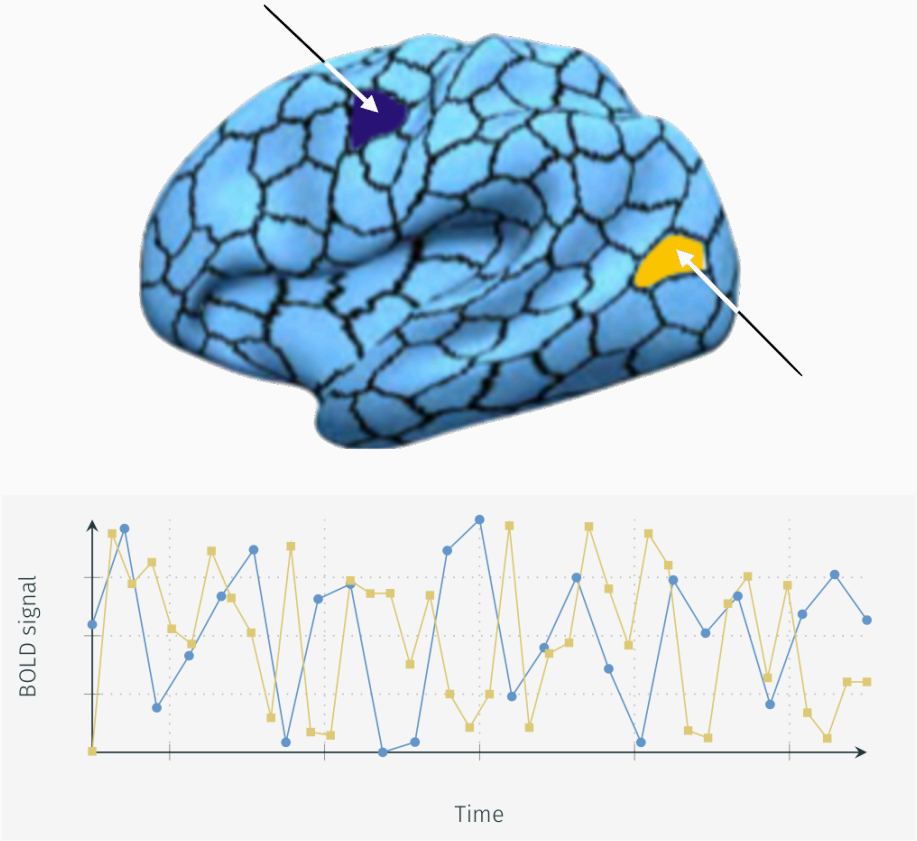
Graphic obtained from 2 Minute Neuroscience

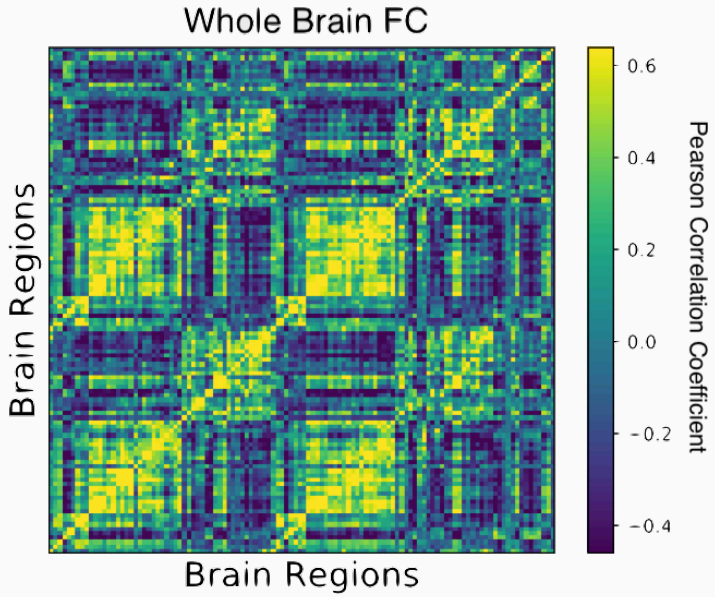
Example of Low and High Correlations of BOLD Time Series





Example Estimation of a Subject's Functional Connectome





Previous Work

- PCA reconstruction of functional connectivity (FC) improves subject-level identifiability[3]
- Geodesic distance, unlike Euclidean distance, preserves FC geometry and improves predictive power.[11]
- Deep Neural Networks have achieved a task identification rate of 0.937 in HCP 8-task fMRI classification.[12]
- Support Vector Machines have achieved MZ and DZ twin identification rate of 0.64 and 0.22, respectively.[5]

Research Motivation and Question

Motivation

- Absence of a gold standard pipeline for analyzing functional connectivity[8]
- Clinical application of fMRI is very limited[10]
- Recent data frameworks for functional connectivity have shown to increase predictive power[2, 11]
- The explosion of publicly available neuroimaging datasets has made deep learning viable[8]

Research Question

- How can we best process and project functional connectivity data for task and twin identification?

Monozygotic (MZ) and Dizygotic (DZ) Twins

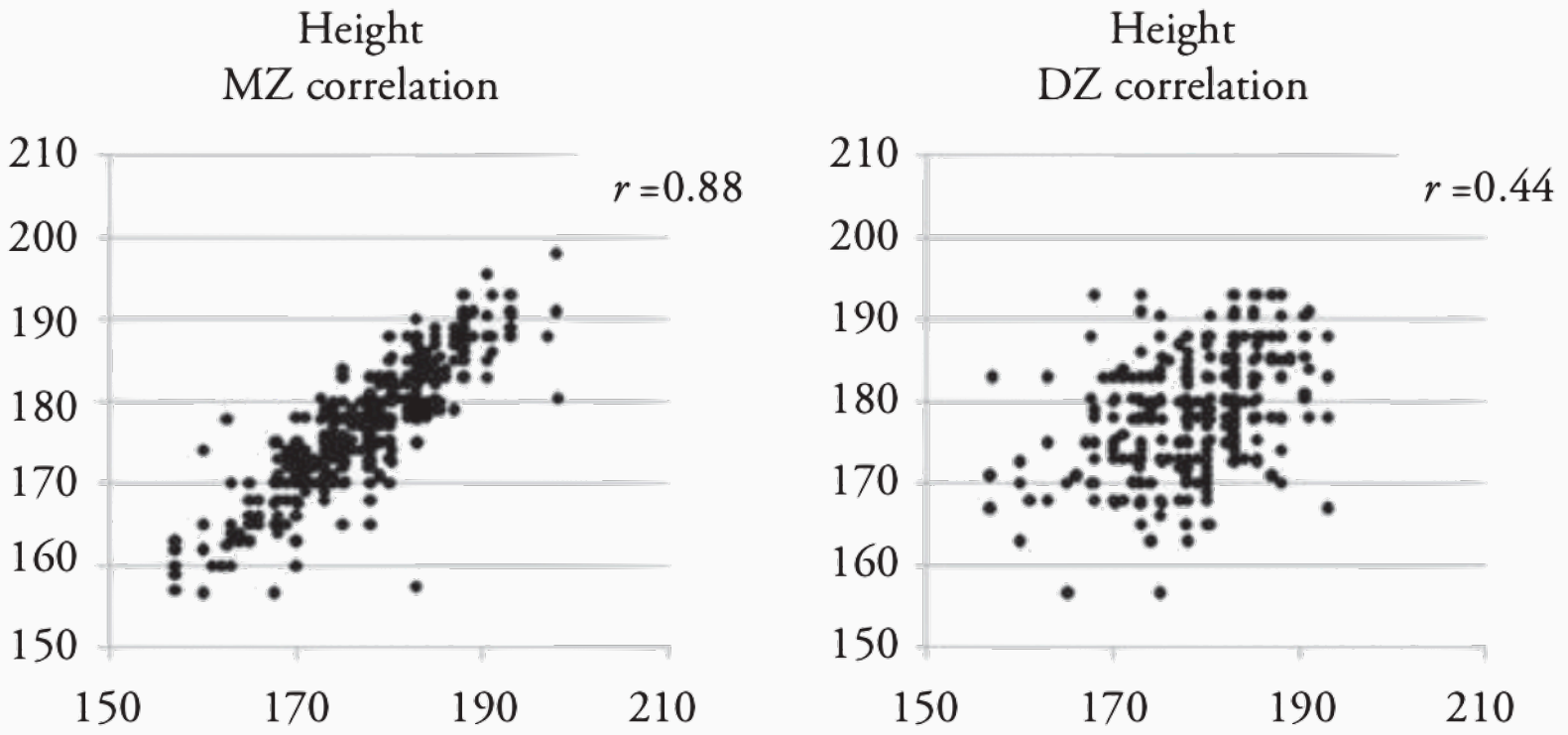


Figure 3: Scatter plot of monozygotic (MZ) and digyzotic (DZ) twin pairfor height (cm) in males. MZ twins share 100% of their parents' genetic material, DZ twins share 50% of genetic material, and unrelated subjects share 0% of the genetic material. MZ and DZ twins also grow up in similar environments.

Methods

Dataset and Minimal Pre-processing

- Human Connectome Project (HCP) 1200 Subjects Release[7]
- HCP functional minimal preprocessing pipeline[4]
  - Artifact removal
  - Motion correction
- 8 cognitive tasks
  - Resting State
  - Working Memory
  - Relational
  - Social
  - Gambling
  - Motor
  - Language
  - Emotion
- Test and retest scans per subject, per task

Task & Twin Subsets of HCP 1200 Subjects Release

Task Identification Dataset

- 424 unrelated subjects
- 8 cognitive tasks
- Total of $424 \times 8 \times 2 =$ **6784 fMRI scans**

Twin Identification Dataset

- 106 pairs of monozygotic (MZ) twins across 8 tasks
- Total of $106 \times 2 \times 8 \times 2 =$ **3392 fMRI scans**
- 58 common dizygotic (DZ) twin pairs across 8 tasks
- Total of $58 \times 2 \times 8 \times 2 =$ **1856 fMRI scans**

Experimental Design

Post-processing Methods

- PCA Reconstruction
- Tangent Space Projection
- Brain Parcellation Granularity

Classifiers

- Convolutional Neural Network
- K-Nearest Neighbors

Adaptation of the Differential Identifiability Framework[3]

- PCA decomposition into orthogonal eigenvectors W sorted by eigenvalues in decreasing explained variance
- Project centered data onto eigenvectors to create PCs Z
- Reconstruct data using subset of *k* PCs and eigenvectors

$$\underset{m\times n}{\mathrm{X_r}} = \underset{m\times k}{\mathrm{W}} \times \underset{k\times n}{\mathrm{Z'}} + \mu$$

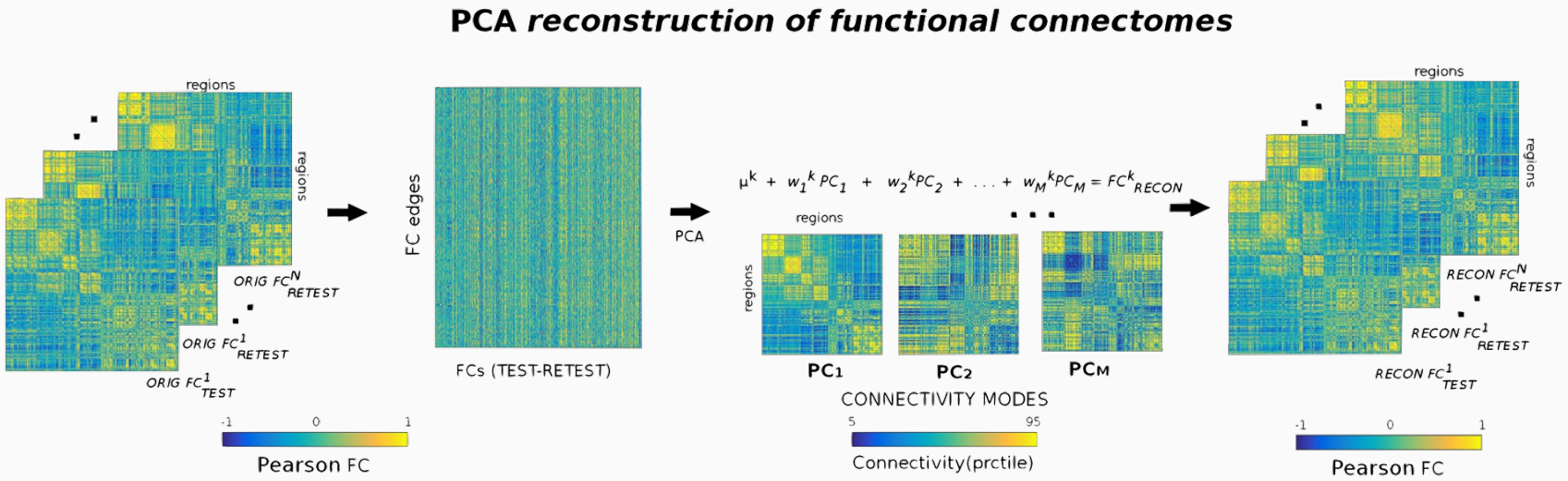


Figure obtained from Amico & Goñi, 2018[3]

Significance of Geodesic Distance

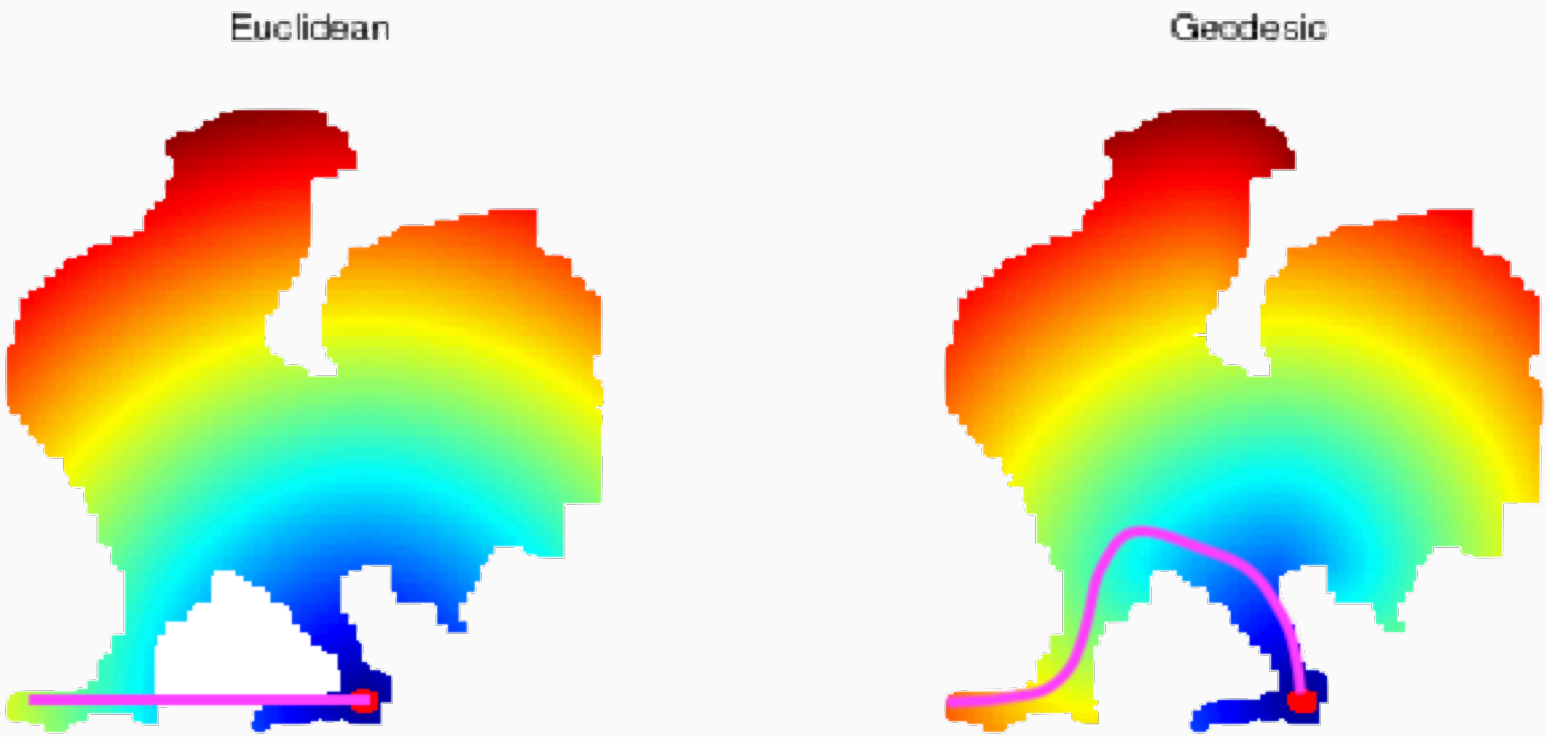


Figure 4: Visualization of Euclidean and geodesic distance between two points on a chicken-shaped surface

Figure obtained from https://dsp.stackexchange.com/

Tangent Space Projection

- FC matrices are positive definite (PD)
- PD matrices do not form a Euclidean space[1]
- Project FC into tangent space to preserve geometry:

$$\hat{\mathrm{C}} = \log_m \left(\mathrm{C}_g^{-\frac{1}{2}} \mathrm{C} \mathrm{C}_g^{-\frac{1}{2}} \right)$$

Table 1: Reference Matrices C_g

Reference	Equation
Euclidean	$\frac{1}{N} \sum_i C_i$
Harmonic	$(\frac{1}{N} \sum_i C_i^{-1})^{-1}$
LogEuclid	$\exp_m(\frac{1}{N} \sum_i \log_m C_i)$
Kullback	$C_e^{\frac{1}{2}} (C_e^{-\frac{1}{2}} C_h C_e^{-\frac{1}{2}})^{\alpha} C_e^{\frac{1}{2}}$
Riemmanian	$\arg\min(\sum_i \delta_R(C_e C_i)^2)$

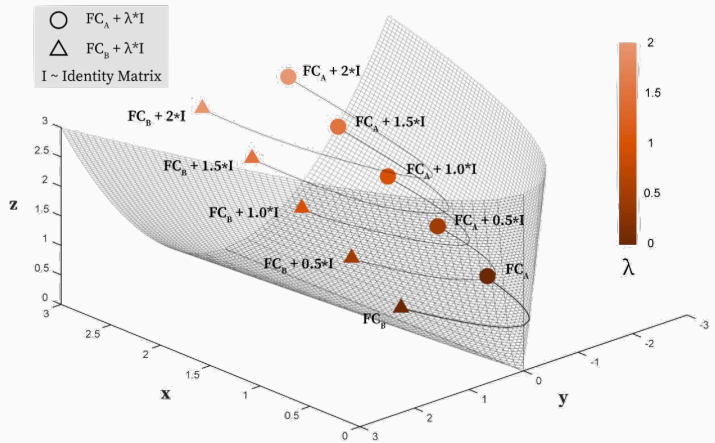


Figure 5: Positive Definite Cone

Figure 4 obtained from Abbas et al. 2020[1]

Mean Reference Matrices for Tangent Space Projection

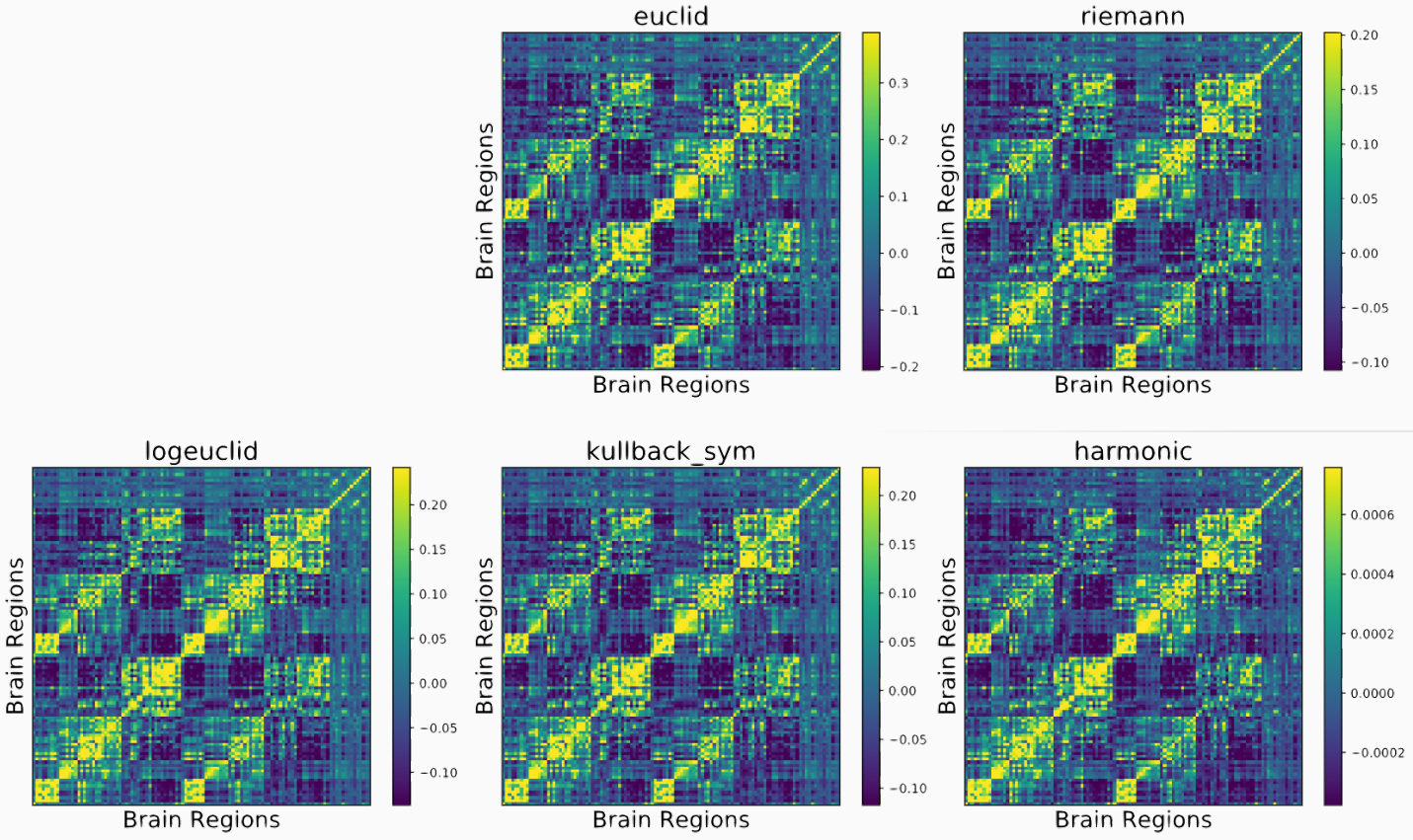


Figure 6: Reference matrices calculated from the Human Connectome Project (n= 424) with Schaefer's 100 brain region parcellation.[9]

Tangent Projected Functional Connectivity Matrices

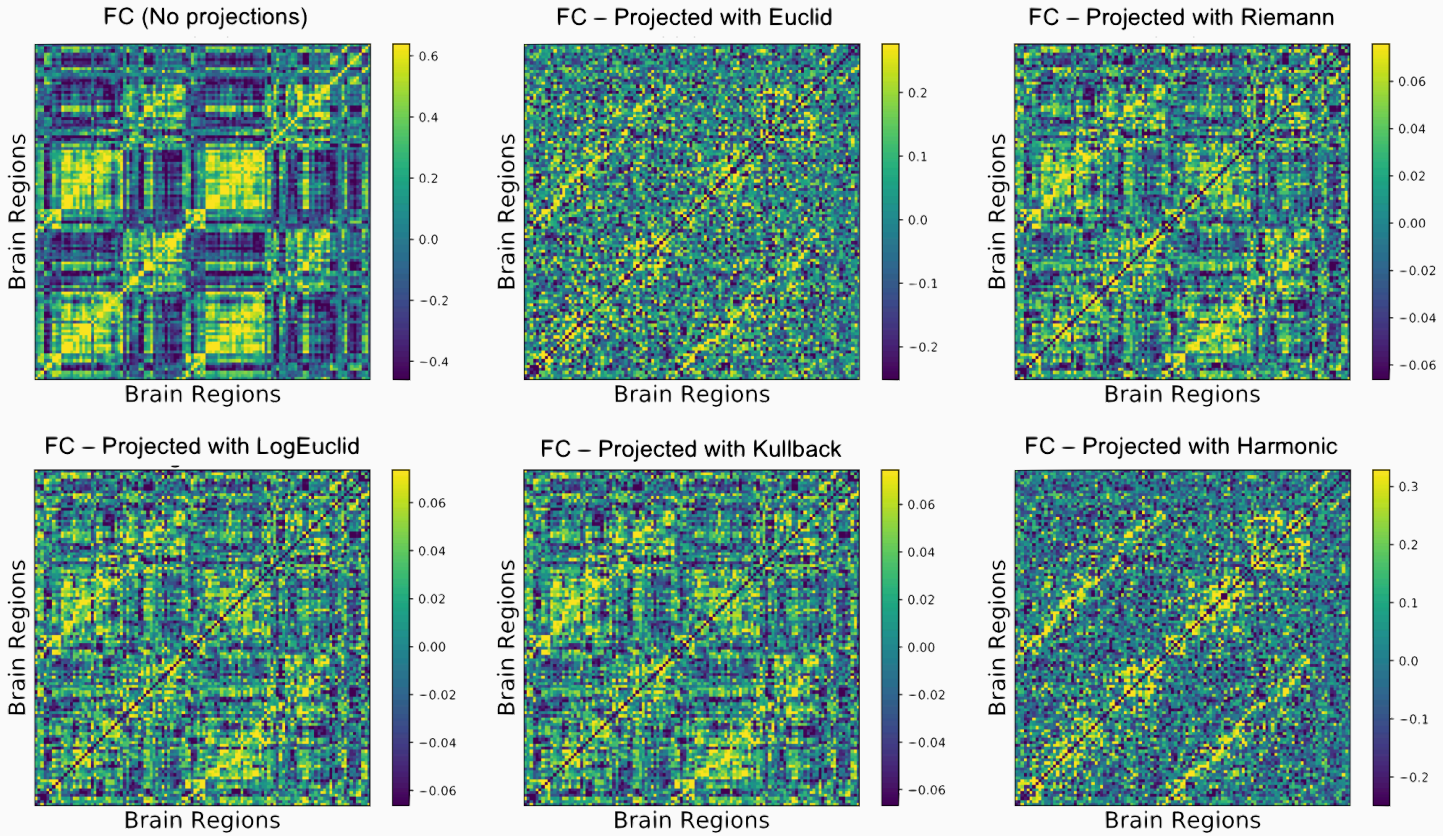


Figure 7: Resting state functional connectomes of Subject 1 after tangent projection with various reference matrices[9]

Brain Parcellation Granularity

Evaluating the effect of parcellation size from 100-500 brain regions

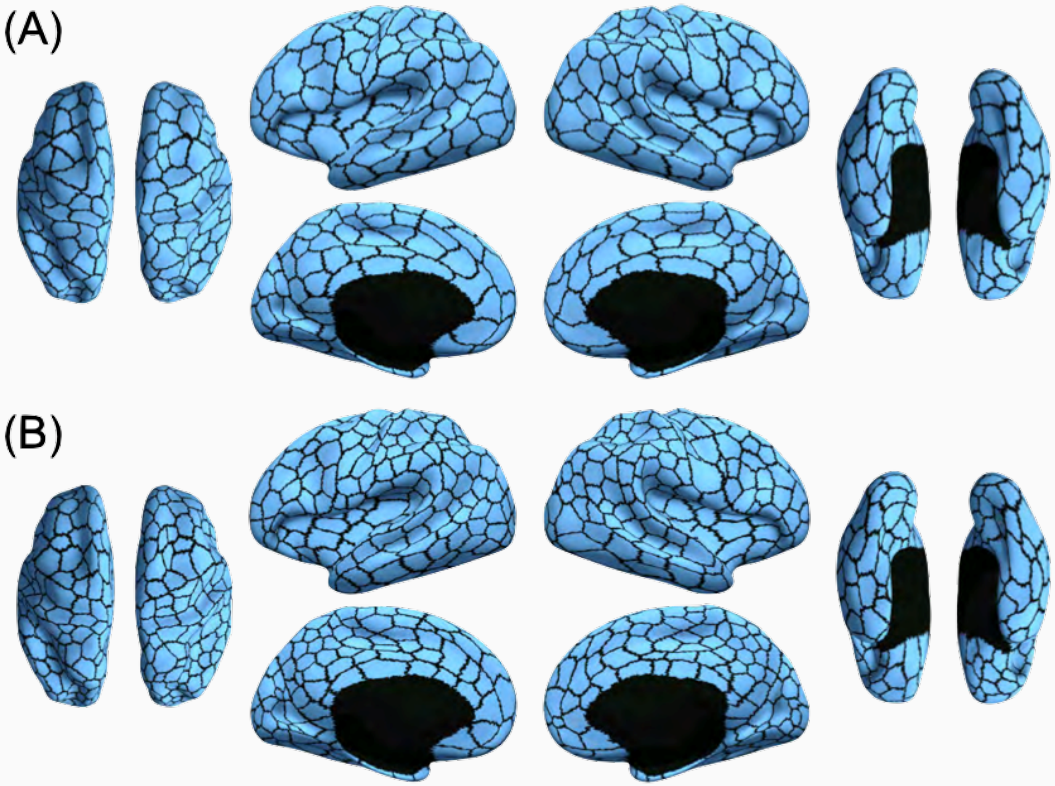


Figure 8: Examples of different Schaefer parcellation granularities[9]

Figure 5 adapted from Schaefer 2018[9]

K-Nearest Neighbors Classifier

1	2	3	4
2	5	6	7
3	6	8	9
4	7	9	10

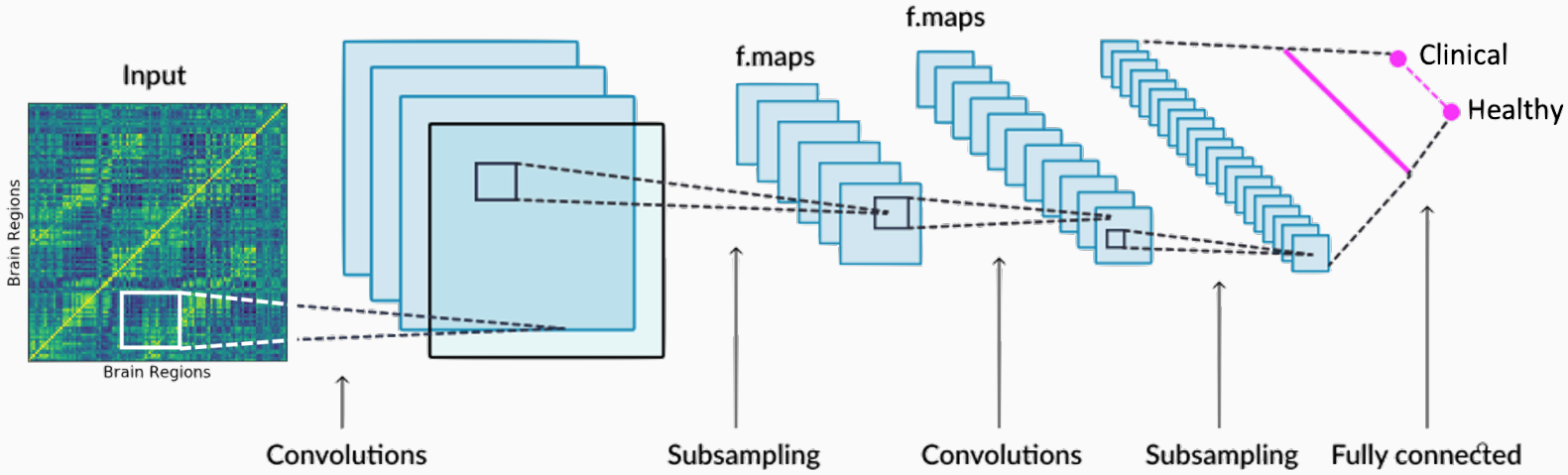
3	1	8	5
1	2	9	10
8	9	8	7
5	10	7	10

2	3	4	6	7	9

1	8	5	9	10	7

Figure 9: Vectorize FCs and predict based on the *k* nearest neighbors

Convolutional Neural Network Classifier



Why Convolutional Neural Networks?

- Preserve spatial information of FC
- Reduces the number of parameters which must be learned
- Does not need feature selection
- Can learn the complex nonlinear relationships of brain functionality
- Recently - enough FC samples to effectively train a model

Example CNN architecture adapted from https://missinglink.ai/guides

Convolutional Neural Network Architecture

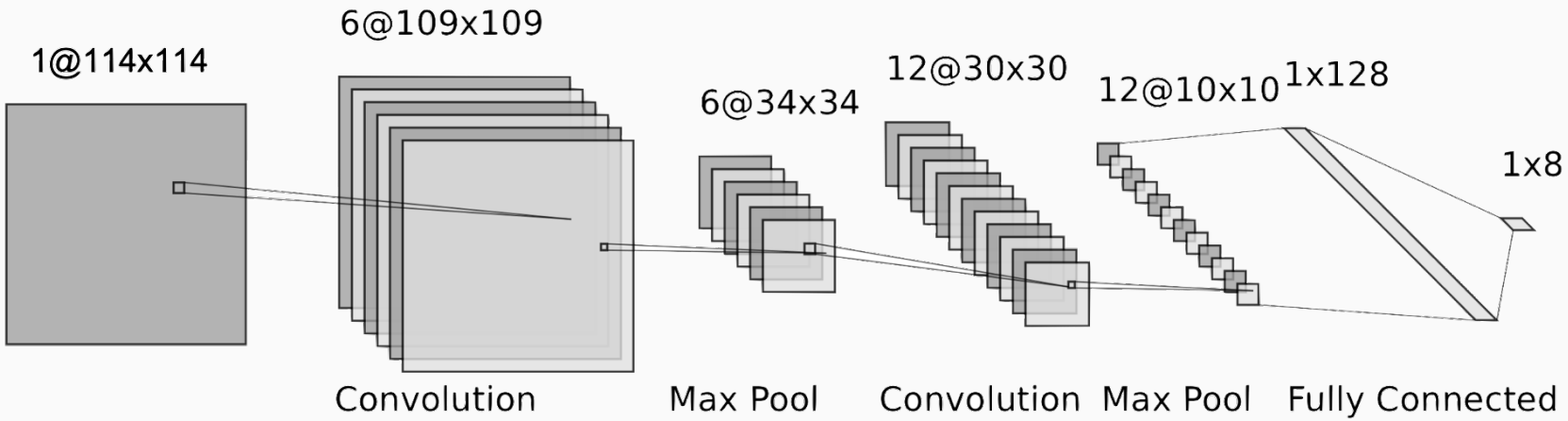


Figure 10: Architecture of Schaefer100 Neural Network

- 80/20 Train/Test split
- Data Normalization
- ReLU activation
- SGD Optimizer
- Cross-entropy Loss
- Learning Rate: $1e-3$
- 200 epochs
- GPU Computing

Results

Task: KNN Classifier Configuration

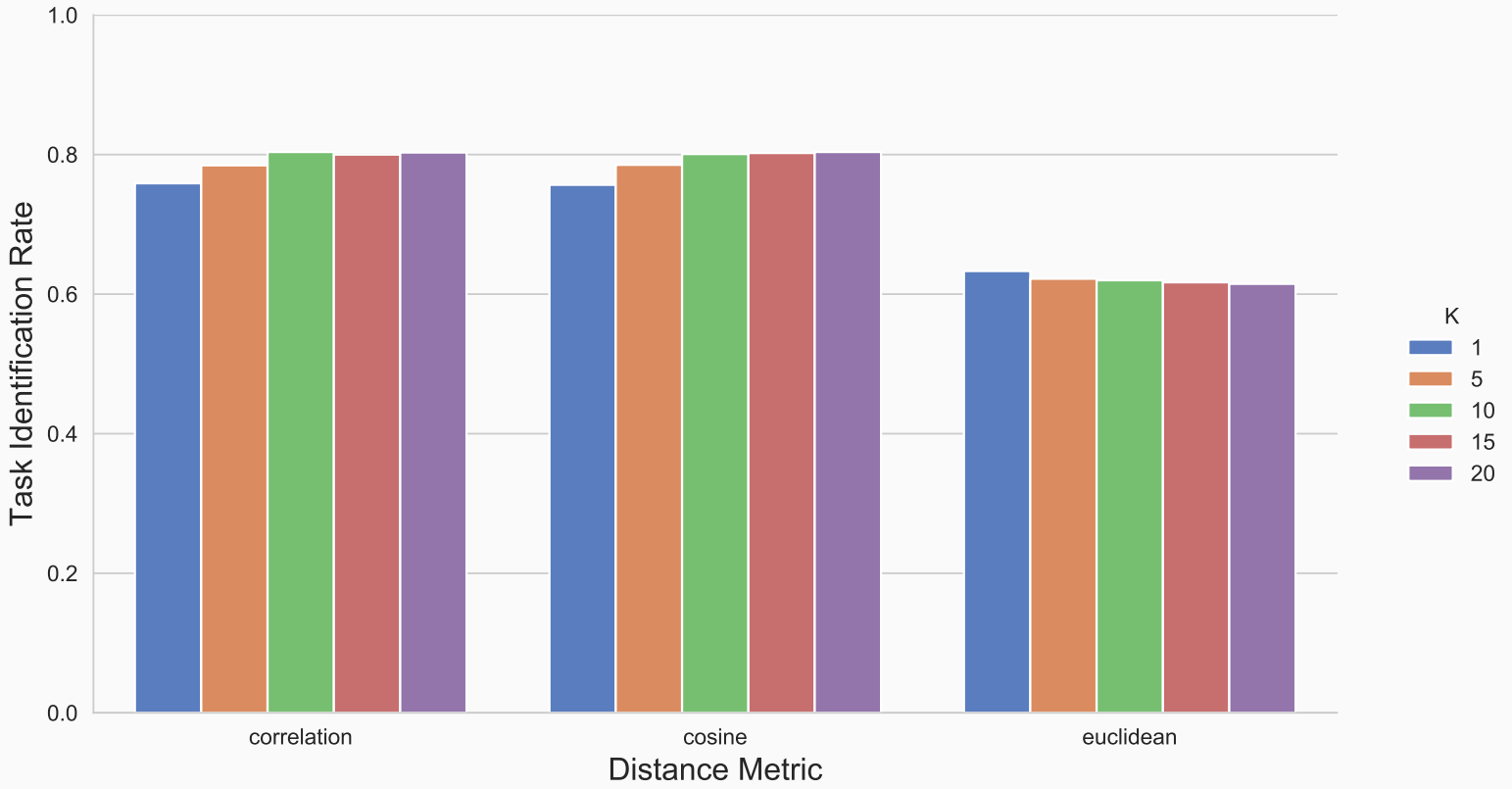


Figure 11: KNN Test Accuracy with various values of *k*

Key Finding

For values of *k* above 10, task identification rate no longer increases significantly. Correlation and cosine perform better than Euclidean.

Task: PCA Reconstruction of Functional Connectomes

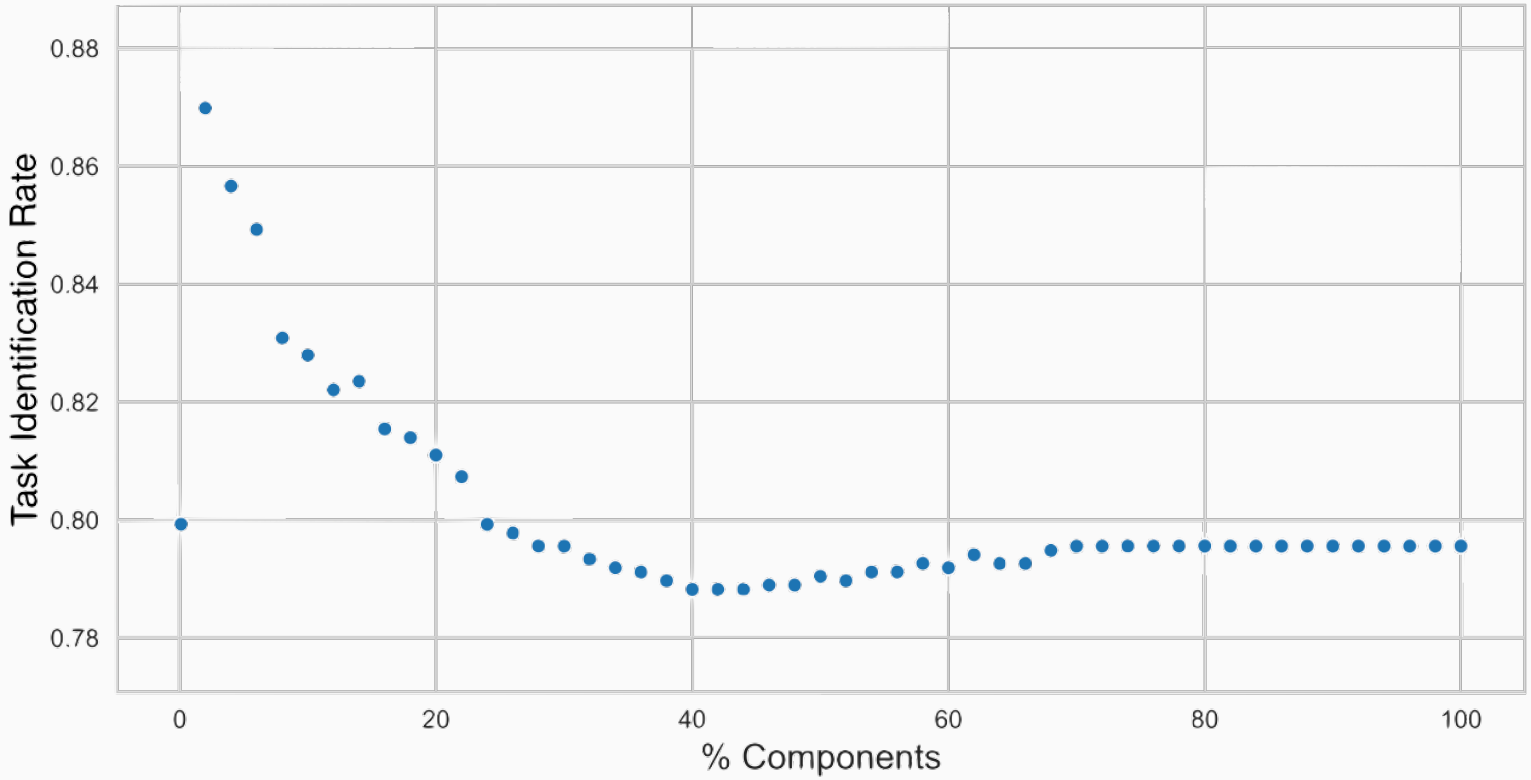


Figure 12: KNN Task Identification Rate with % Principal Components

Key Finding

Optimal PCA reconstruction at roughly **2%** of total PCs

Fine grain search: Optimal reconstruction at **80** of 6,874 PCs

Task: Tangent Space Projection of Functional Connectomes

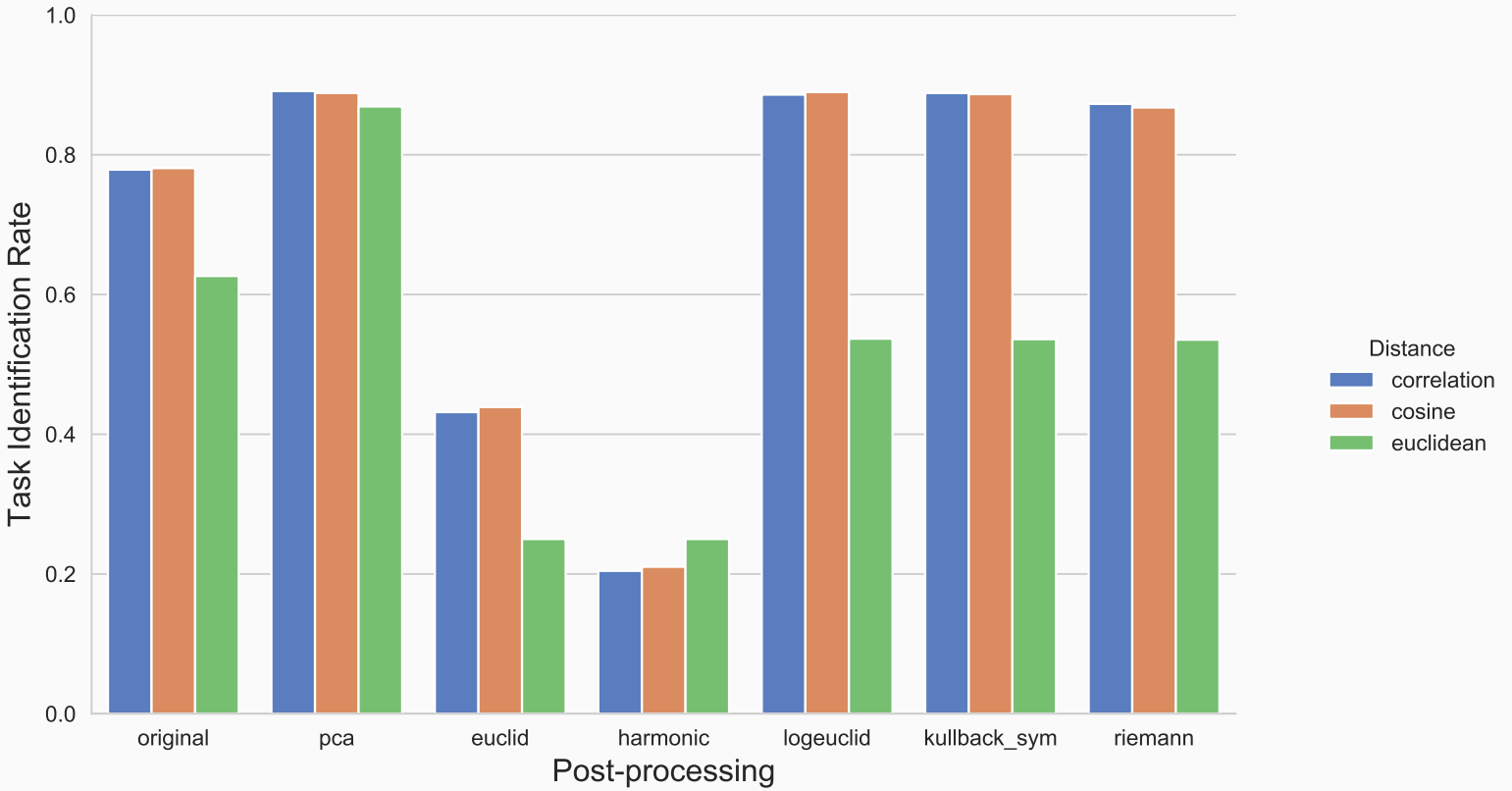


Figure 13: KNN Task Identification Rate with Tangent References

Key Finding

PCA reconstruction and tangent projection with LogEuclid, Kullback, and Riemann reference matrices improve task identification rate

Task: Convolutional Neural Network Training and Validation

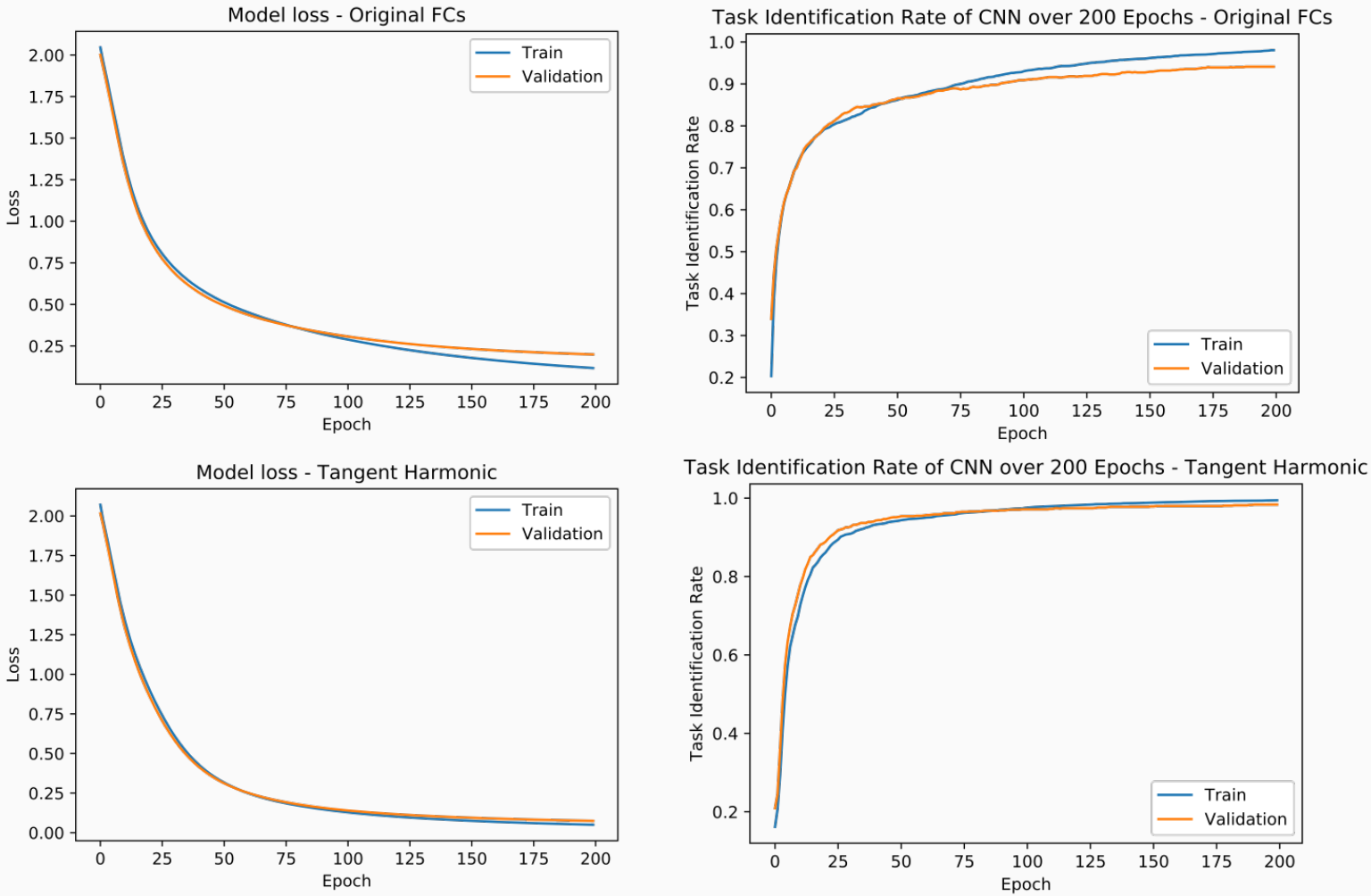


Figure 14: CNN Learning Curves of Original and Tangent Harmonic FCs

Task: Convolutional Neural Network Results

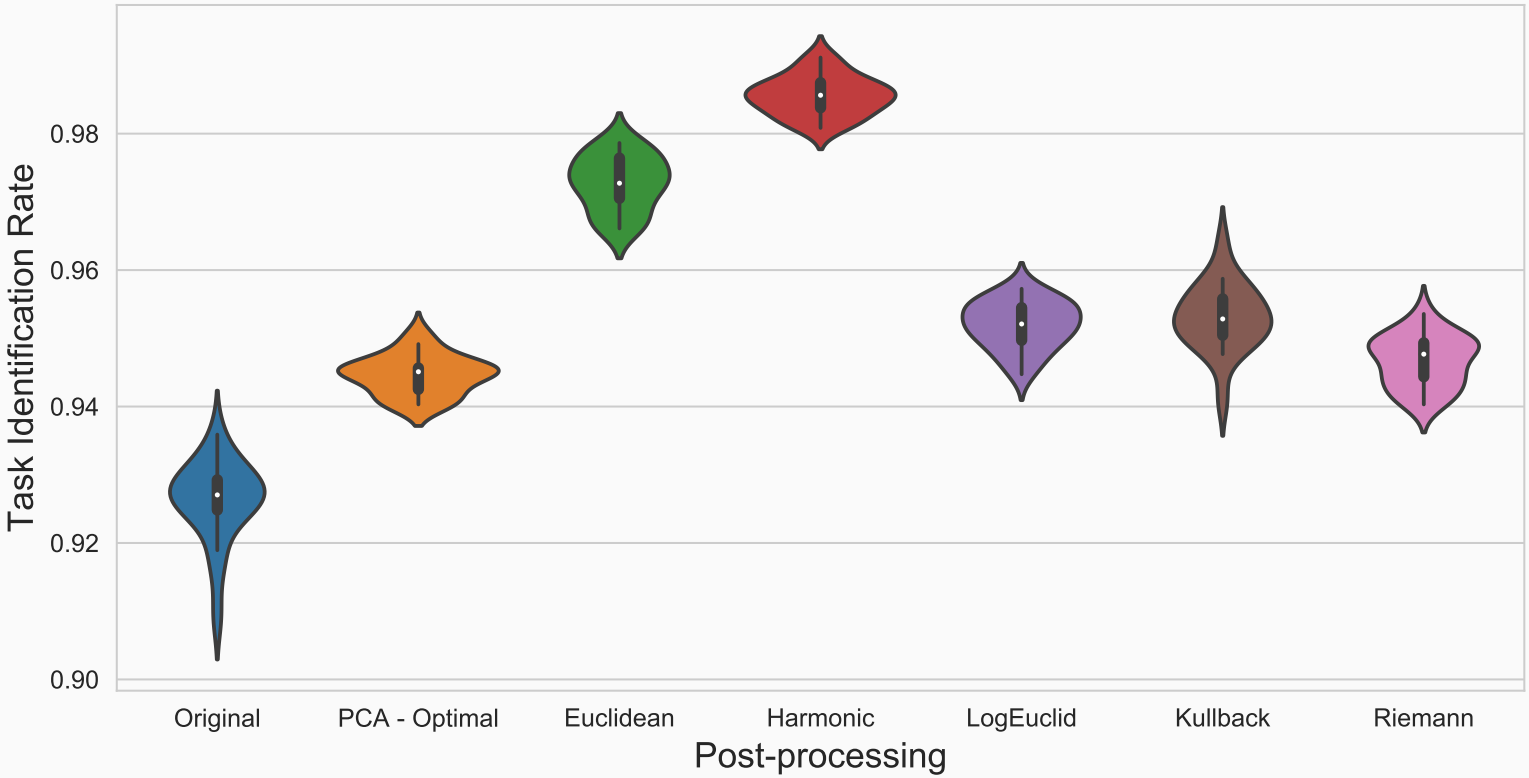


Figure 15: CNN Task Identification Rate with Post-Processing Methods

Key Finding

Tangent projected FCs with harmonic mean reference perform best

Twin: PCA Reconstruction for MZ Twin Identification

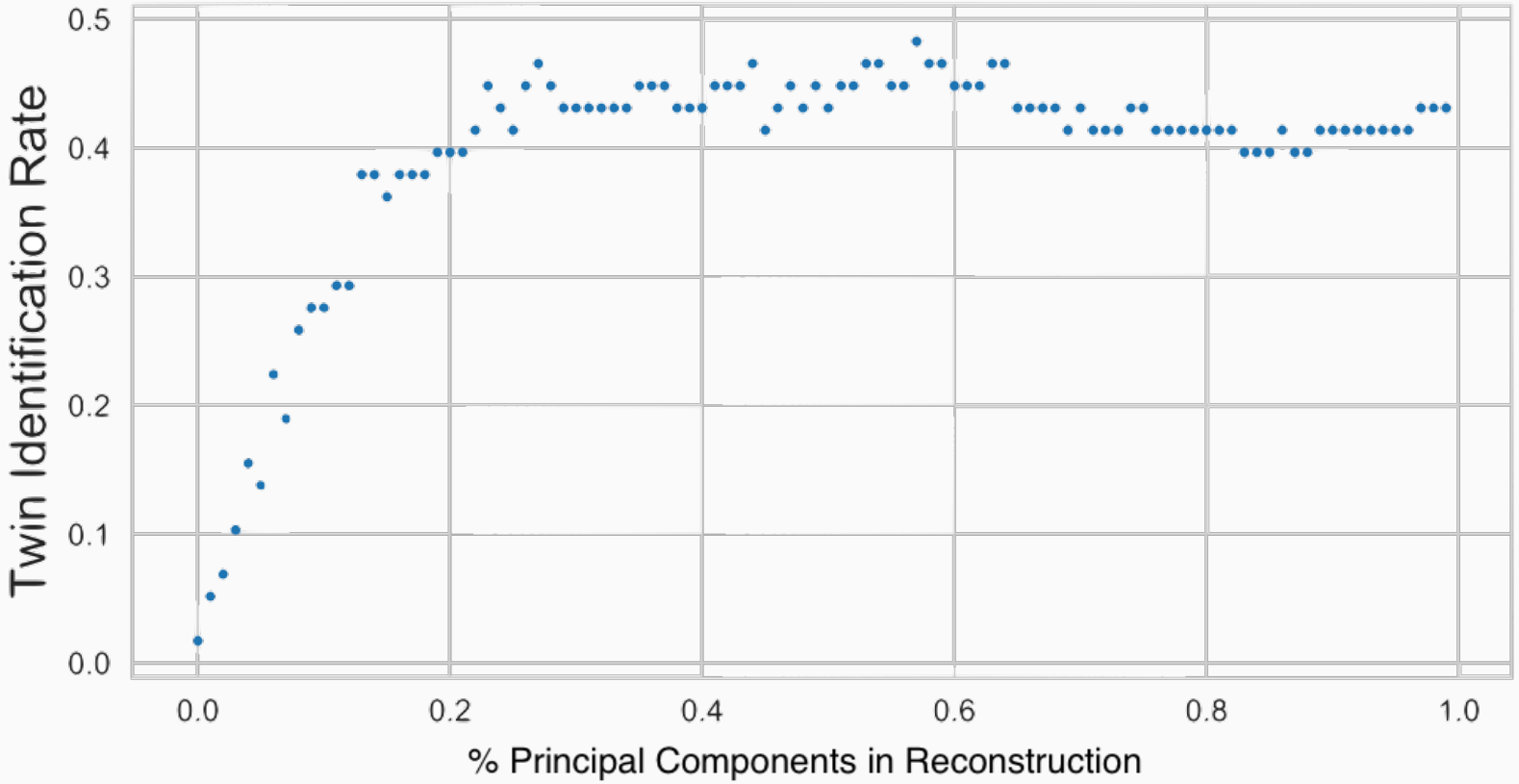


Figure 16: MZ Identification Rate vs. % Components ($n = 106$)

Key Finding

Optimal reconstruction at approximately 50% of total PCs

Twin: MZ Identification with Tangent Reference Matrices

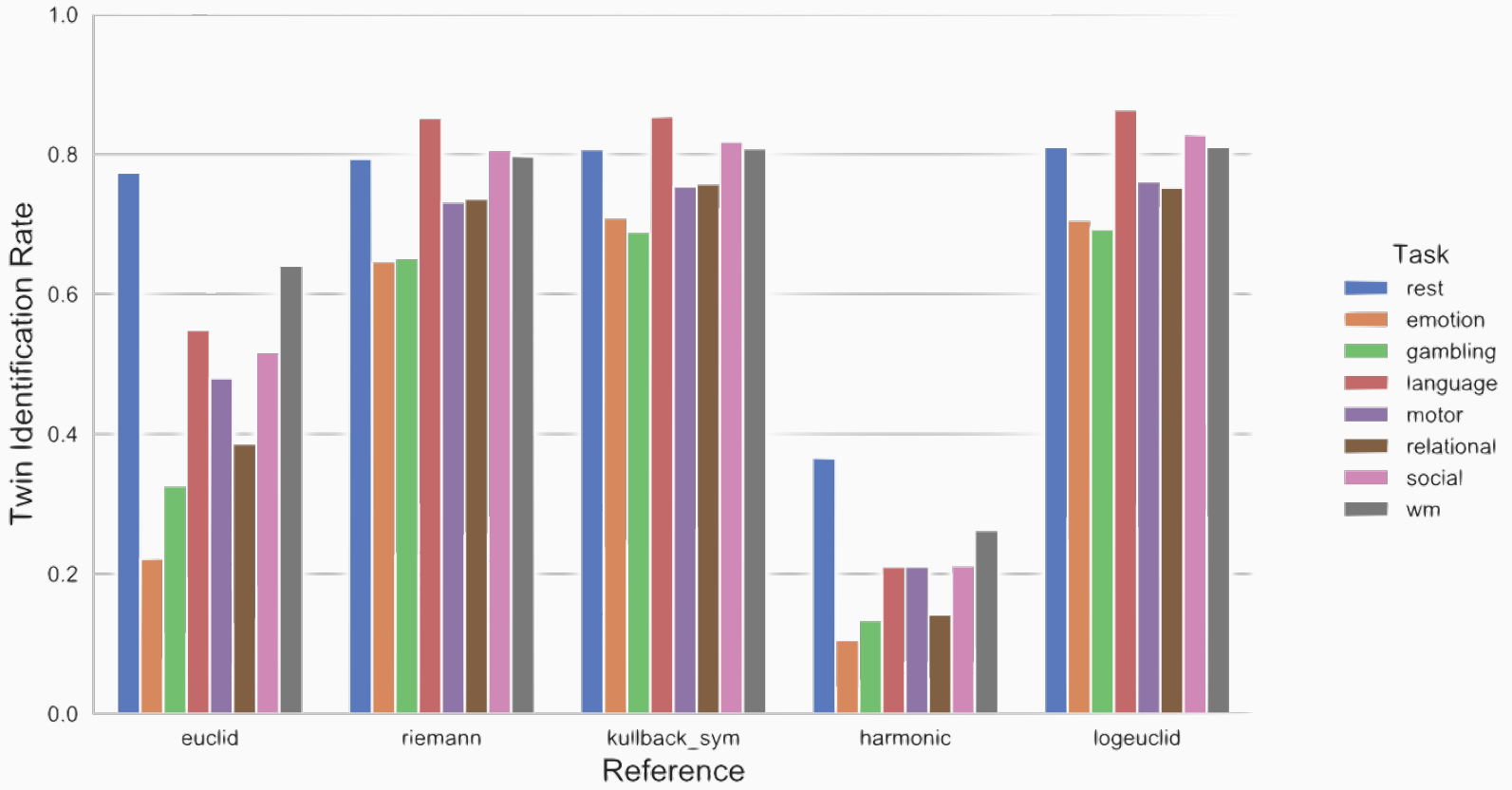


Figure 17: KNN MZ Twin Identification Rate ($n = 106$) with References

Key Finding

LogEuclid, Kullback, and Riemann references perform the best

Twin: MZ Identification across Parcellation Granularities

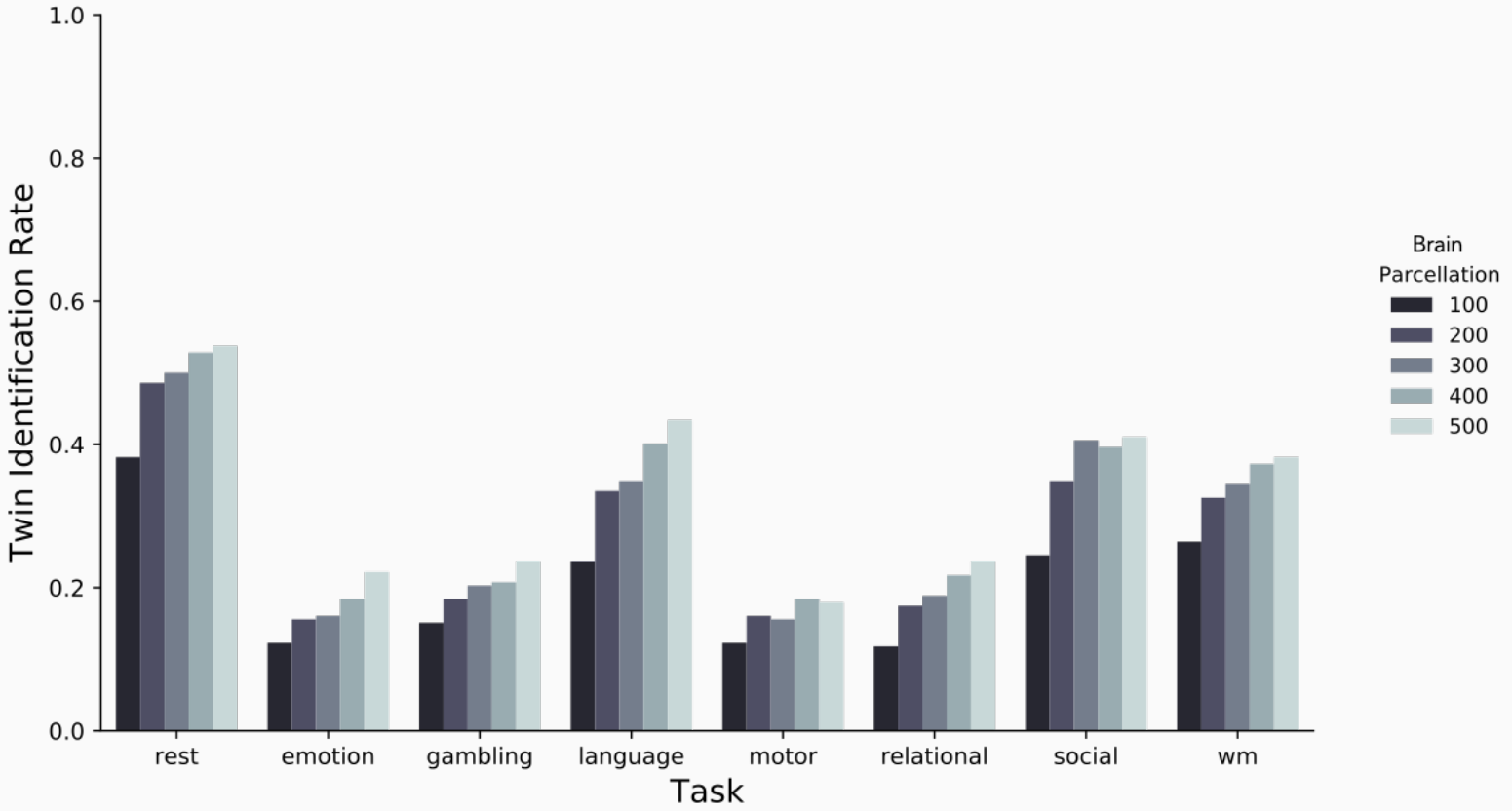


Figure 18: KNN MZ Identification Rates ($n = 106$) with Original FCs

Key Finding

Increasing parcellation granularity increases identification rate

Twin: MZ Identification across Post-processing Methods

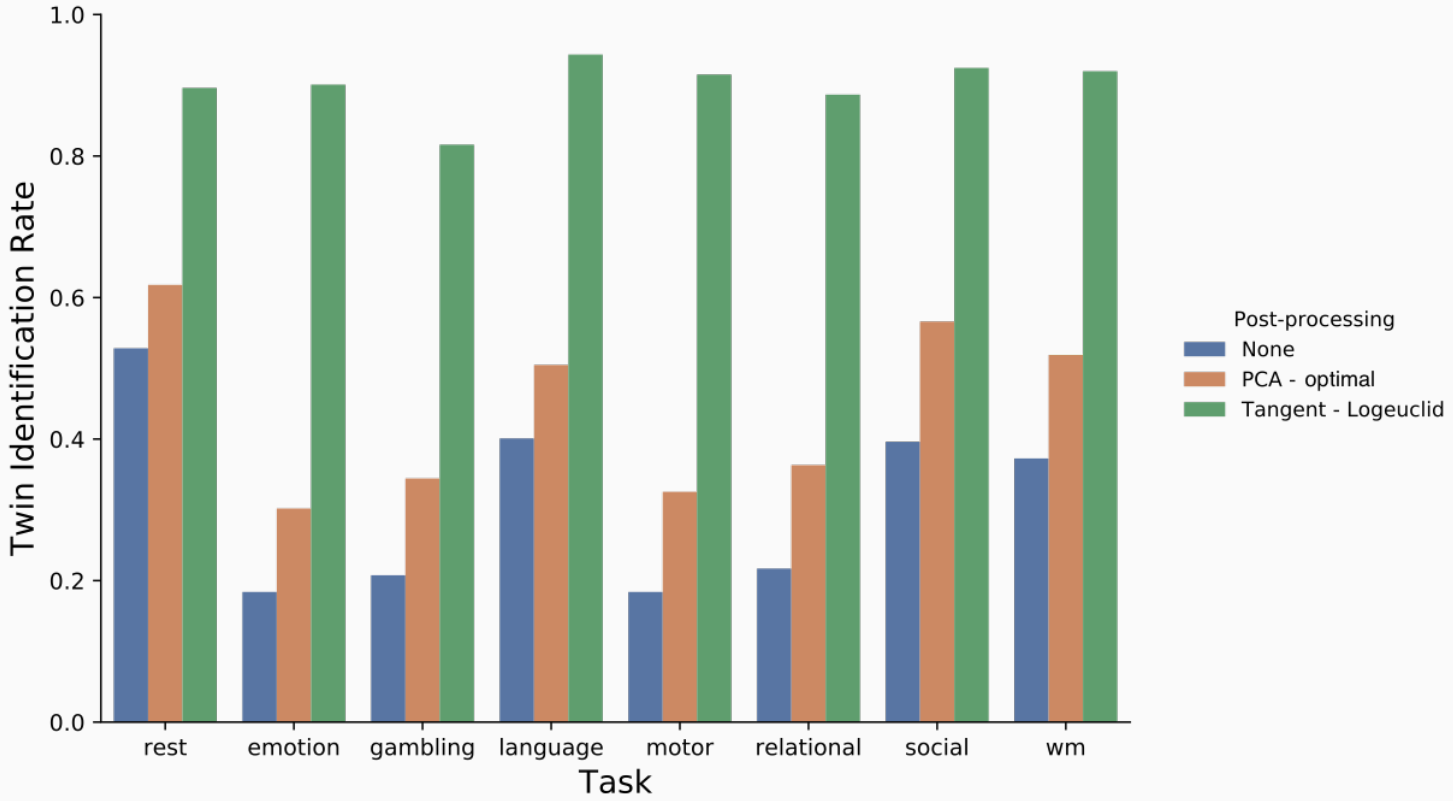


Figure 19: KNN MZ Identification Rates ($n = 106$) with Schaefer400

Key Finding

Tangent projection (LogEuclid) and optimal PCA reconstruction increase KNN MZ twin identification rate compared to original FCs

Twin: PCA Reconstruction for DZ Twin Identification

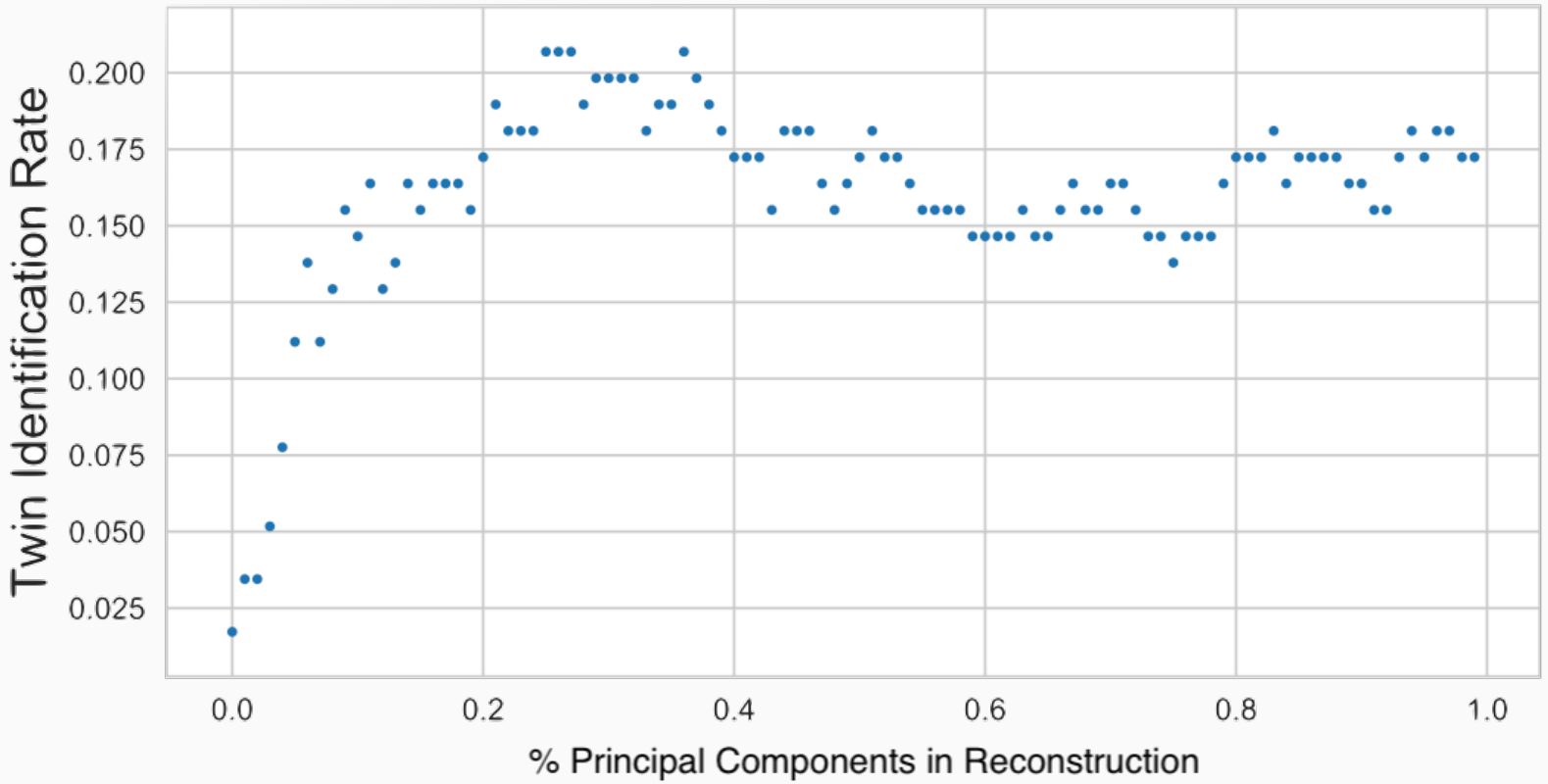


Figure 20: DZ Identification Rate vs. % Components ($n = 58$)

Key Finding

Optimal reconstruction at approximately 30% of total PCs

Twin: DZ Identification with Tangent Reference Matrices

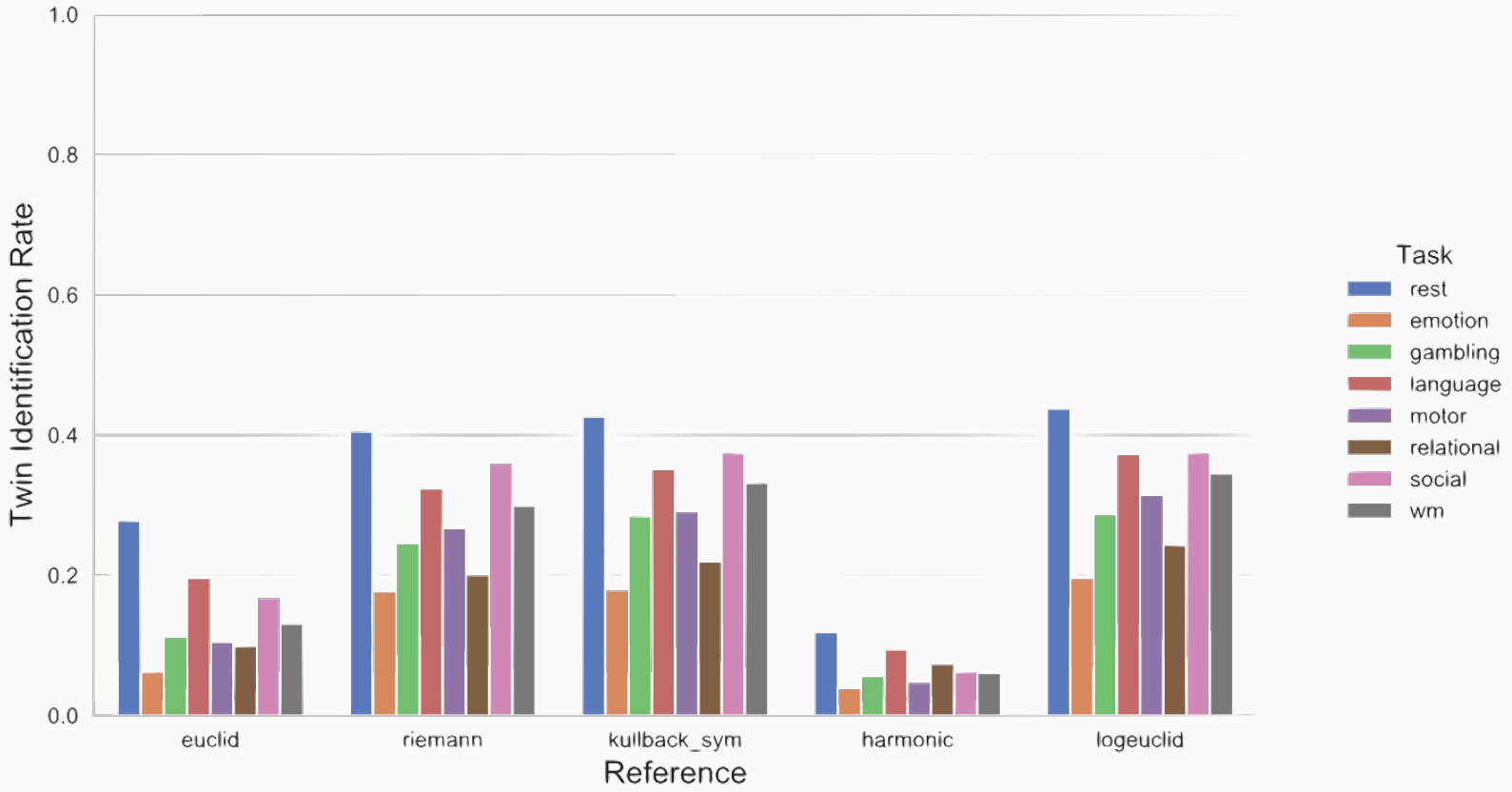


Figure 21: KNN DZ Twin Identification Rate ($n = 58$) with References

Key Finding

LogEuclid, Kullback, and Riemann references perform the best

Twin: MZ vs. DZ Identification Rates

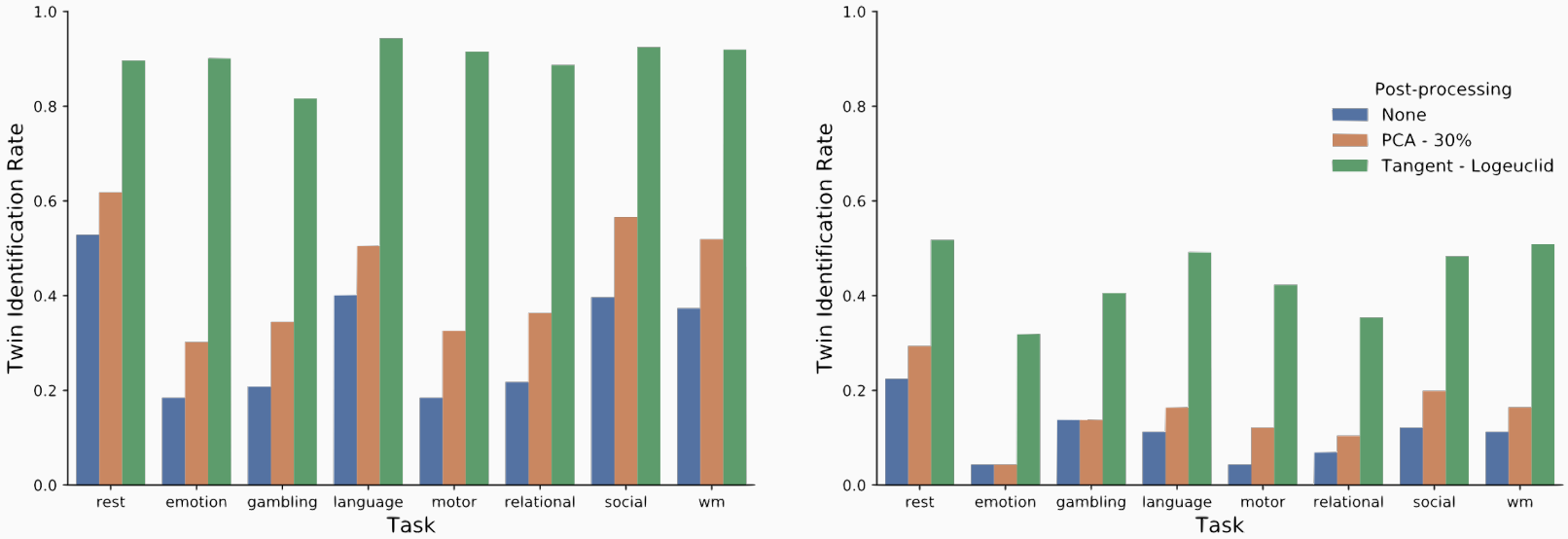


Figure 22: KNN MZ and DZ Twin Identification Rates with Schaefer400

Key Finding

DZ twin identification Rate with Tangent projected (LogEuclid) FCs surpasses MZ twin identification rate with original FCs

Subject: Identification across Parcellation Granularities

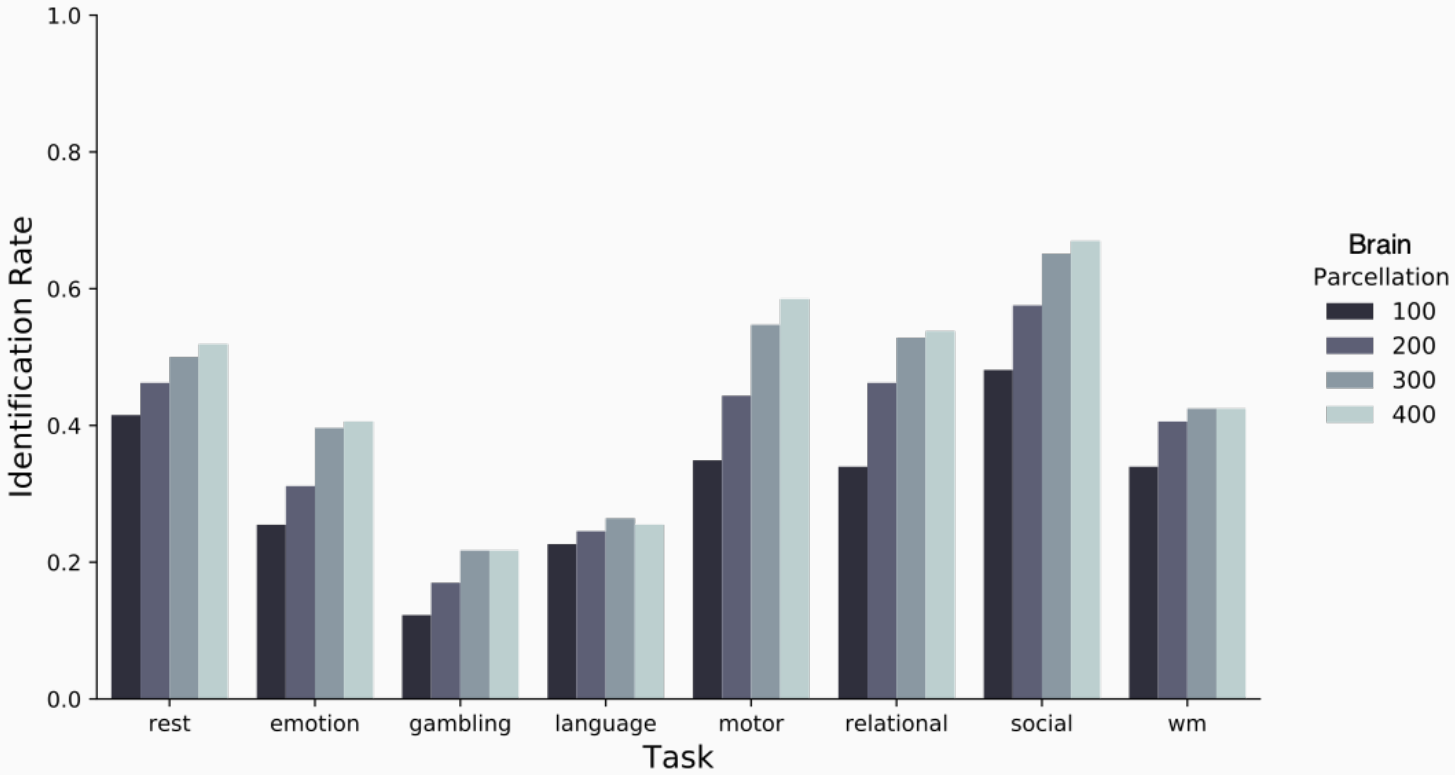


Figure 23: KNN Subject Identification Rate ($n = 106$) with Original FCs

Key Finding

Increasing parcellation granularity increases identification rate

Subject: Identification with Post-processing Methods

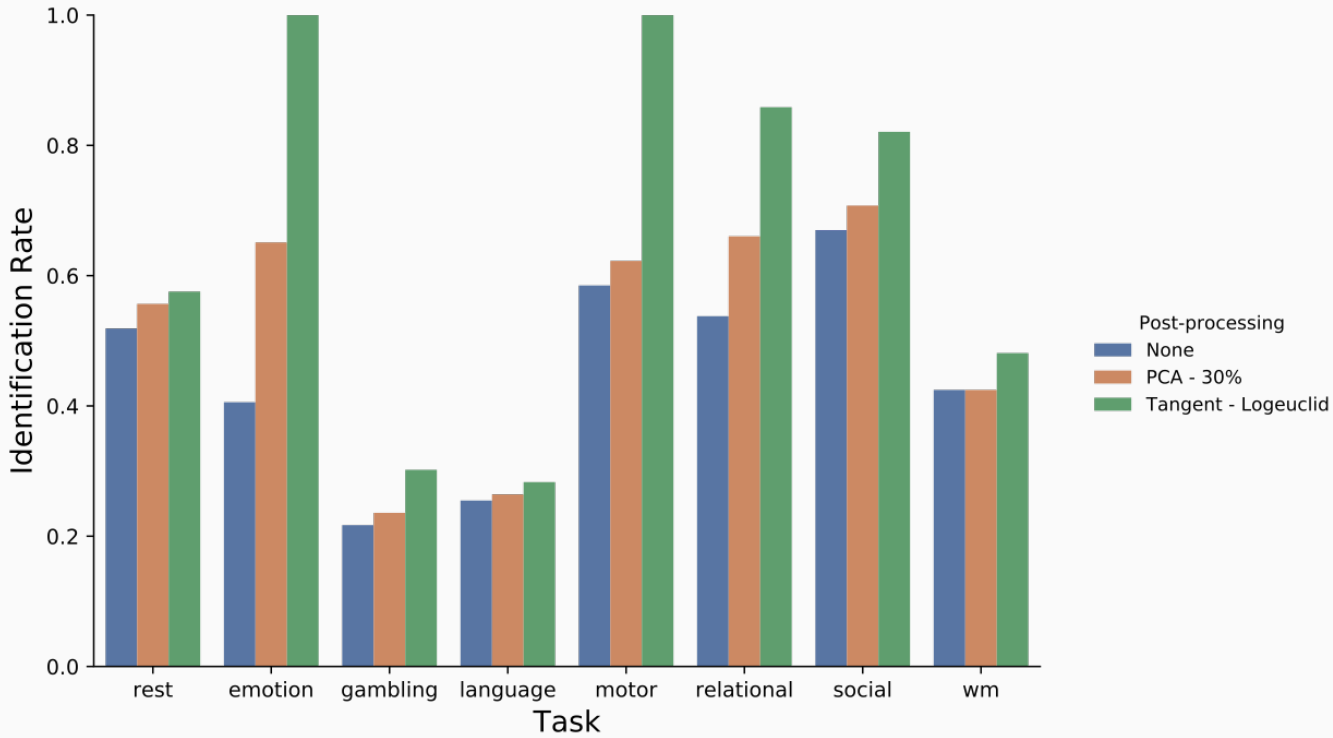


Figure 24: KNN Subject Identification Rate ($n = 106$) with Schaefer400

Note

Similar results to MZ twins but with lower performance in gambling and language FCs. Likely due to fewer FC samples per label.

Discussion

Research Outcomes

- Tangent space projection preserves the geometry of functional connectivity matrices (+0.061 identification rate vs. original FCs).
- Optimal PCA reconstruction depends on the dataset. For task, this is at 80 of 6,784 principal components. For MZ and DZ twins, optimality is at 50% and 30% of total components, respectively. PCA reconstruction increases task and twin identification rates.
- Increasing parcellation granularity from 100 brain regions to 400 brain regions improves both task and twin identification rates with the KNN classifier.
- Tangent projected FCs with the harmonic reference allow CNN classifier to outperform the state-of-the-art task identification rate by 0.049.
- Tangent projected FCs with the logarithmic Euclidean reference allow KNN classifier to outperform state-of-the-art MZ and DZ twin identification rates by 0.303 and 0.297, respectively.

Areas of Improvement and Future Work

- Perform a structured hyperparameter search for the convolutional neural network with open-source packages
- Investigate effect of brain parcellation size on task identification rate with the CNN classifier
- Understand the significance of the reference matrix in tangent projection and its effect on classifier performance (e.g., harmonic reference vs. logarithmic Euclidean)
- Apply the proposed post-processing methods to clinical fMRI cohorts and obtain classifiers for diagnosis and disease progression

Acknowledgements

- Dr. Joaquín Goñi
- Dr. Kausar Abbas
- Duy Duong-Tran
- Uttara Tipnis
- Dr. Mario Ventresca
- Dr. Juan Wachs
- Purdue School of Industrial Engineering

References I

K. Abbas, M. Liu, M. Venkatesh, E. Amico, J. Harezlak, A. D. Kaplan, M. Ventresca, L. Pessoa, and J. Goñi.
Regularization of functional connectomes and its impact on geodesic distance and fingerprinting.
page 16.

E. Amico, A. Arenas, and J. Goñi.
Centralized and distributed cognitive task processing in the human connectome.
Network Neuroscience, 3(2):455–474, Jan. 2019.

E. Amico and J. Goñi.
The quest for identifiability in human functional connectomes.
Scientific Reports, 8(1):8254, Dec. 2018.

M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson.
The minimal preprocessing pipelines for the Human Connectome Project.
NeuroImage, 80:105–124, Oct. 2013.

A. Gritsenko, M. Lindquist, and M. K. Chung.
Twin Classification in Resting-State Brain Connectivity.
In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1391–1394, Apr. 2020.
ISSN: 1945-8452.

P. M. Matthews, G. D. Honey, and E. T. Bullmore.
Applications of fMRI in translational medicine and clinical practice.
Nature Reviews Neuroscience, 7(9):732–744, Sept. 2006.

References II

NIH.
Human Connectome Project.

U. Pervaiz, D. Vidaurre, M. W. Woolrich, and S. M. Smith.
Optimising network modelling methods for fMRI.
NeuroImage, 211:116604, May 2020.

A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo.
Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI.
Cerebral Cortex, 28(9):3095–3114, Sept. 2018.

K. Specht.
Current Challenges in Translational and Clinical fMRI and Future Directions.
Frontiers in Psychiatry, 10, 2020.
Publisher: Frontiers.

M. Venkatesh, J. Jaja, and L. Pessoa.
Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification.
NeuroImage, 207:116398, Feb. 2020.

X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, J.-H. Gao, and B. Qiu.
Decoding and mapping task states of the human brain via deep learning.
Human Brain Mapping, 41(6):1505–1519, 2020.
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24891.

Questions?

PCA Reconstruction Visualized

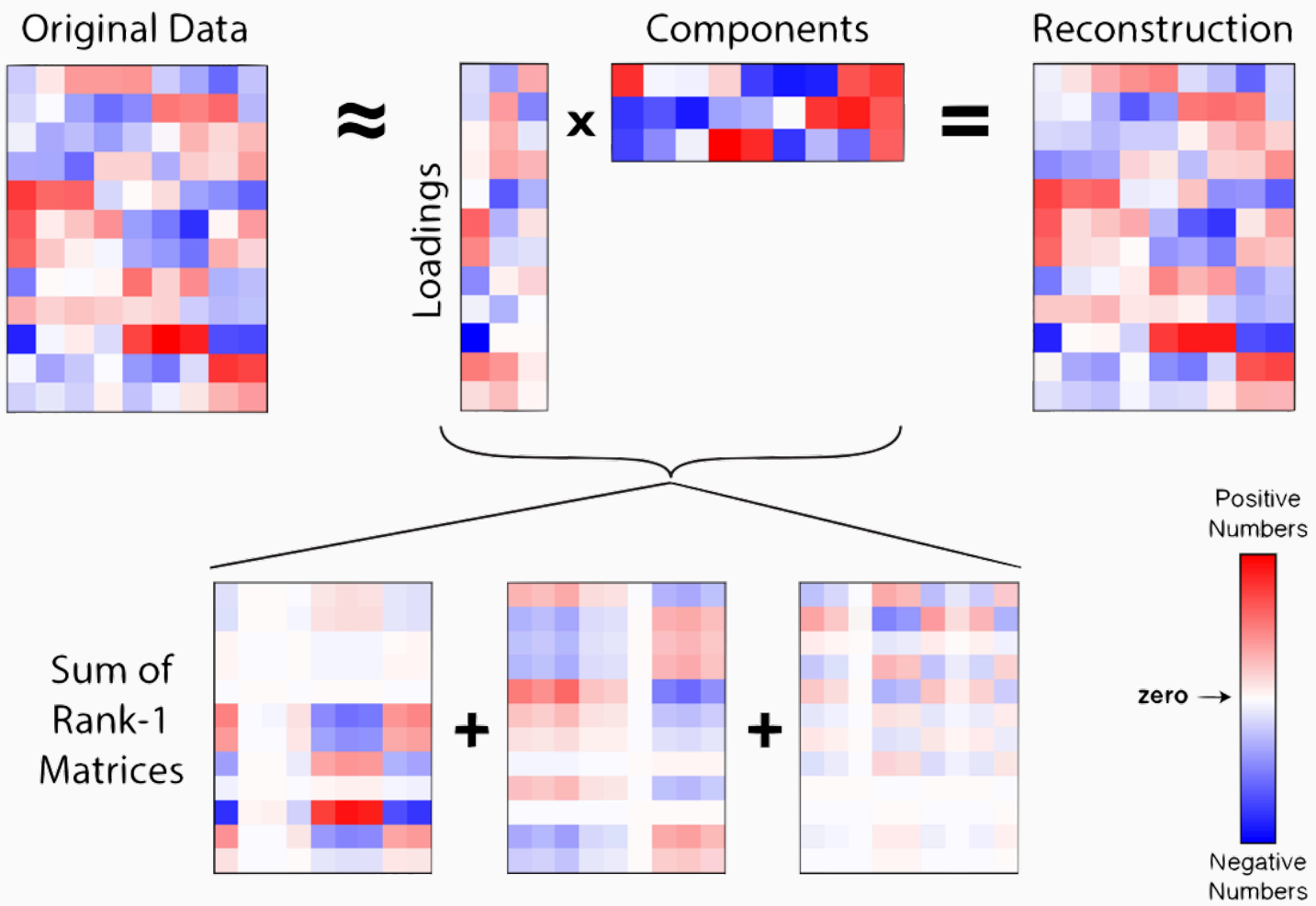


Figure 25: PCA reconstruction with 3 PCs

Figure obtained from http://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/

Twin: DZ Twin Identification across Post-processing Methods

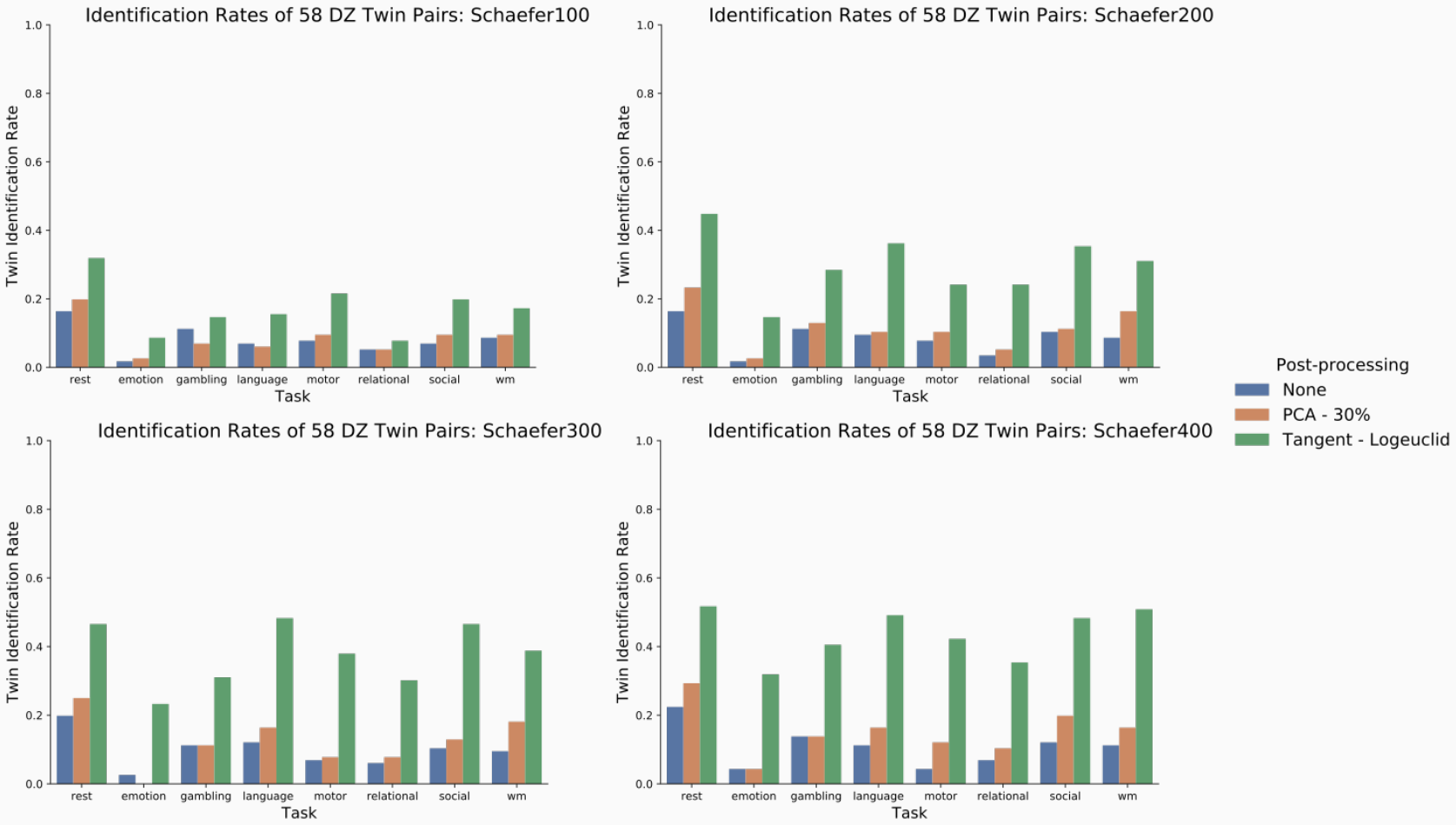


Figure 26: CNN Task Identification Rate with Post-Processing Methods

Twin: DZ Identification across Parcellation Granularities

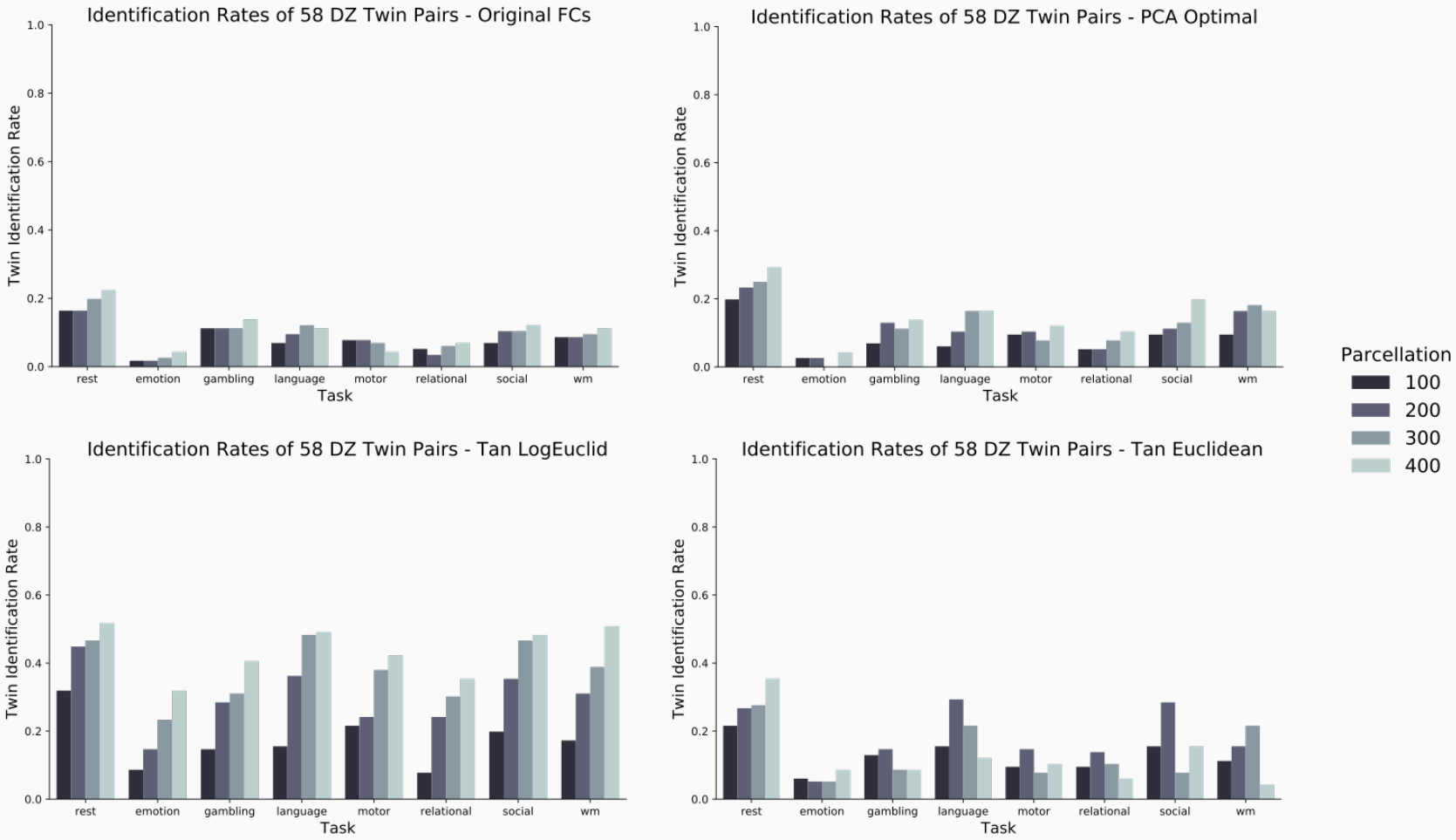


Figure 27: CNN Task Identification Rate with Post-Processing Methods

Twin: MZ Identification across Post-processing Methods

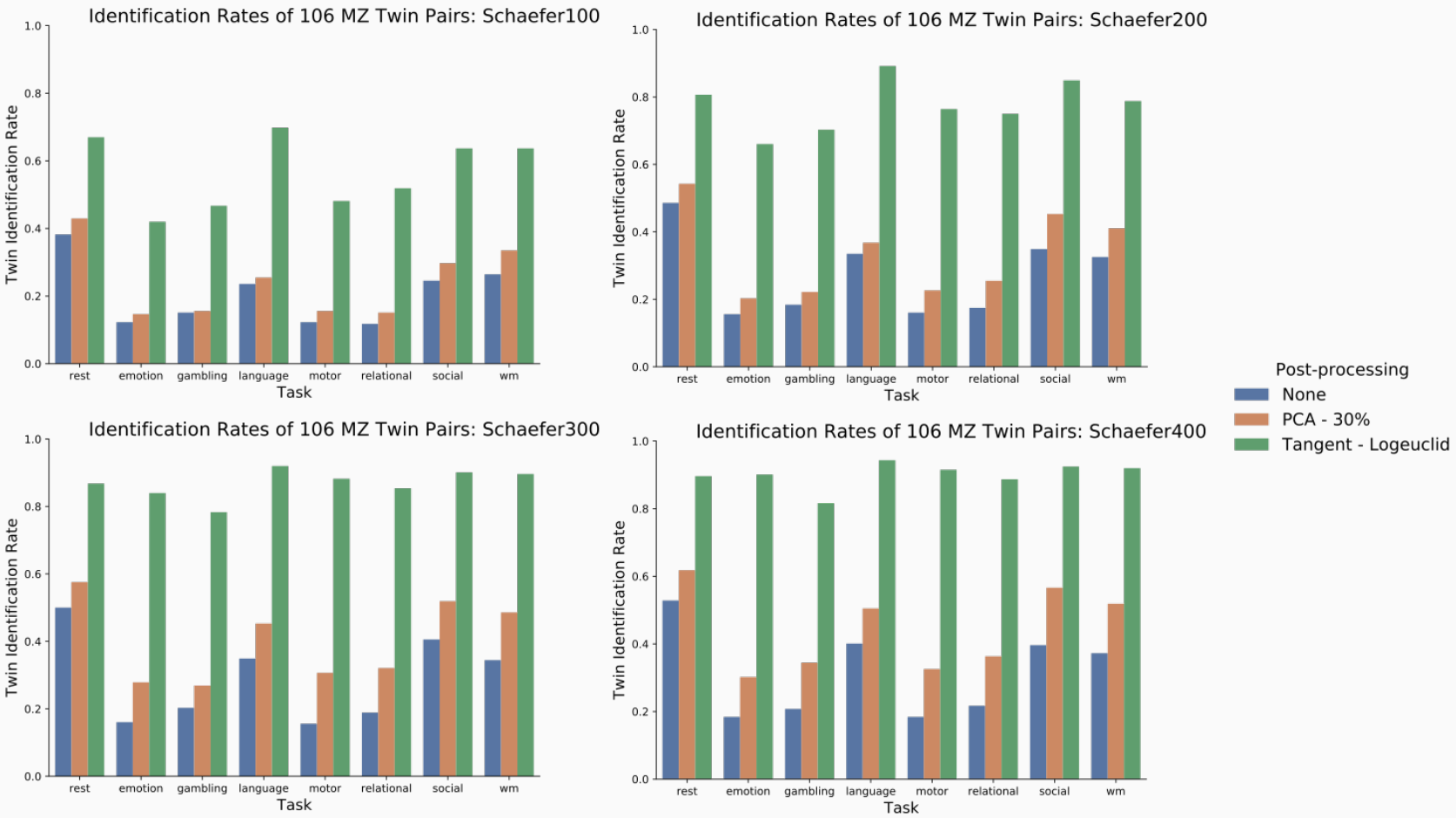


Figure 28: KNN Task Identification Rate with Post-Processing Methods

Twin: MZ Identification across Parcellation Granularities

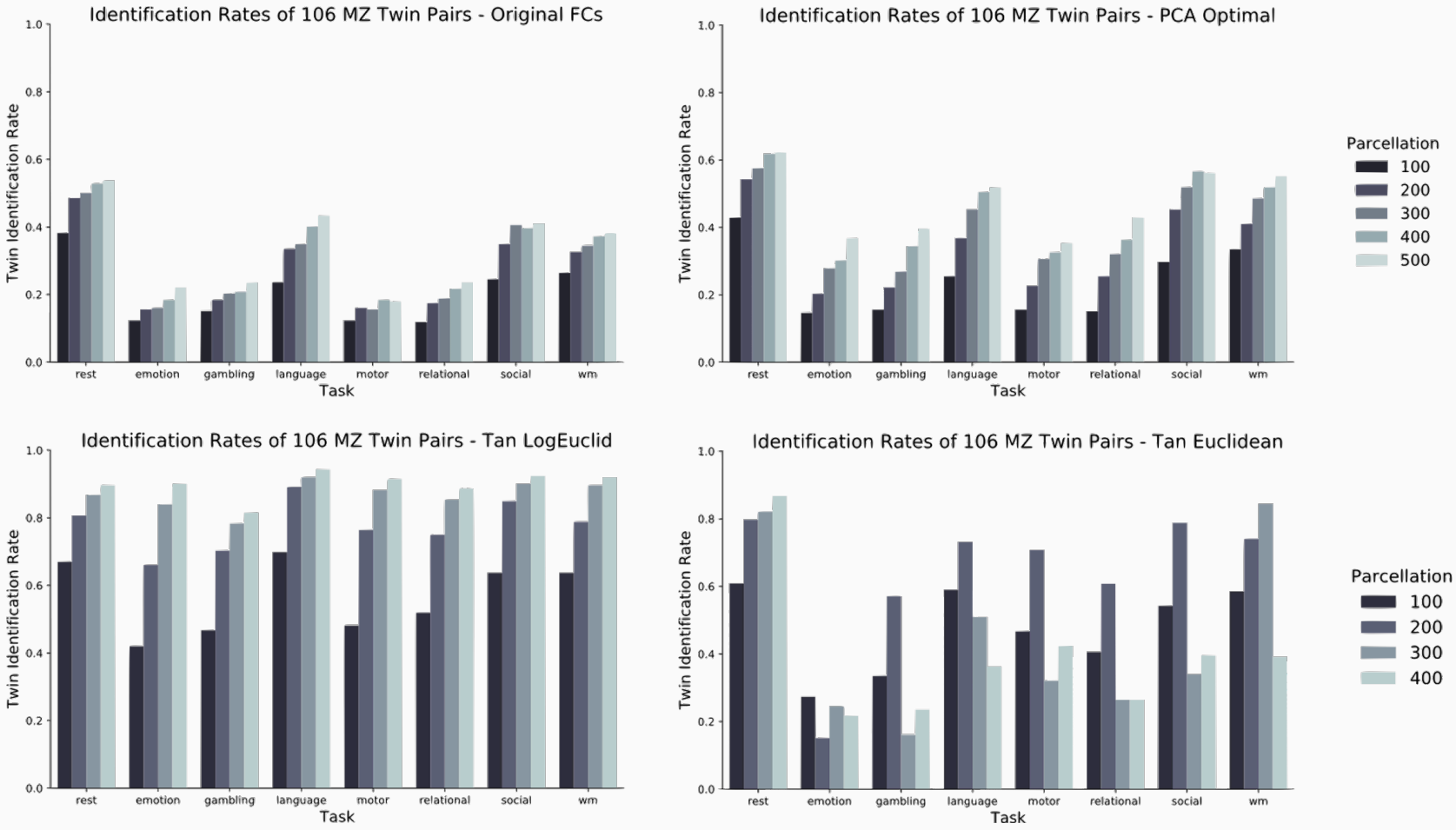


Figure 29: KNN MZ Twin Identification Rate with Post-Processing Methods

Subject: Identification with Post-processing Methods

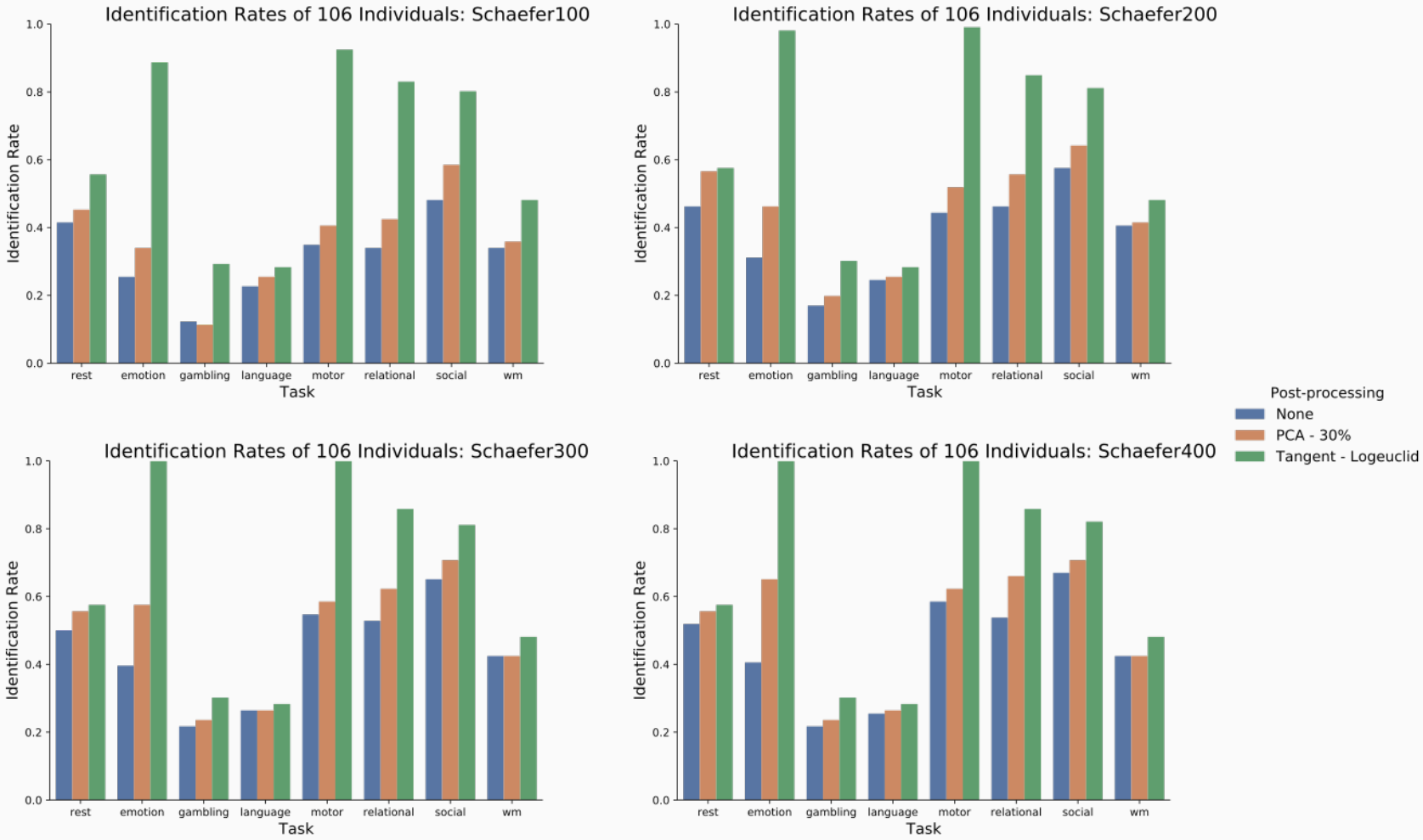


Figure 30: KNN Subject Identification Rate with 106 Individuals

Subject: Identification across Parcellation Granularities

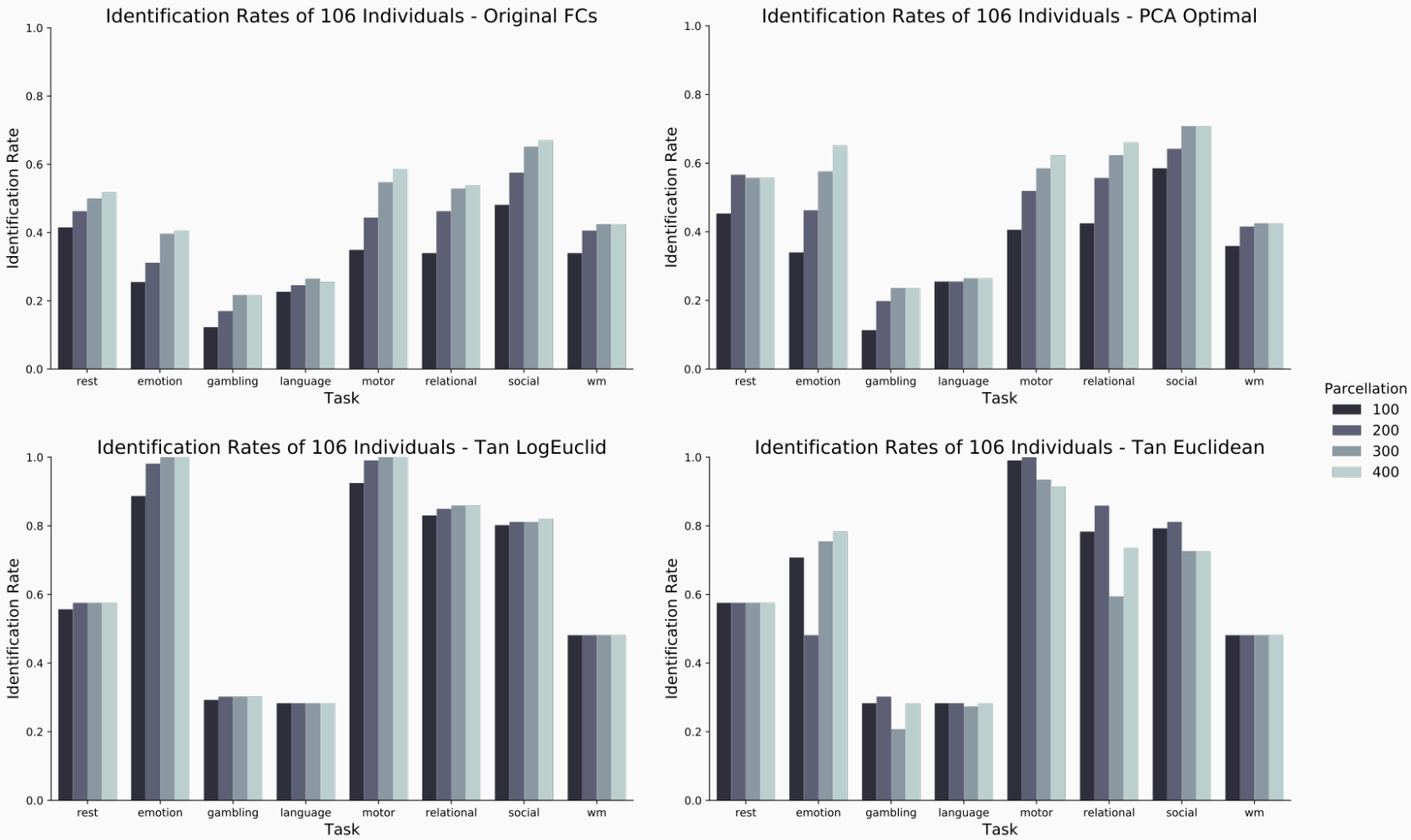


Figure 31: KNN Subject Identification Rate with 106 Individuals

Task: CNN Identification Rates

Post-processing	Task Identification Rate	SD (σ)
Original FCs	0.926	0.006
PCA - Optimal	0.945	0.003
Tan - Euclidean	0.973	0.004
Tan - Harmonic	0.986	0.003
Tan - LogEuclid	0.952	0.003
Tan - Kullback	0.953	0.005
Tan - Riemann	0.947	0.004