A Radiogenomics-based Coordinate System to Quantify the Heterogeneity of Glioblastoma

Fanyang Yu
Department of Bioengineering
University of Pennsylvania
yfy@seas.upenn.edu

Anahita Fathi Kazerooni
Department of Radiology
University of Pennsylvania
anahita.kazerooni@pennmedicine.upenn.edu

Pratik Chaudhari
Department of Electrical and Systems Engineering
University of Pennsylvania
pratikac@seas.upenn.edu

Christos Davatzikos
Department of Radiology
University of Pennsylvania
christos.davatzikos@pennmedicine.upenn.edu

Abstract

Glioblastoma (GBM) is an aggressive brain tumor with median patient survival of about 15 months. The key reason for our poor understanding of GBM is that it is a highly heterogeneous disease with molecular heterogeneity and spatiotemporal heterogeneity. Radiogenomics is a rapidly emerging field that seeks to develop non-invasive imaging signatures associated with genetic mutations of such cancers from magnetic resonance imaging (MRI) scans. This paper develops a technique to quantify the molecular heterogeneity of GBM using radiogenomic features. We fit a probabilistic model that predicts the likelihood of 13 different genetic mutations and use a technique called Intensive Principal Component Analysis (InPCA) to visualize the predictions of this model. The principal components of InPCA form an interpretable coordinate system for characterizing the imaging signatures of different GBM pathways; this coordinate system is consistent with clinical research. We quantify the overlap of different pathway groups to characterize the molecular heterogeneity of GBM. Such analysis can potentially be used in the future for targeted treatments, e.g., when patients present with one or more of these overlapping pathways.

1 Introduction

Glioblastoma (GBM) is an aggressive brain tumor with a high likelihood of recurrence. The median patient survival is 14.6–16.7 months [1] in spite of sophisticated treatment options. One of the major reasons for treatment failure in GBM is the intra-tumor heterogeneity, implying various genetic sub-populations may cohabit the tumor and diffuse through the neurophil [2–4]. This makes treatments or surgical resection of tumors extremely challenging. The molecular heterogeneity indicates that multiple mutations across complex pathways with significant cross-talk renders single-target therapies ineffective. There is also spatiotemporal heterogeneity where diffuse and migrating glioma cells make even *ex vivo* assays challenging [5].

Radiogenomics is a rapidly emerging field that seeks to develop signatures of cancers by exploiting the differences in the magnetic resonance imaging (MRI) features corresponding to cells with different mutations [6,7]. This is very powerful because it enables non-invasive approaches to examine the tumor and track its progression during treatment. If we can extract imaging biomarkers that capture

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the phenotypic characteristics of the tumor and its surroundings based on descriptors for image intensity, morphology, texture, etc. from MRI features, then we can hope to develop a more holistic understanding of the tumor and eventually aid treatment[8]. Although radiogenomics has led to promising results for some types of cancers[9], the molecular heterogeneity of GBM is somewhat unique and it is the key hurdle to instantiating this program for GBM.

This paper develops a technique to quantify the molecular heterogeneity of GBM from radiogenomic features. We use a dataset of multi-parametric MRI (mpMRI) scans (both conventional and advanced modalities including T1, T1-Gd, T2, T2-FLAIR, DSC and DTI) of 286 subjects who were diagnosed with GBM. Corresponding to these scans, we have genomics data from next generation sequencing (NGS), which is an ultra-high throughput sequencing technology used in clinical practice to determine the genetic buildup for each subject. NGS data is the result of an assay of tumorous tissue which provides "annotations" for the presence or absence of a set of 13 gene mutations (EGFR, PDGFRA, MET, FGFR2, PIK3CA, PIK3R1, PTEN, NF1, BRAF, TP53, MDM4, CDKN2A and RB1) underlying 5 commonly identified pathways (RTK, PI3K, MAPK, P53, RB1) in GBM. Table 1 describes our dataset. In addition to the biological complexity of the problem, the small sample size ($n = 286$) and class imbalance (0.7% subjects have FGFR2 alteration vs. 36.4% subjects have PTEN alteration) make this data challenging for machine learning. Moreover, subjects typically have multiple gene mutations and multiple pathways activated. This is a unique dataset for studying radiogenomic features of GBM and the first of its kind.

Roughly speaking, our goal is to identify imaging signatures of known prominent genetic pathways and quantify the "overlap" between different pathway groups. For pathways that are relatively well clustered, we hope to identify precise signatures, and for pathways have large overlaps with other pathways, we hope to quantify the amount of overlap across these clusters. Our technical contributions are summarized as follows.

1. We fit a probabilistic model, i.e., a deep neural network, which predicts the likelihood of the 13 different gene mutations (multiples one can be active for each subject) using over 3000 different radiogenomic features obtained from mpMRI images as inputs. The output of this model is probability distribution and it can be thought of as a "feature vector" pertaining to each subject.
2. We leverage a nonlinear manifold learning technique called Intensive Principal Component Analysis (InPCA) for visualizing such feature vectors[10]. This technique creates a global isometry and thus it can visualize the global geometry of the predictions because it preserves pairwise distances between points (unlike t-SNE[11], UMAP[12] or diffusion maps[13] which only preserve local distances).
3. The principal components of InPCA form an interpretable coordinate system for characterizing the imaging signatures of different GBM pathways. We show how these principal components can imply potential clinical findings, e.g., the first principal component (PC1) is associated with P53 and then RTK/PI3K; the second principal component (PC2) is associated with P53 then MAPK, etc. Our approach therefore enables to use the labels of the genetic mutations to discover the imaging signatures of the pathways. Our data-driven results can provide clinical value to the investigation of the cancer pathways.
4. We quantify the overlap of the pathways to characterize the molecular heterogeneity of GBM. The quantitative analysis can potentially facilitate the development of targeted therapy especially for treating patients with multiple pathways involved.

2 Methods

We have NGS-based annotations $y_i \in \{0,1\}^{13}$ and corresponding inputs x_i for subjects $i = \{1, \ldots, N\}$. Labels y_i can have multiple genetic alterations. We train a multi-layer neural network by minimizing the binary cross-entropy loss

$$\ell(w) = -N^{-1} \sum_{i=1}^{N} \sum_{k=1}^{13} (y_i)_k \ \log p_w(k \mid x_i) + (1 - (y_i)_k) \ \log\left(1 - p_w(k \mid x_i)\right).$$

where $\log p_w(k \mid x_i) - \log(1 - p_w(k \mid x_i)) \equiv (\hat{y}_i)_k$ is the output of the network. This objective is minimized using stochastic gradient descent (see §5.2). We normalize the 13 logits $\hat{y}_i$ using a softmax operation and set $\log p'_w(k \mid x_i) \propto (\hat{y}_i)_k$ to obtain a 13-dimensional "feature vector" $(z_i)_k = p'_w(k \mid x_i)$ for each datum x_i. We can also cluster z_i using the von Mises-Fisher (vMF) mixture model[14] by observing that the vector $\sqrt{(z_i)_k}$ for $k = 1, \ldots, 13$ has unit norm and therefore lies on sphere of radius 1. Since unsupervised clustering cannot adequately identify radiogenomics signatures of different pathways due to significant inter-pathway overlapping, we fit vMF model for each pathway in a

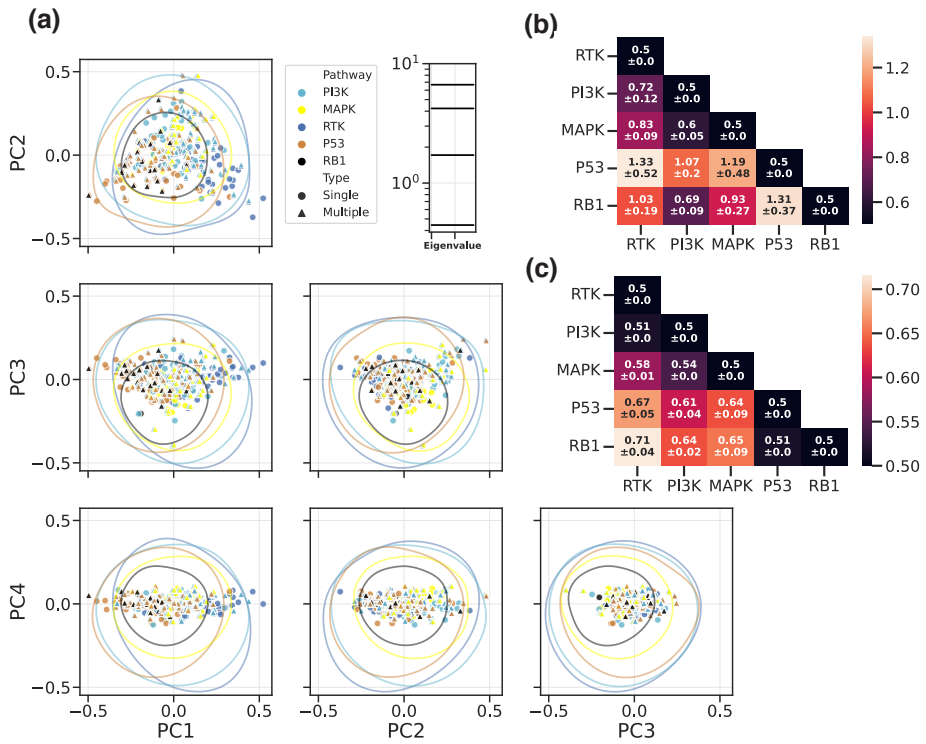

Figure 1: (a) InPCA embeddings of top 4 principal components of subjects with both single pathway and multiple-pathway alterations. Subjects with multiple-pathway alterations (presented by multiple triangle points close to each other) typically fall in-between different pathway groups. Clusters pertaining to different pathways can be distinguished using the colors of the markers. This indicates that even if the model was trained using annotations of the altered genes, its features are useful to predict the genetic pathways. **(b) Inter-pathway distances for subjects with single pathway alteration.** The distances are calculated as inter-pathway Bhattacharyya distances between points normalized by the average of the intra-pathway Bhattacharyya distances (error bars from 5-fold nested CV). Entries that are smaller than one indicate a large overlap between the pathways (PI3K-MAPK, PI3K-RTK); entries that larger than 1 are relatively well-separated pathways (MAPK-P53). This is a quantitative exposition of the heterogeneity of GBM features: pathways that are relatively well-separated in this matrix have distinctive radiogenomic signatures. **(c) Inter-pathway distances for subjects with multiple-pathway alterations.** The distances are similarly calculated as the above. It can be observed that the inter-pathway distances are smaller among multiple-pathway subjects than these among single pathway subjects. This result suggests multiple-pathway subjects are more heterogeneous than single-pathway subjects.

supervised fashion by selecting the number of components as 1. We will next analyze these feature vectors z_is.

Intensive Principal Component Analysis (InPCA)[10] is a generalization of multi-dimensional scaling (MDS)[15]. It works by creating a matrix $D \in \mathbb{R}^{n\times n}$ whose entries are the Bhattacharyya distances $\mathrm{dB}(z_i, z_j) = -\log \sum_k \sqrt{(z_i)_k (z_j)_k}$ and calculating $W = -LDL/2$ where $L = \delta_{ij} - 1/n$. An eigen-decomposition of $W = U\Sigma U^\top$ with eigenvalues sorted in descending order of their magnitude gives the InPCA embeddings $\mathbb{R}^{n\times n} \ni X = U\sqrt{\Sigma}$. Unlike standard PCA where eigenvalues are non-negative, eigenvalues of InPCA can be both positive and negative[10]. This allows the InPCA embedding to be an isometry:

$$\sum_{k=1}^{13} \left((X_i)_k - (X_j)_k\right)^2 = \mathrm{dB}(z_i, z_j) \geq 0.$$

for two columns X_i, X_j of X. We use the InPCA embedding for visualizing the clusters corresponding to each pathway. We will use the intra-cluster and inter-cluster pairwise Bhattacharyya distances to quantify how different imaging features can correspond to the same genetic pathway. Specifically, if inter-cluster pairwise distances divided by the average of the intra-cluster pairwise distances is large, then we will call the two genetic pathways far away from each other.

3 Results

We have trained a multi-label classification network implemented with a multi-layer perceptron (MLP). Our network obtained a training accuracy of 88.5 ± 0.4 % and 83.2 ± 1.1 %. accuracy on held-out test data (see §5.2). We use the feature vectors z_is (see §2) for subjects with single-pathway alterations to calculate InPCA and obtain the top 4 principal components. New data, e.g., features of test subjects, or features of subjects with multiple-pathway alterations can be embedded into the same coordinates using eigenvectors of InPCA (see §5.3).

See Fig. 1. The top 3 principal components are associated with distinct pathway mutation patterns: (1) Increasing values of PC1 are associated with primary involvement of P53 then RTK/PI3K; (2) Increasing values of PC2 were associated with P53 then MAPK; (3) PC3 generally distinguishes MAPK from other pathways. The overlaps between pathway groups indicated these pathways were potentially correlated through their imaging phenotypes. From the average pairwise Bhattacharyya distance matrix, we observed that RTK and P53, as well as RTK and RB1 are the two most distant pathway pairs, while the distance between PI3K and MAPK is the smallest. These findings can be potentially validated in clinical literature on signaling pathways. For instance, there is evidence that a significant amount of cross-talk exists between the PI3K and MAPK pathways [16]. It is also known that RTKs are major upstream regulators of PI3K/Akt signaling [17], where our results show the inter-pathway distances between RTK and PI3K pathways are as small as 0.72 and 0.51 for subjects with single pathway and multiple-pathway alterations respectively.

4 Potential Negative Social Impact

Glioblastoma (GBM) is the most common malignant tumor with a grim prognosis. In spite of extensive efforts, new drugs and technologies over the past few decades, little has changed in terms of patient outcome. This paper seeks to develop new techniques to extract signatures of mutations from *in vivo* imaging and thereby shed light upon genetic and spatial heterogeneity of GBM. We believe that this line of research can lead to a holistic understanding of GBM that goes beyond the single tissue sample analysis in the current clinical practice.

This research is in very preliminary stages and its merit lies in acting as an aid to scientific investigations of mutations that cause GBM. The ideas discussed in our paper are very far from any kind of deployment. We therefore do not foresee any negative social impacts of this research. Our methods can also be applied to other heterogeneous diseases, such as different types of cancer or neurodegenerative diseases to study the genetic underpinnings of imaging phenotypes.

That said, the sample size of our data is fundamentally small. Although the size of the dataset keeps growing as more samples from the clinicians are obtained and nested cross-validation has been applied to ensure that our findings are not over-fit to the data, it is possible that some of our conclusions could change with more available data. Therefore, the biological findings obtained from this work should be carefully validated by clinicians and researchers. More broadly, small sample size is a hallmark of problems in the clinical sciences. To address it effectively, we need to inject knowledge from biology into machine-learning based methods [18], e.g., known relationships between pathways and co-occurring mutations in the clinical literature.

References

[1] Roger Stupp, Sophie Taillibert, Andrew Kanner, William Read, David M Steinberg, Benoit Lhermitte, Steven Toms, Ahmed Idbaih, Manmeet S Ahluwalia, Karen Fink, et al. Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: a randomized clinical trial. *Jama*, 318(23):2306–2316, 2017.

[2] Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *cell*, 155(2):462–477, 2013.

[3] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[4] Andrea Sottoriva, Inmaculada Spiteri, Sara GM Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.

[5] Anahita Fathi Kazerooni, Spyridon Bakas, Hamidreza Saligheh Rad, and Christos Davatzikos. Imaging signatures of glioblastoma molecular characteristics: a radiogenomics review. *Journal of Magnetic Resonance Imaging*, 52(1):54–69, 2020.

[6] Hugo JWL Aerts. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA oncology*, 2(12):1636–1642, 2016.

[7] C Carl Jaffe. Imaging and genomics: is there a synergy?, 2012.

[8] Anahita Fathi Kazerooni and Christos Davatzikos. Computational diagnostics of gbm tumors in the era of radiomics and radiogenomics. In *International MICCAI Brainlesion Workshop*, pages 30–38. Springer, 2020.

[9] Apurva Singh, Rhea Chitalia, and Despina Kontos. Radiogenomics in brain, breast, and lung cancer: opportunities and challenges. *Journal of Medical Imaging*, 8(3):031907, 2021.

[10] Katherine N Quinn, Colin B Clement, Francesco De Bernardis, Michael D Niemack, and James P Sethna. Visualizing probabilistic models and data with intensive principal component analysis. *Proceedings of the National Academy of Sciences*, 116(28):13762–13767, 2019.

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[13] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[14] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[15] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347. 2008.

[16] Wei Zhang and Hui Tu Liu. Mapk signal pathways in the regulation of cell proliferation in mammalian cells. *Cell research*, 12(1):9–18, 2002.

[17] Nahal Haddadi, Yiguang Lin, Glena Travis, Ann M Simpson, Najah T Nassif, and Eileen M McGowan. Pten/ptenp1:'regulating the regulator of rtk-dependent pi3k/akt signalling', new targets for cancer therapy. *Molecular cancer*, 17(1):1–14, 2018.

[18] Mohamed Omar, Lotte Mulder, Tendai Coady, Claudio Zanettini, Eddie Luidy Imada, Wikum Dinalankara, Laurent Younes, Donald Geman, and Luigi Marchionni. Biological constraints can improve prediction in precision oncology. *bioRxiv*, 2021.

[19] Christos Davatzikos, Saima Rathore, Spyridon Bakas, Sarthak Pati, Mark Bergman, Ratheesh Kalarot, Patmaa Sridharan, Aimilia Gastounioti, Nariman Jahani, Eric Cohen, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of medical imaging*, 5(1):011018, 2018.

[20] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

Table 1: Key driver pathways and genes: Each pathway contains several gene alterations and if one of the genes is altered, then we regard the pathway which the gene belongs being altered. The numbers of patients with each gene alteration are summarized as following.

Pathway	Gene alteration	# Subjects (Total: 286)	Fraction (%)
RTK	EGFR	75	26.2
	PDGFRA	14	4.9
	MET	12	4.2
	FGFR2	2	0.7
PI3K	PIK3CA	26	9.1
	PIK3R1	26	9.1
	PTEN	104	36.4
MAPK	NF1	48	16.8
	BRAF	2	0.7
P53	TP53	100	35
	MDM4	2	0.7
RB1	CDKN2A	3	1.0
	RB1	21	7.3

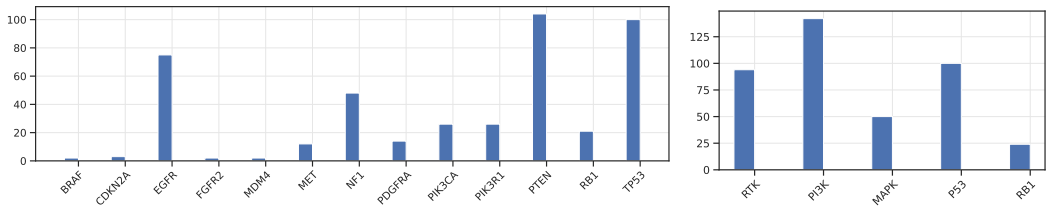

Figure 2: Left: Number of subjects which have alterations for the 13 genes in out dataset. The dataset has a large imbalance which makes learning a good representation difficult, this is even more challenging because of the limited sample size. **Right:** Number of subjects which have alterations for the 5 different pathways considered in this paper.

5 Appendix

5.1 Dataset

Multi-parametric MRI (mpMRI) scans (T1, T1-Gd, T2, T2-FLAIR, DSC, DTI) of 286 subjects diagnosed with glioblastoma were retrospectively collected. Radiomics features, including histograms, morphologic and textural descriptors, were derived using Cancer Imaging Phenomics Toolkit (CaPTk)[19]. For the data preprocessing, the subjects with any missing values in the imaging features are excluded. Highly correlated imaging features ($\rho > 0.95$) and features with identical values for every subject are removed. The total number of the imaging features included is 3480 after all above processing steps.

For the genomics data, genetic markers were obtained through next generation sequencing (NGS) panel on the resected tumor samples from the patients. A total number of 27 genes are included in the NGS panel. Patients with IDH1 or IDH2 mutations are excluded. Our final cohort consists of $n = 286$ IDH-wildtype patients with mpMRI data available.

5.2 Network and Implementation Details

The multi-label classification network is implemented as a five-layer multi-layer perceptron (MLP) to avoid overfitting in the small data regime. Each of the first four layers consists of a linear layer, rectified linear units (ReLU) nonlinear activation function, batch-normalization and dropout for regularization. After training the model, the logits from the final layer will be the input of a softmax function to convert into probability distributions, which facilitates the application of InPCA. The

Table 2: Nested CV accuracies (%)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
Train	88.78	88.65	88.89	88.45	87.72	88.50 ± 0.42
Test	81.95	83.27	81.93	84.41	84.40	83.19 ± 1.10

network is trained in a nested cross-validation (CV) scheme with a 5-fold inner loop CV for hyper-parameter tuning and 5-fold outer loop to estimate the test error on completely held out data. Through this procedure, we found the optimal hyper-parameters to be a dropout rate of 0.2 and learning rate of 0.01 for stochastic gradient descent (SGD). The training and validation accuracies for the 5 outer loops are shown in the Table 2.

5.3 Embedding new data into InPCA

Denote N as the number of training subjects, L as the number of test subjects, $\mathbf{1}_N$ as a $N \times N$ matrix with each element as $\frac{1}{N}$, $\mathbf{1}_{NL}$ as a $N \times L$ matrix with each element as $\frac{1}{N}$, X and X_{test} as the probability matrix obtained from softmax function by input training and test subjects, E as the selected eigenvectors derived from training data and the diagonal of Λ as the corresponding eigenvalues, the embedding can be calculated similarly to that of Kernel PCA[20], which is the following:

$$\begin{aligned} W &= \tilde{\varphi}(X)^\top \tilde{\varphi}(X) \\ &= [\varphi(X) - \varphi(X)\mathbf{1}_N]^\top [\varphi(X) - \varphi(X)\mathbf{1}_N] \\ &= D - D\mathbf{1}_N - \mathbf{1}_N^\top D + \mathbf{1}_N^\top D \cdot \mathbf{1}_N \end{aligned}$$

$$\begin{aligned} W_{\text{test}} &= \tilde{\varphi}(X_{\text{test}})^\top \tilde{\varphi}(X) \\ &= [\varphi(X_{\text{test}}) - \varphi(X)\mathbf{1}_{NL}]^\top [\varphi(X) - \varphi(X)\mathbf{1}_N] \\ &= D_{\text{test}} - D_{\text{test}}\mathbf{1}_N - \mathbf{1}_{NL}^\top D + \mathbf{1}_{NL}^\top D\mathbf{1}_N \end{aligned}$$

$$T_{\text{proj}} = W_{\text{test}} E \Lambda^{-\frac{1}{2}}$$