Group Presentation

12.24 by Xiao Zhou

Outline

Part I: ImageGenomics Project

Part II: Kolmogorov–Arnold Networks (Paper)

ImageGenomics
Integrative Analysis of Genomic and fMRI Data for Brain Disorder Prediction

Background

Brain-related Diseases

Previous and Related Works

- Multi-modal deep learning from imaging genomic data for schizophrenia classification
- Deep Learning with Neuroimaging and Genomics in Alzheimer’s Disease
- Multimodal deep learning to predict prognosis in adult and pediatric brain tumors

Motivation of Integration Analysis

- Brain disorders are influenced and/or shown in both data
- Limited related studies on combining these two specific data, on a large scale

Primary Goals

Predict brain-related diseases (e.g., PD, AD, SCZ, BPD, ASD) by integrating genomic and fMRI data, thus showing the potential connections between (specific) genes and fMRI information.

Specific Objectives

Create robust models that used both data, with enhance performance.

Association analysis and interpretability of models for clinical insights.

Data Sources

UK Biobank (~500k)

- Imaging data - Functional connectivity matrices derived from fMRI scans (~40k)
- Genomic data - TOPMed imputed SNP data. (~490k)

Imaging Data

fMRI (functional Magnetic Resonance Imaging):

Measures brain activity by detecting changes in blood oxygen levels.

FC matrices (419*419 - 400 cortical + 19 non-cortical areas):

Converted from fMRI that shows brain regions' communications (rest)

https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=25741

Genomic data

TOPMed Imputation (~490k):

TOPMed uses a diverse reference panel to fill in missing genetic data, enhancing SNP data quality. (Number of SNPs: 321517)

Filtering (~40k):

Dropped entries that does not have fMRI data.

https://biobank.ndph.ox.ac.uk/crystal/field.cgi?id=21007

Labels Distribution

Before filtering, there are ~5k combined labels for the diseases (i.e. AD, PD, …)

After filtering, there are <300 combined labels available.

Label Enlarged

Enlarged group based on 29000 (https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=29000)

Group A: Severe mental health, Group B: Anxiety-related, Group C: Eating disorders (~31k intersections)

Total: ~40k

Data Status

Group A: Severe mental health conditions (~6,000 cases).

Ratio: 85-15

Group B: Anxiety-related disorders (~5,300 cases).

Ratio: 87-13

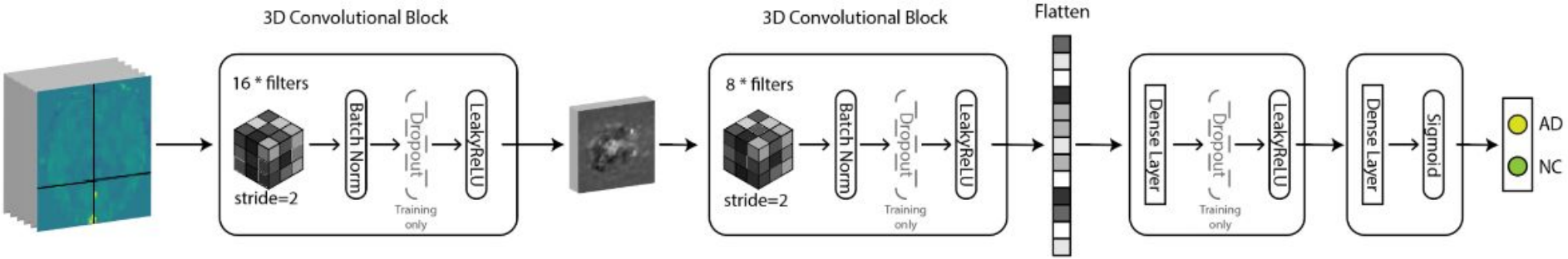
Group C: Eating disorders (~300 cases).

Ratio: 99-1

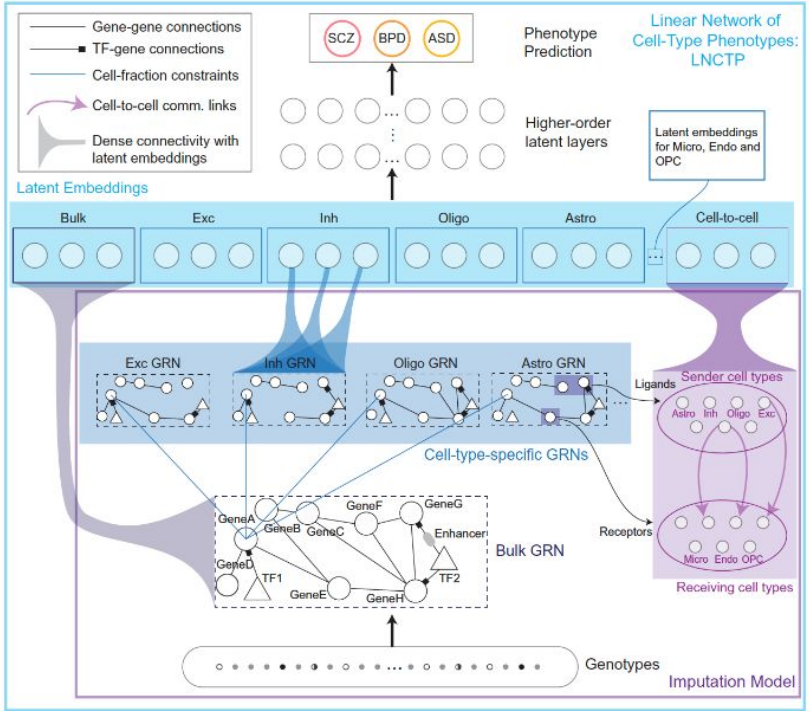
Input = SNPs, FC matrices; output = Group X (X = A | B | C)

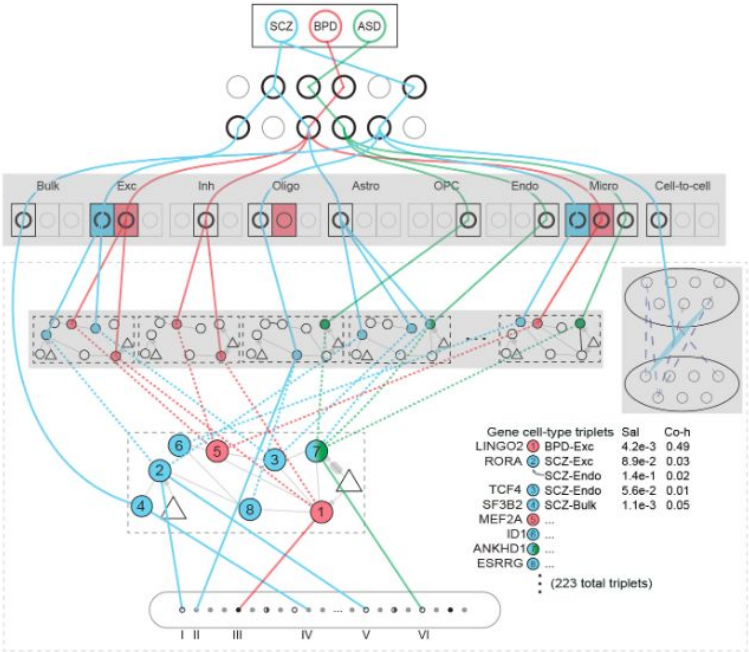
Modeling Strategies

Imaging Model

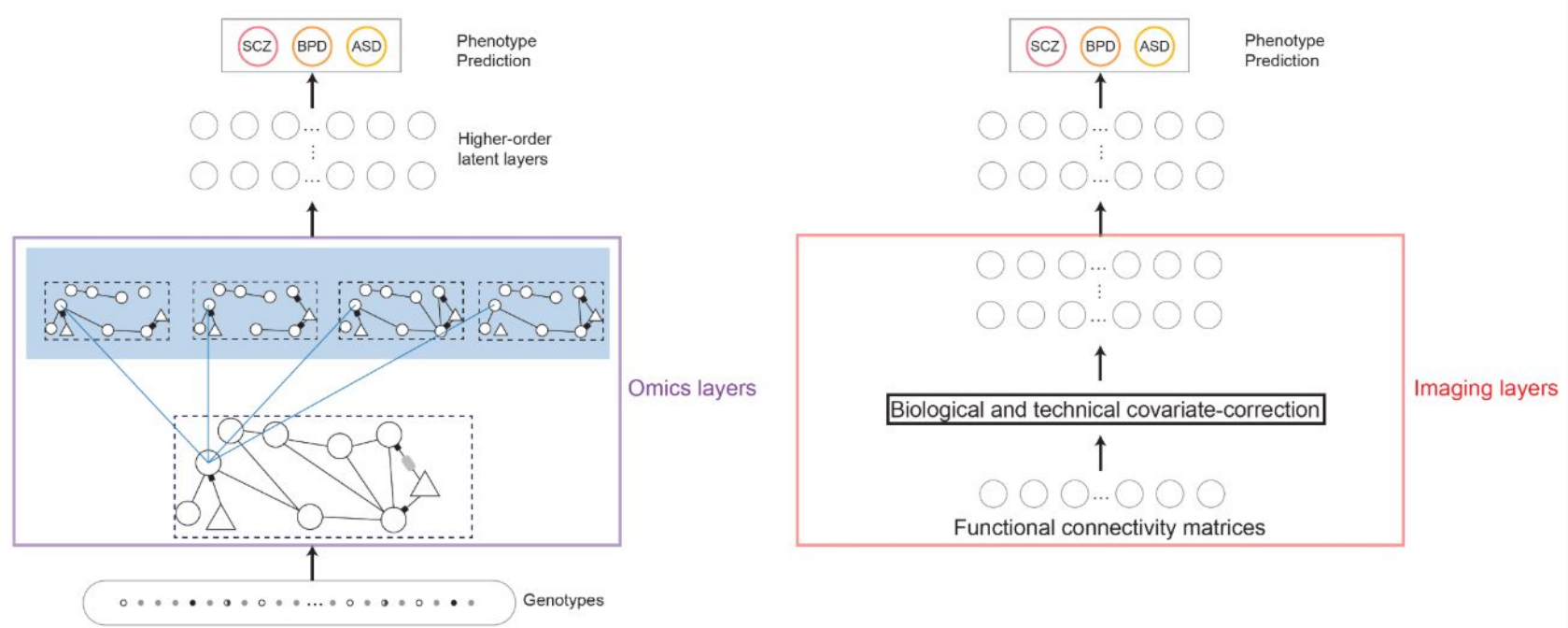


Genomic Model

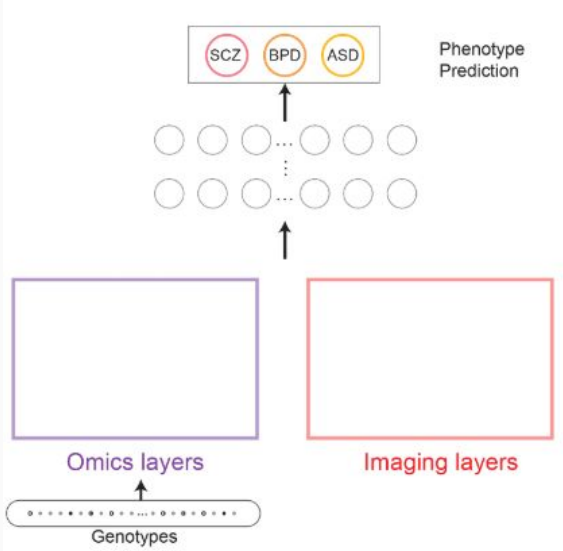


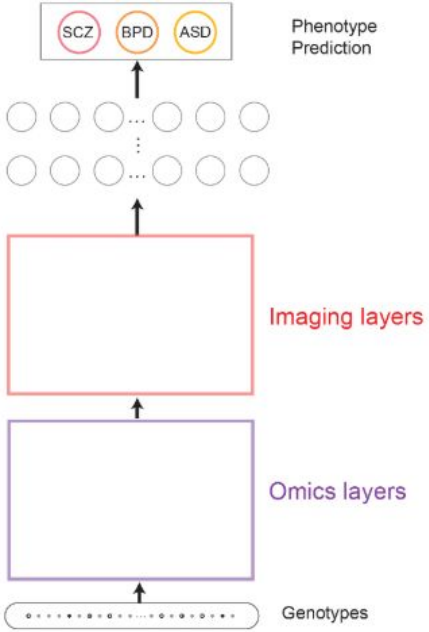


Integration



Integration





Ablation Studies

Batch Correction

PCA*

KAN*

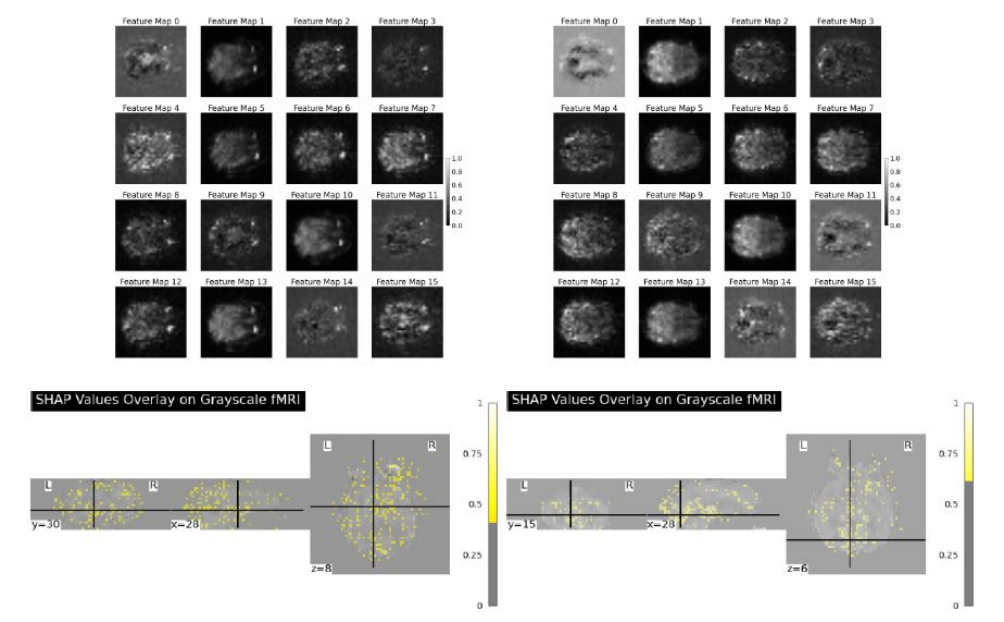
Interpretability Techniques

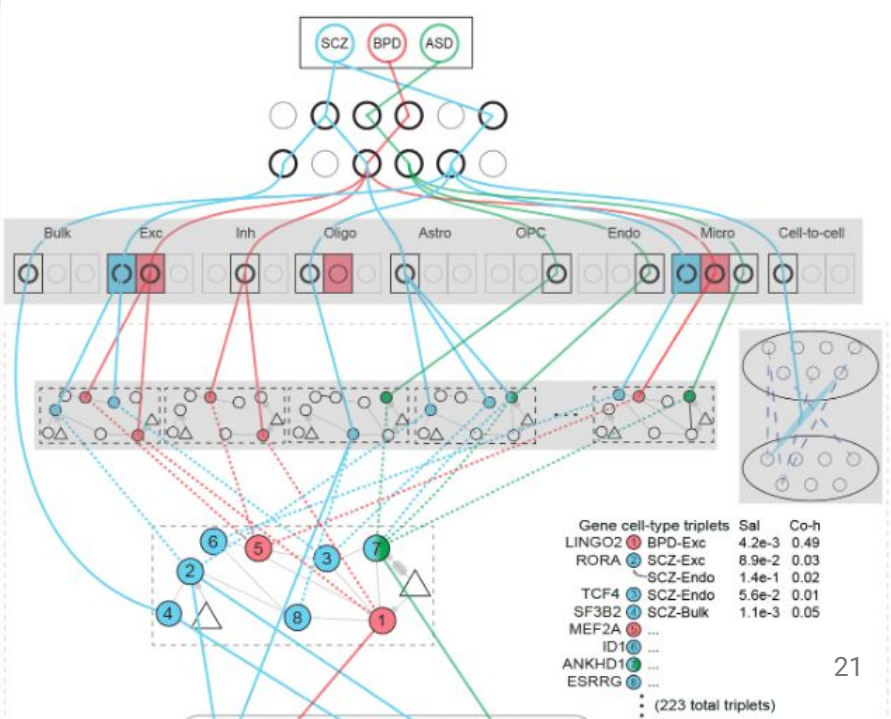
Feature Maps - Captured by CNN layers

SHAP Values - Captures each features’ contribution to prediction

LNCTP Weights - Captures the connection between genes and disorders

Interpretability





Performance Metrics

Current:

Model’s prediction performance

Need:

Model’s interpretability performance

Current (early) Results

	Data	AUC	ACC
Sleep Duration	FC matrices	0.56	0.65
Group A	FC matrices	0.57	0.85

Comparative Analysis with Other Models

	XAI Paper	BT Paper
Objective	Classify schizophrenia	Classify brain tumor
Methodology	sMRI, fMRI, Gene	histopathology+gene
Result	79.1% accuracy	83.6% accuracy

Interpretation of Results

Class imbalance

- Resampling, class weighting

Interesting relevance between fMRI and sleep duration

Limitations

Results are not as good as in similar studies

High computational demands due to size of the cohort

Single Cohort (UKBB)

Plans for Model Improvement

Genomic modeling results

Integration

Advanced models like KAN for better interpretation of the input and outputs.

Extensive optimizations

Other datasets (i.e. ADNI, AIBL)

KAN:Kolmogorov-Arnold Networks

Background

Kolmogorov-Arnold Theorem:

Multivariate continuous function = sum of univariate functions.

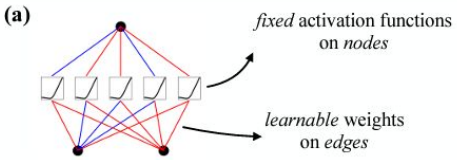
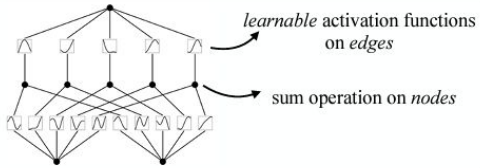
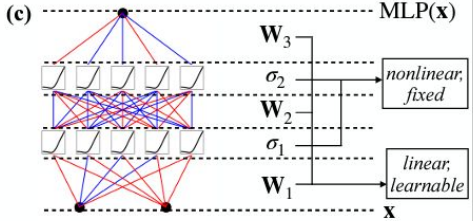
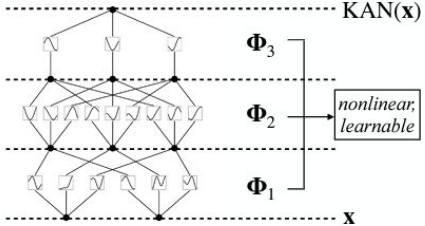
Motivation of KAN

Improve basic NN architecture (MLP) by using the KAT.

Claims:

- Better performance than MLP
- Better interpretability than MLP

Architecture (MLP vs KAN)

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	**Universal Approximation Theorem**	**Kolmogorov-Arnold Representation Theorem**
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model (Shallow)		
Formula (Deep)	$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\mathrm{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)		

$$\mathrm{spline}(x) = \sum_i c_i B_i(x)$$

Implementation and Training

Initialize training, with regularization

Investigate, then prune (automatically or manually)

Interpret & evaluate

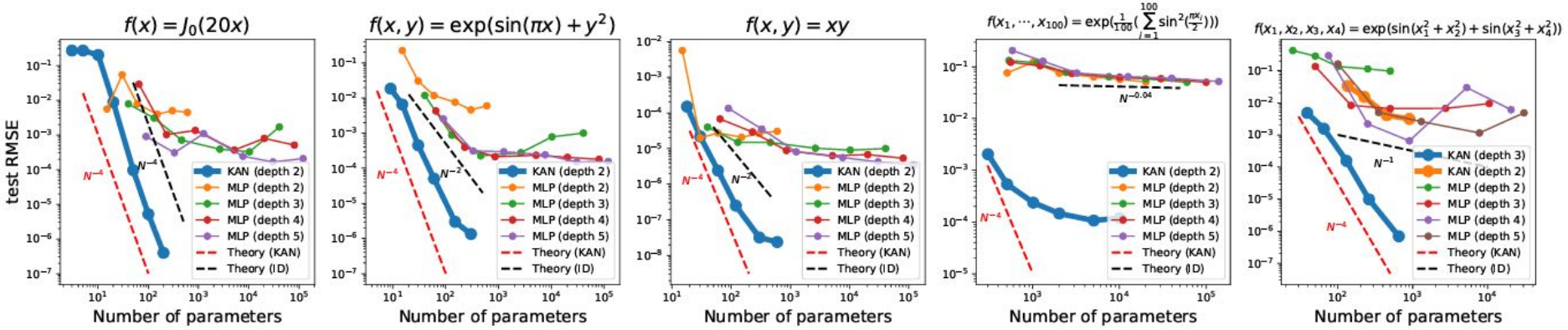
Regularization

$$\ell_{\text{total}} = \ell_{\text{pred}} + \lambda \left(\mu_1 \sum_{l=0}^{L-1} |\mathbf{\Phi}_l|_1 + \mu_2 \sum_{l=0}^{L-1} S(\mathbf{\Phi}_l) \right)$$

$$|\mathbf{\Phi}|_1 \equiv \sum_{i=1}^{n_{\text{in}}} \sum_{i=1}^{n_{\text{out}}} |\phi_{i,j}|_1$$

$$S(\mathbf{\Phi}) \equiv -\sum_{i=1}^{n_{\text{in}}} \sum_{j=1}^{n_{\text{out}}} \frac{|\phi_{i,j}|_1}{|\mathbf{\Phi}|_1} \log \left(\frac{|\phi_{i,j}|_1}{|\mathbf{\Phi}|_1} \right)$$

Results


$f(x) = J_0(20x)$
$f(x, y) = \exp(\sin(\pi x) + y^2)$
$f(x, y) = xy$
$f(x_1, \cdots, x_{100}) = \exp(\frac{1}{100}(\sum_{i=1}^{100} \sin^2(\frac{\pi x_i}{2})))$
$f(x_1, x_2, x_3, x_4) = \exp(\sin(x_1^2 + x_2^2) + \sin(x_3^2 + x_4^2))$
test RMSE
Number of parameters
N^{-4}
N^{-2}
$N^{-0.04}$
N^{-1}
KAN (depth 2)
KAN (depth 3)
MLP (depth 2)
MLP (depth 3)
MLP (depth 4)
MLP (depth 5)
Theory (KAN)
Theory (ID)

Parameters comparison

KAN: $O(N^2LG)$

MLP: $O(N^2L)$

L: depth; N: layers; G: #intervals

Integration Strategies

Replacing the Dense layer in the final layers

Feasibility study on raw data

Evaluate interpretability enhancement

Limitations

Computational intensity - large number of univariate functions

Trains ~10x slower than MLP, given same number of parameters.

Implementation complexity

Mathematical understanding is limited

Acknowledgements

Special thanks to Jonathan, Prashant, Suchen, and Kexing.

QA