Topological Data Analysis (TDA) for Identification of miRNAs as Biomarkers for Human Performance

Christopher Dean, Christopher Chia, Rajesh Naik & Ryan Kramer

MINEDXAI

October 24, 2024

ABOUT US

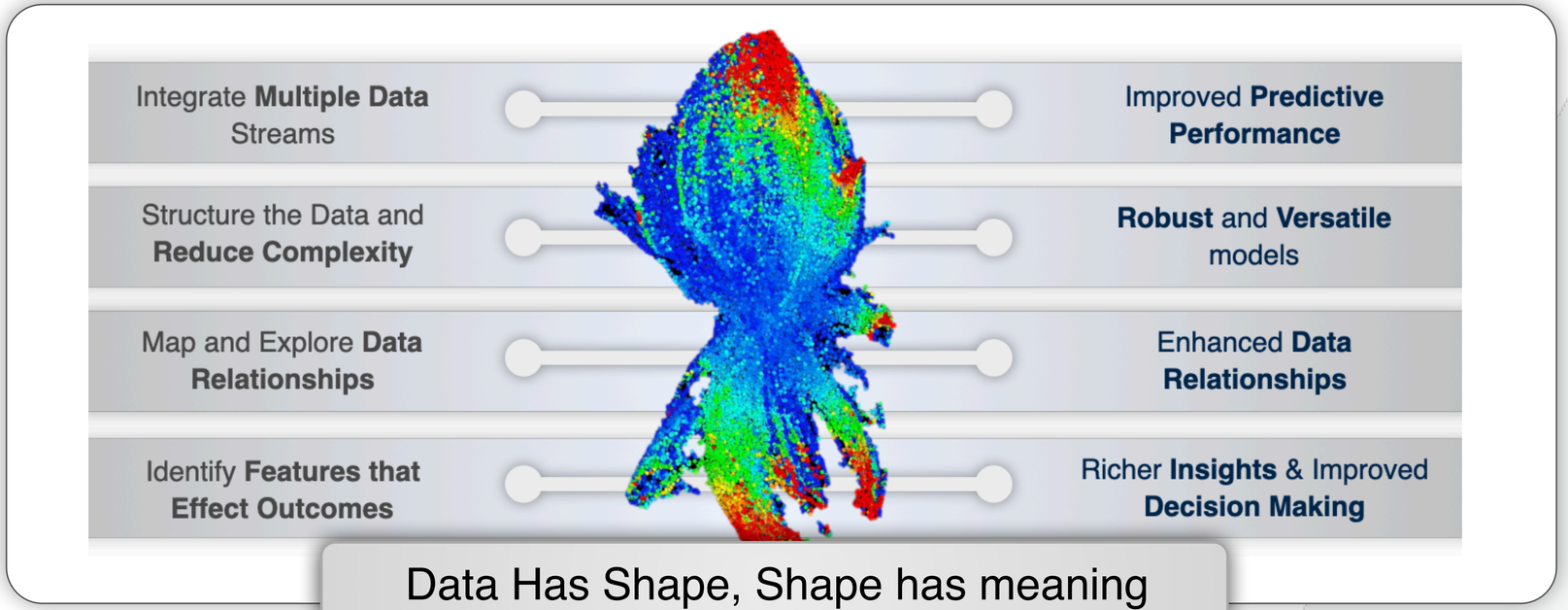
- Creating next-generation machine intelligent solutions using explainable artificial intelligence (XAI).
- Technology seeded from a DARPA program and demonstrated within the US Air Force Research Laboratory.
- Deep topological modeling to integrate and synthesize multi-domain knowledge for direct human interaction.
- Providing benchmarking, customized analysis, data walks, and interactive dashboard solutions and platforms.
- Serving information technology, food & beverage, wholesale distributors, and non-profit clients.

Mined XAI's Business Verticals

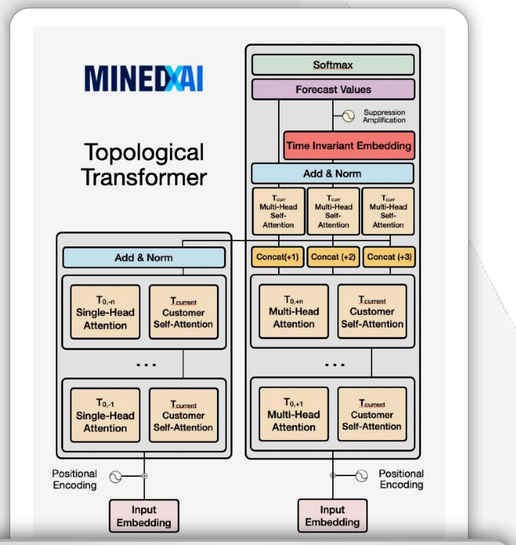
Veteran-owned Business
Estb. Oct 2020

WHY TOPOLOGICAL DATA ANALYSIS (TDA)?

- Formulate dimensionality reduced understanding of complex signature distributions
- Develop deep topological modeling capabilities that have encoder/decoder transformer-like architectures that integrate topological hierarchical decomposition (base AI sub-unit)
- Utilize unsupervised, self-supervised, attention, and supervised insights in explainable framework
- Deep topological models easily capture non-linear temporal and spatial prediction capabilities that enable easy multi-modal integration (ideal for complex biological systems)



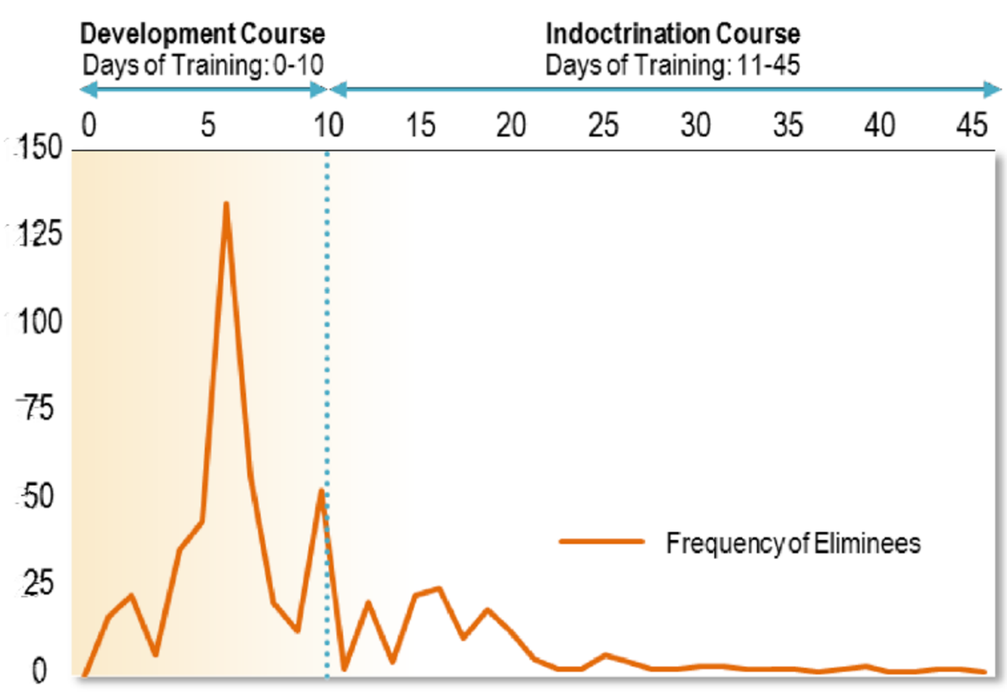
Data Has Shape, Shape has meaning



Glass-Box Architectures

AF RELEVANCE

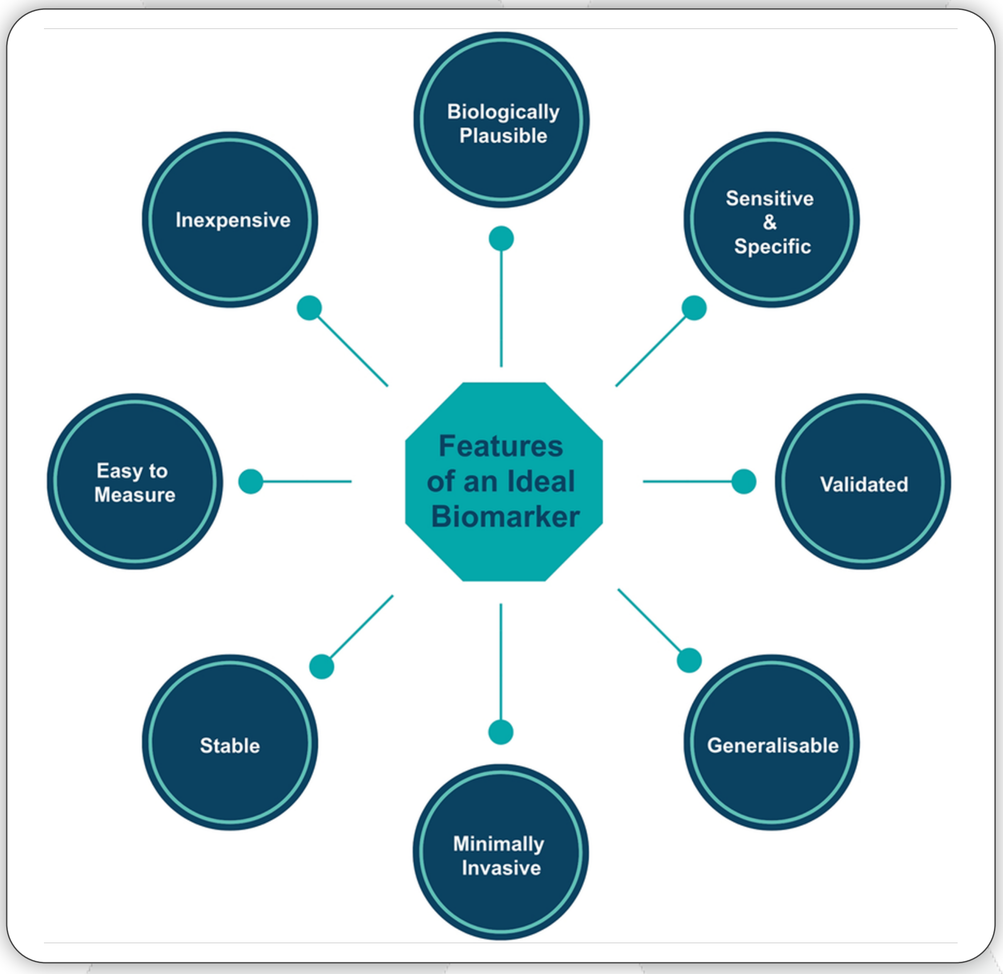
- Inadequate physical fitness continues to be a challenge for warfighter readiness
- The annual cost of attrition is roughly $620 million across the US military.
- Washout rates are significantly higher with special operators (~ 80% washout within first 10 days).
- Air Force Combat Readiness Officer (CRO) training, 20% drop occurs early in training on the day of the first physical test, with another 20% dropping out due to injury.
- Need for biomarkers as prognostic indicators of physical performance and injury.



USAF CRO Washout Data from Jul 2011- Oct 2013
Chappelle, Ramstein Aerospace Medicine Summit (2015)

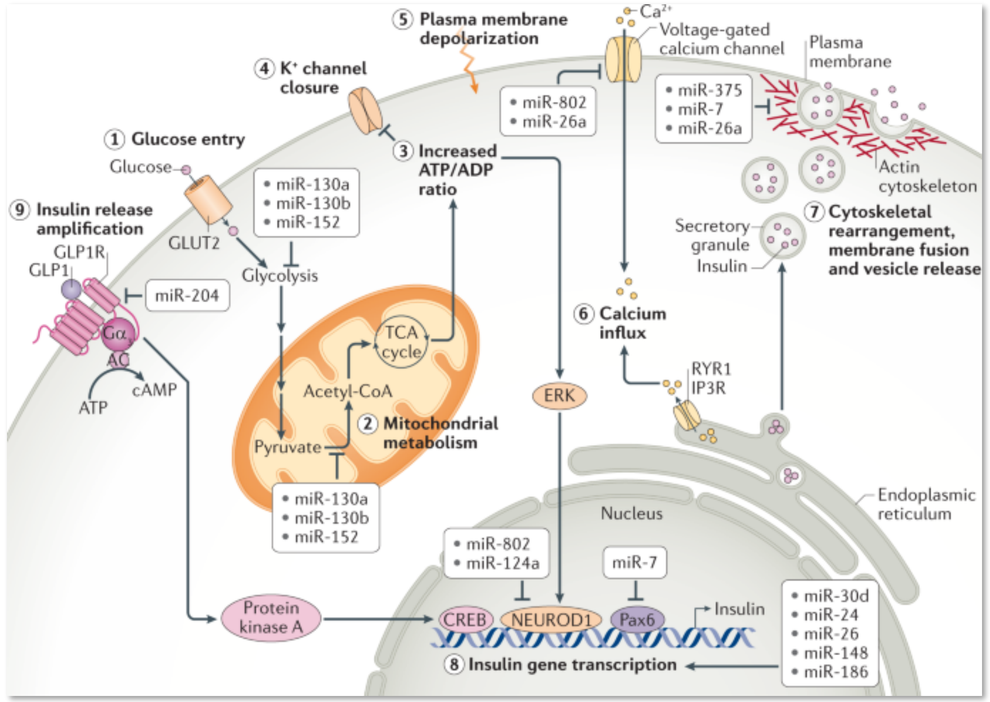
CHALLENGES WITH BIOMARKERS

- Standard statistical methods often struggle to separate meaningful signals from noise in increasingly complex biological systems.
- Challenges arise particularly with human subject studies, where sample sizes can vary widely (small vs. large), individual variability affecting the reliability and generalizability of results.
- High-dimensional datasets present difficulties in analysis, making it harder to recognize or interpret patterns effectively.
- There is a growing need for explainable models that enhance the potential for biomarker discovery and provide interpretable explanations for their predictions.



BACKGROUND ON Micro(mi)RNAS

- miRNAs are key epigenetic regulators that influence signaling pathways and transcript and protein expression in response to exercise stimuli
- Play a critical role in regulating essential biological processes, including cellular proliferation, differentiation, and cell death
- ~ 2,600 miRNAs have been identified in the human genome.
- miRNAs are also involved in exercise-induced adaptations, influencing physical performance.
- Attractive as biomarkers due to its stability and presence in blood, urine, saliva.



Nature Reviews Molecular Cell Biology 22, 425–438 (2021)

STUDY DESIGN & SAMPLE COLLECTION

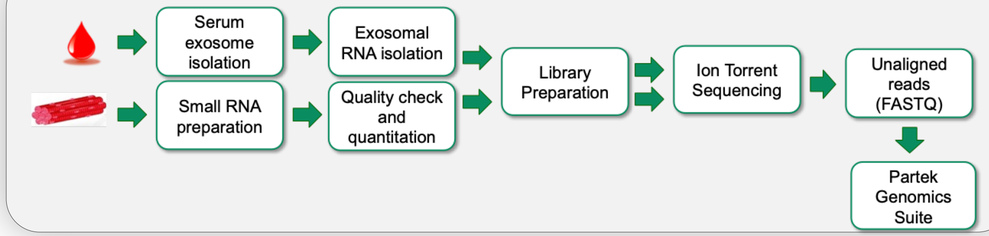
MINEDXAI

- Study previously conducted under ONR MURI “Precision High Intensity Training through Epigenetics (PHITE)” - ended in 2022.
- 2-arm blinded study randomized, exercise dose-response trial of Healthy untrained men and women 18–27 years old subjects who completed 12 weeks of traditional-intensity (TRAD) or high-intensity (HIIT) training followed by a 4-wk deconditioning period.
- Blood and muscle specimens were collected around acute exercise bouts at baseline and after 12 weeks of training. Additional biospecimens were collected after 4 weeks of deconditioning.
- Data obtained from Dr. Madhavi Kadakia (WSU) and Dr. Tim Broderick (IHMC).

Human Subjects

	Female	Male	Total
TRAD	21	22	43
HIIT	27	20	47
Total	48	42	90

Sample Collection and Processing



159 chips / 90 subjects
10-12 samples pooled per run

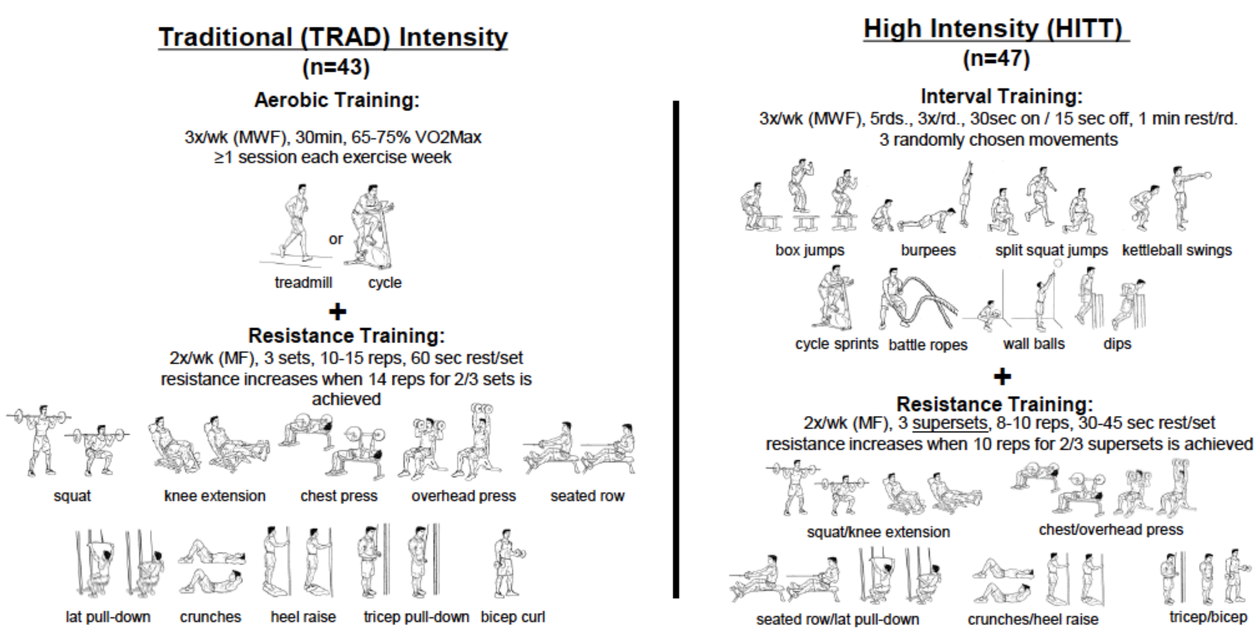
Total: 626 muscle, 806 serum samples

(Kadakia Lab, WSU)

PHITE STUDY DATASET

MINEDXAI

Experimental Setup



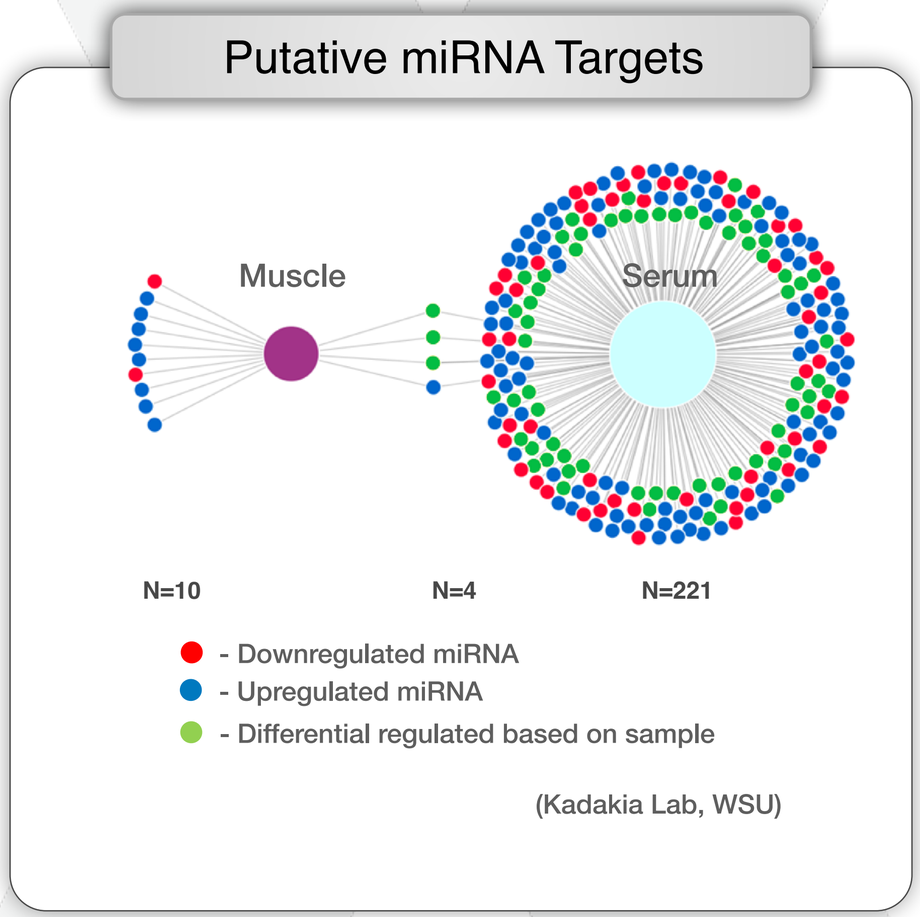
Data Collection Overview

DATA	CATEGORY	TIMEPOINTS
Functional/Phenotypic	Demographics	w0, w12
	VO2 Peak	w0, w6, w12
	Anaerobic Power	w0, w6, w12
	Strenght (1rm)	w0, w6, w12
	Biodex (Toruq)	w0, w6, w12
	Vertical Jump	w0, w6, w12
	Body Composition	w0, w12
	Strength Test	w0, w6, w12
Physiological	Myofiber Typing	w0, w6, w16
	Capillary Density	w0, w6, w16
	Mitochondia Stats	w0, w6, w16
	Serum Cytokines	w0, w12
	Serum Sex Hormones	w0
	Serum Growth Factor	w0, w12
miRNAs	Tissue miRNA Reads	w0pre, w0h3, w0h24, w12pre, w12h3, w12h24, w16rst
	Serum Exosone miRNA Read	w0pre, w0h3, w0h24, w12pre, w12h3, w12h24, w16rst

Trillions of testable hypotheses across datasets

PRELIMINARY FINDINGS FROM PHITE MURI

- Differences in baseline expression of miRNA were found between sexes.
- Upregulation of many serum miRNA immediately following exercise.
- A muscle miRNA signature of the detraining response was identified, with a majority of these miRNA upregulated versus baseline.
- miRNA signatures linked to exercise response in literature (mTOR and AMPK pathways) identified.
- MOD and HI exercise elicited distinct miRNA changes.
- *Need machine-learning models to predict performance improvement and max performance group using expression, demographics and functional features.*



SEEDLING PROJECT GOALS (1 YR EFFORT)

MINEDXAI

- Integrate functional/phenotypic, physiological and biological variables to provide more interpretable network of physical and functional interactions.
- Use machine learning (Deep Topological Modeling) to identify and link miRNAs expressed during exercise with metadata demographics, physiological and functional/phenotypic metrics.
- Identify the most important features to predict physical training outcomes

SYSTEM ARCHITECTURE OVERVIEW

MINEDXAI

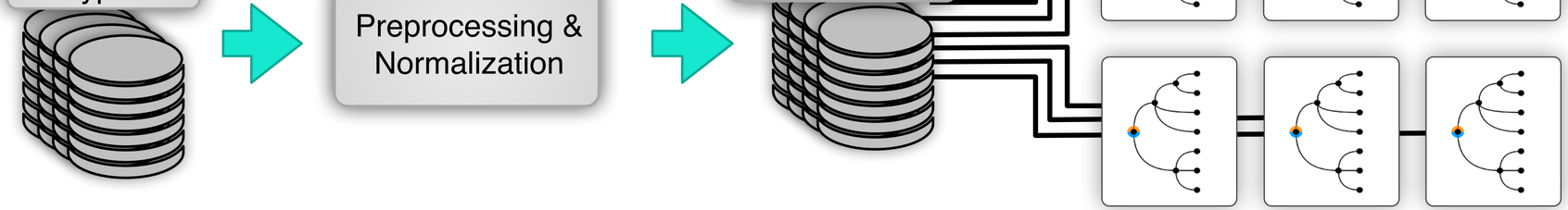
Subject Space Analysis

Individual Metadata Types

Preprocessing & Normalization

Modality Segmentation

Subject Segmentation



Capture unique distributions of complex metafeature

Feature Space Analysis

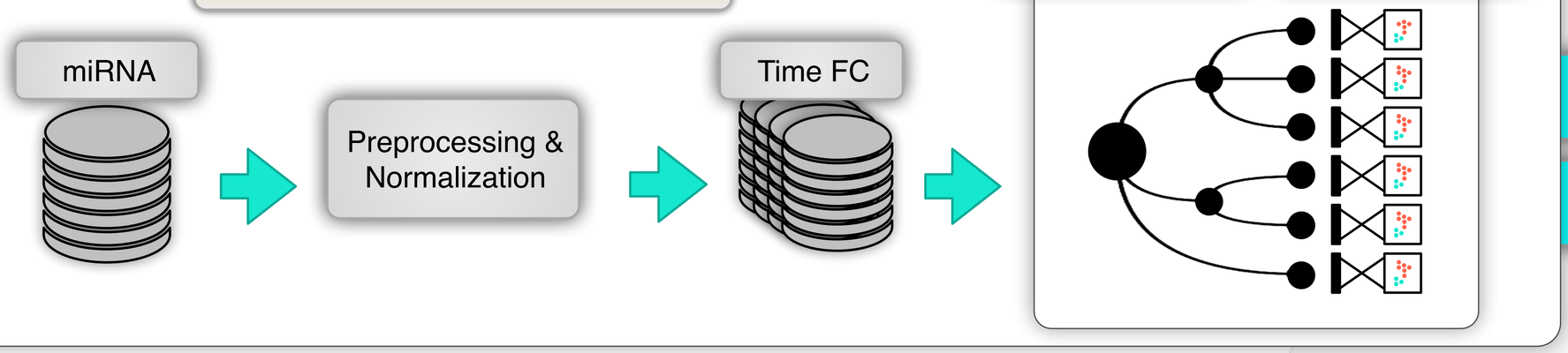
miRNA

Preprocessing & Normalization

Time FC

Feature Segmentation

Local Transpose



Advanced Hierarchical Merge Querying

Unsupervised Subject Reconstruction

Supervised Subject Reconstruction

SYSTEM ARCHITECTURE - METAFEATURE PROCESSING

MINEDXAI

Subject Space Analysis

Individual Metadata Types

Preprocessing & Normalization

Modality Segmentation

Subject Segmentation

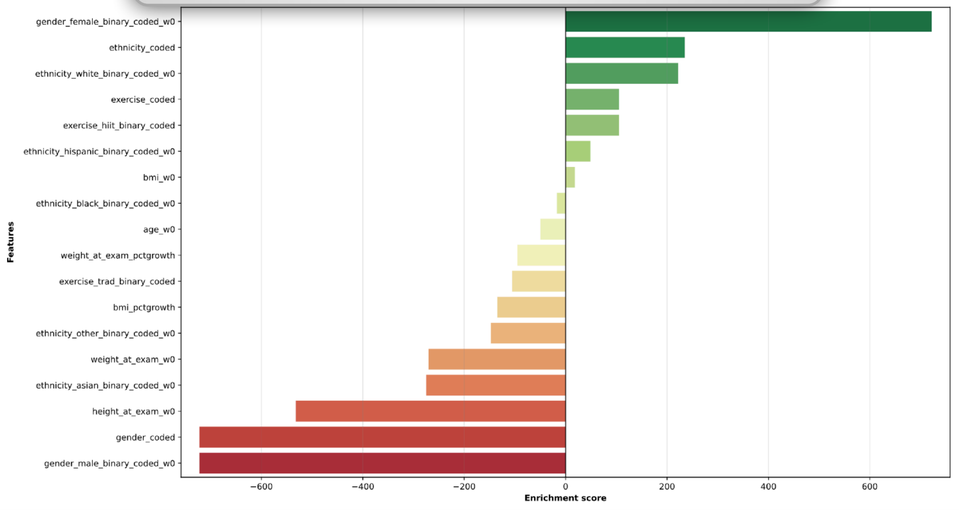
Capture unique distributions of complex metafeature

Demographic Segmentation

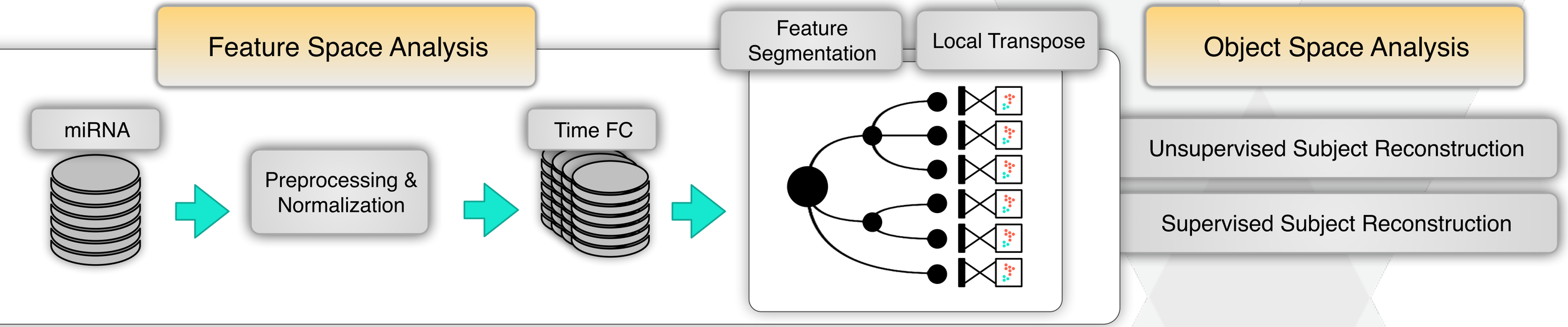
Colored by: Age

Colored by: Gender

THD Group Description



SYSTEM ARCHITECTURE- FEATURE SPACE

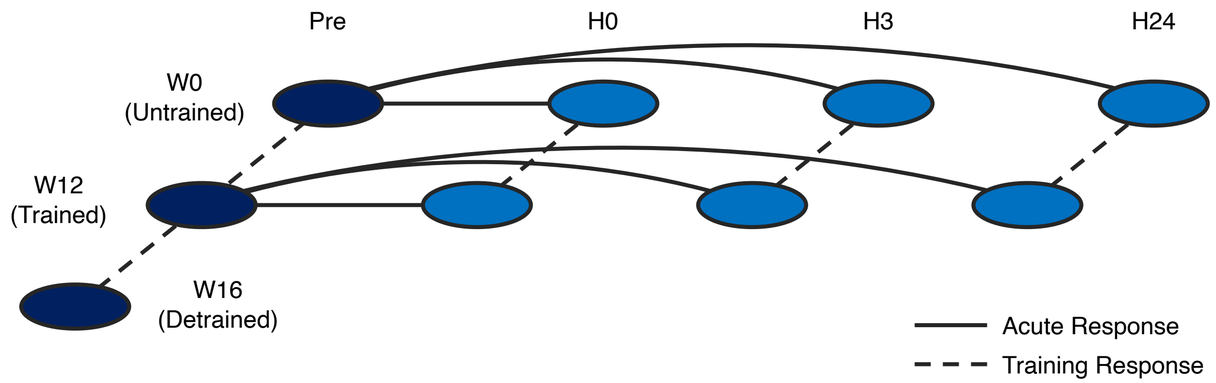


- Feature Space Analysis - Understand complex signatures across spatial or temporal patterns followed by dimensionality reduction
- Ideal for systems biological approaches that involve complex interacting pathways that are observable in population or sub-populations
  - Relies on transposing data to cluster features that share a similar signatures
  - Local subsets of features are then re-transposed to return a subject clustering
  - Total subject understanding can be developed through basic feed-forward networks

MIRNA PREPROCESSING AND NORMALIZATION

- **Traditional normalization found to be inadequate for machine learning input**
- **Input:** raw aligned quantized reads for all 90 subjects over time
- **Normalization Process**
  - RLE Normalization
  - Break out by analysis time frame
  - Convert to log-fold-change representation
  - Probabilistically down-weighted low-count/ fold-change contribution to overall signature

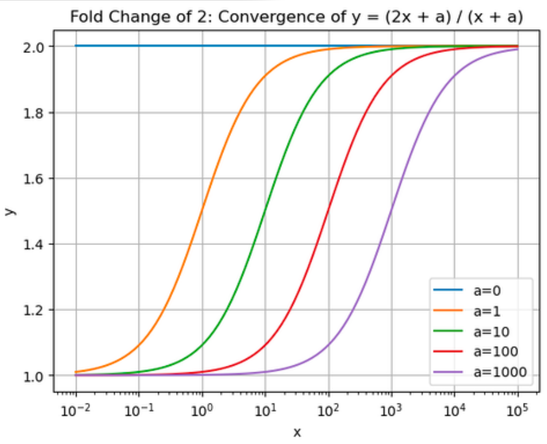
Temporal Model of miRNA Data



Log-Fold-Change Transformation

$$X_{ij}^{t_0 \to t_1} = \log_2 \left(\frac{X_{ij}^{t_1} + \alpha}{X_{ij}^{t_0} + \alpha} \right)$$

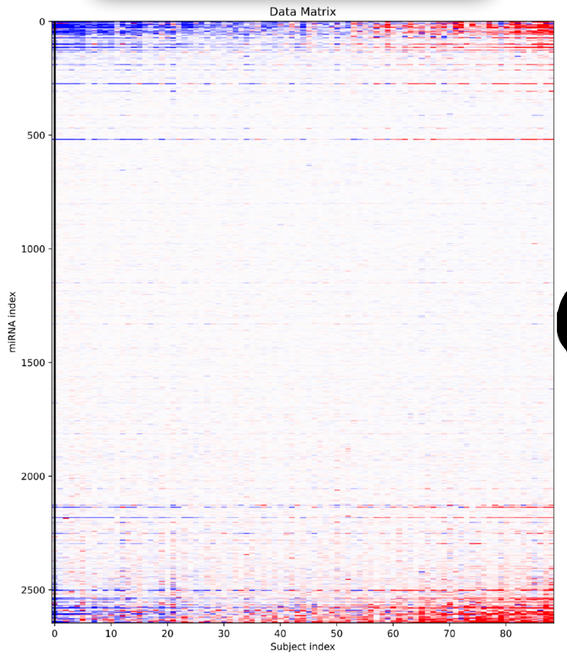
Above: log fold change from t0 to t1. Additive pseudocount parameter alpha reduces the impact of miRNA with low counts.
Right: impact of different choices of alpha on the effective fold change.



FEATURE SPACE ANALYSIS

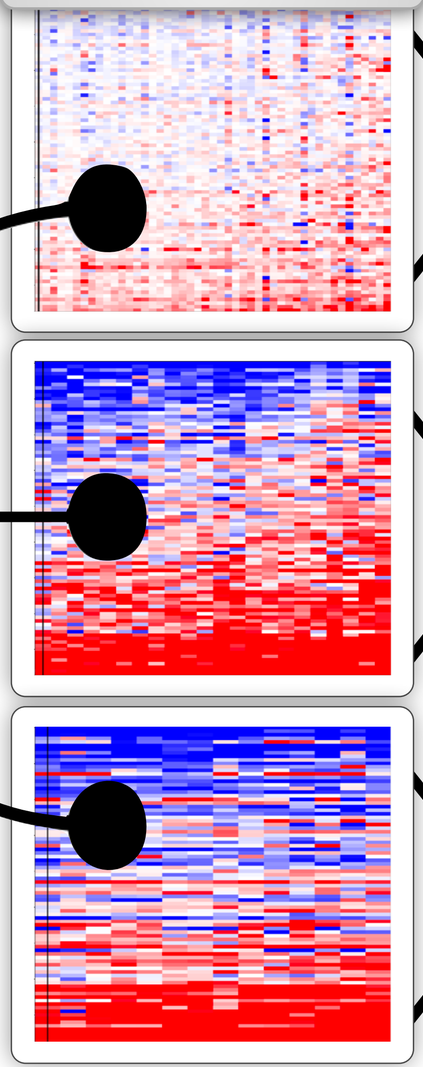
MINEDXAI

miRNA Array (ALL)

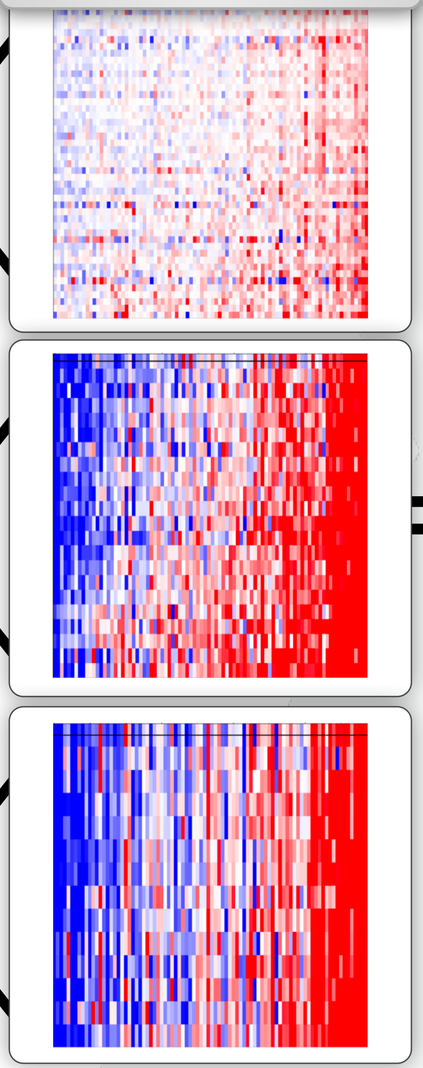


Wk 0 Pre - Wk 0 3hr Post Exercise

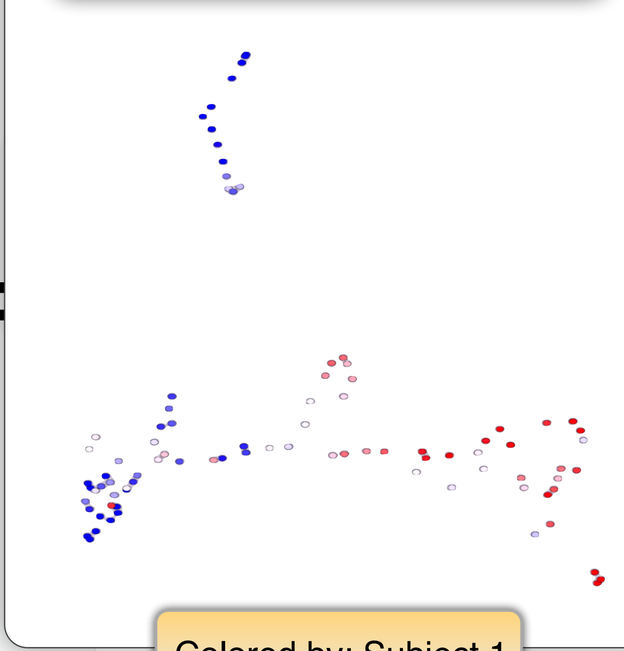
miRNA Array (Subsets)



Transposed Array

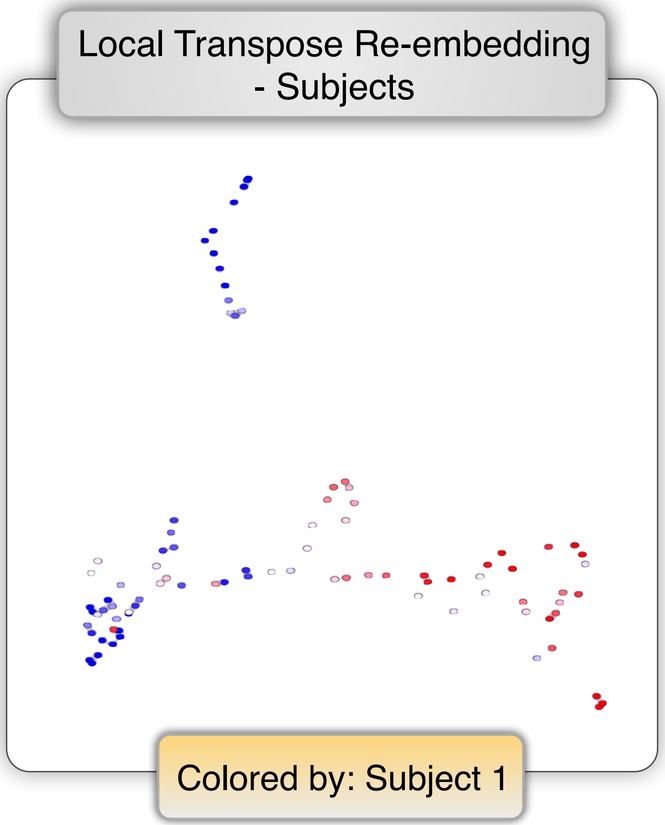


Local Transpose Re-embedding - Subjects



Colored by: Subject 1

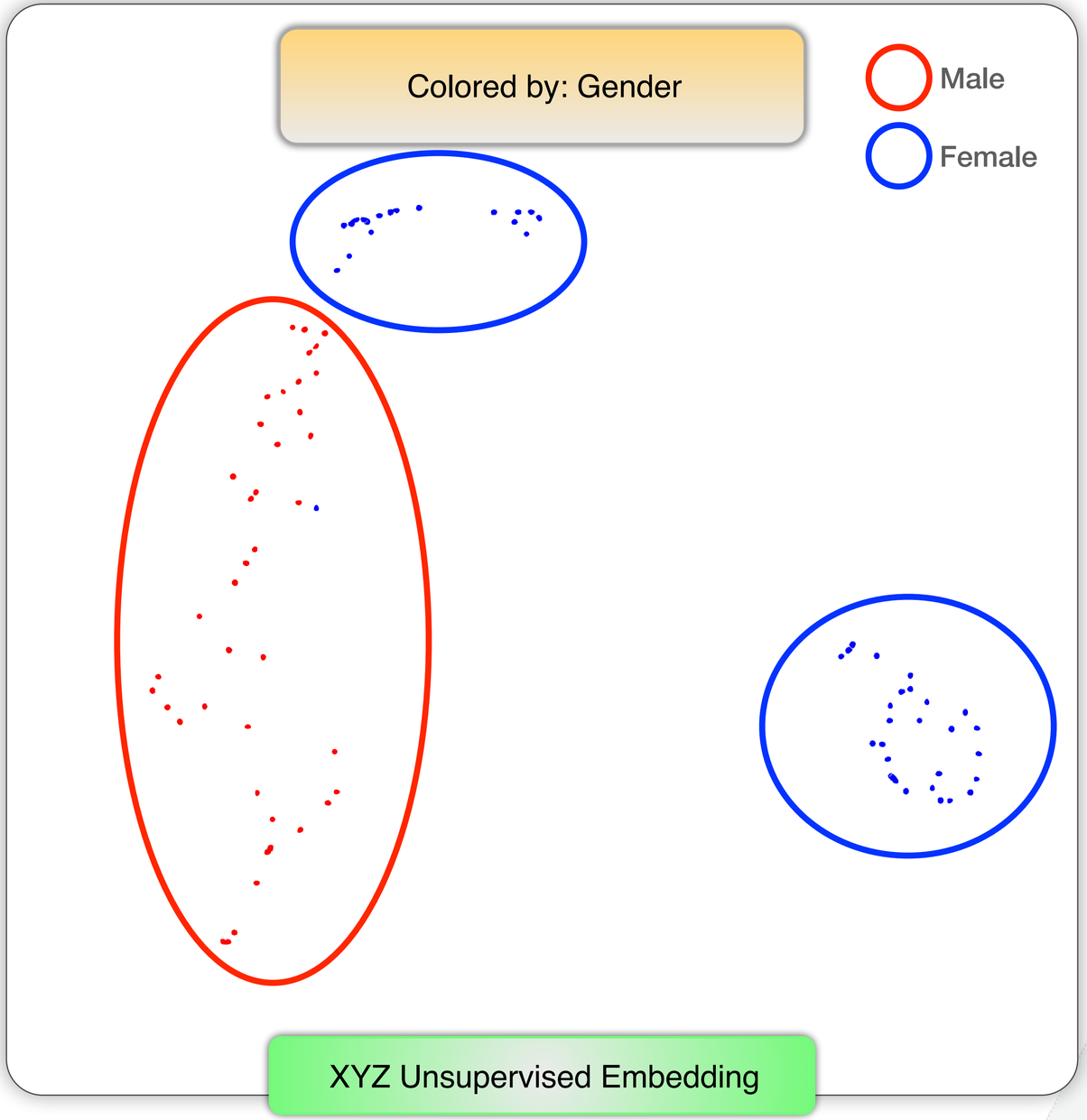
FEATURE SPACE ANALYSIS

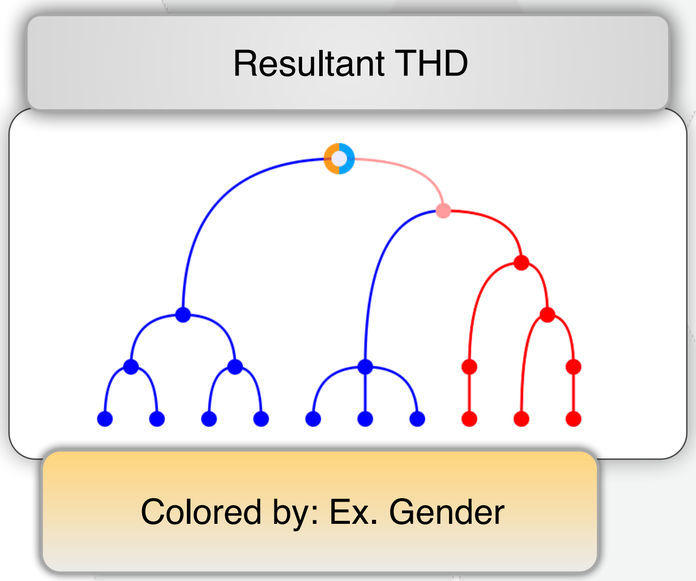


- The local transpose operation converts each feature space to a local object (subject) space understanding
- Every subject within each local transpose can be transformed as a relative position within the embedding (XYZ feature set)
- Local transpose subject spaces are then concatenated
- Fed forward within the architecture to deliver an unsupervised summary of the overall feature space

UNSUPERVISED TIMEPOINT DIFFERENCE ANALYSIS

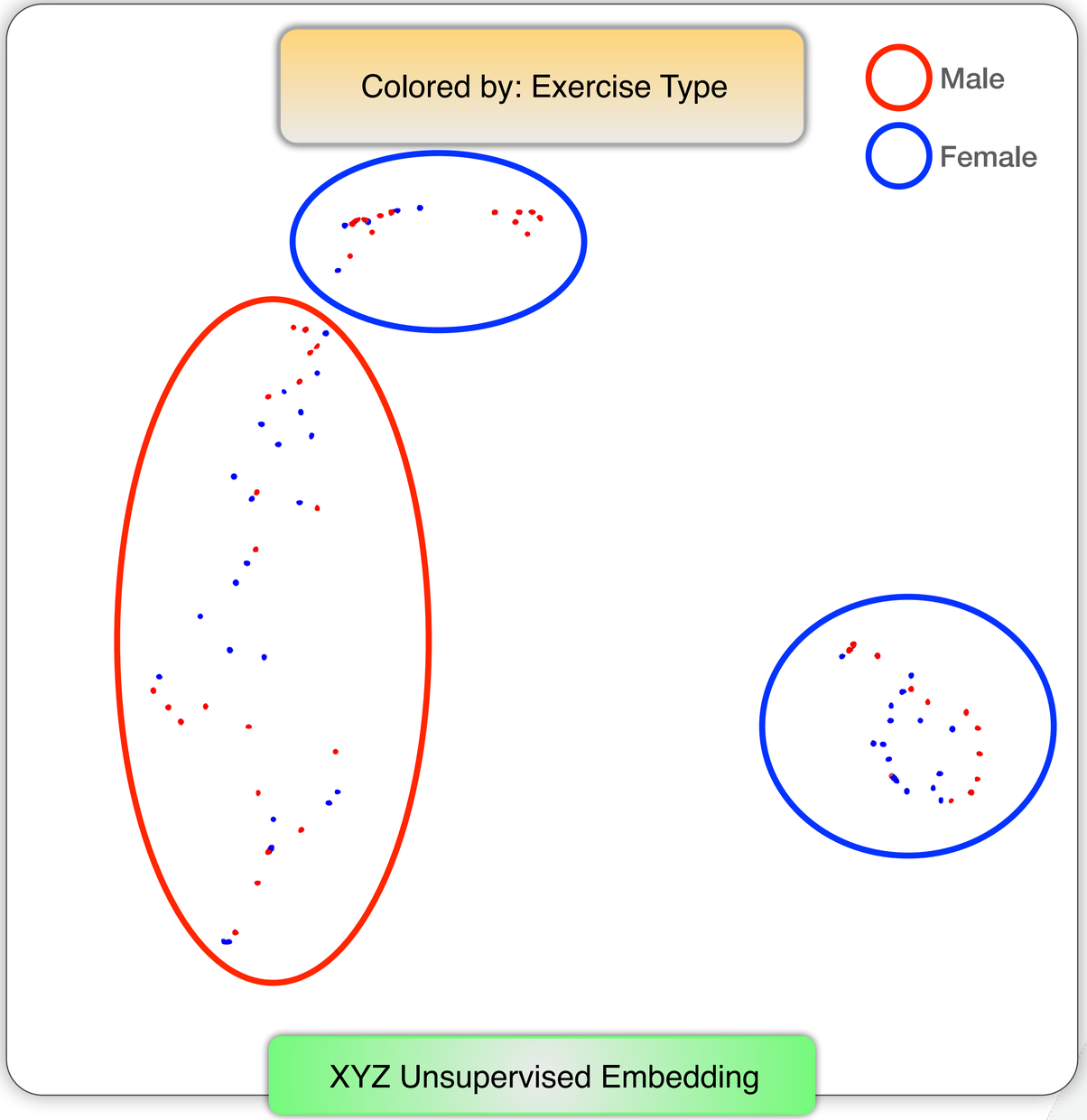
MINEDXAI

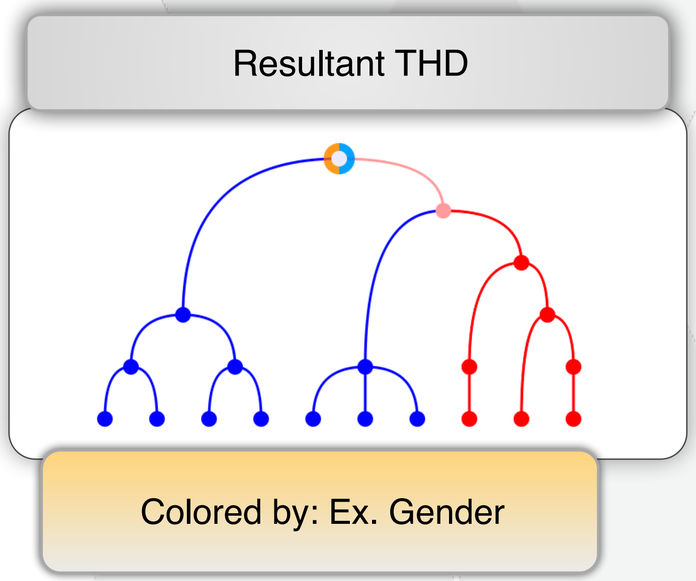




UNSUPERVISED TIMEPOINT DIFFERENCE ANALYSIS

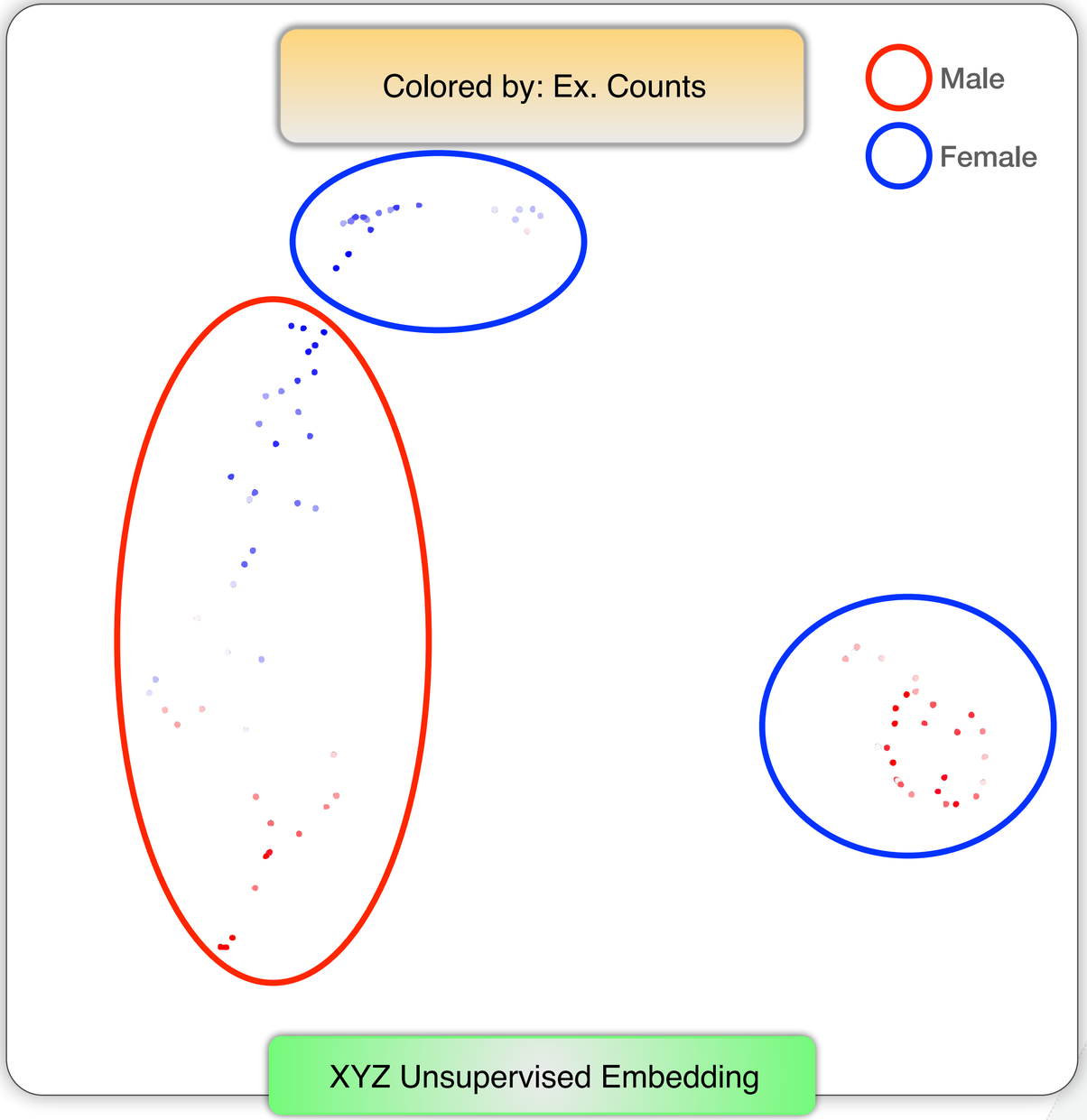
MINEDXAI

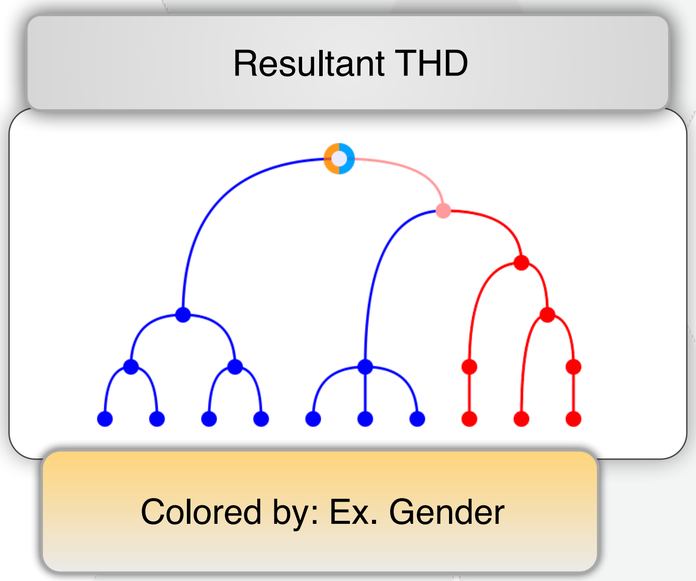




UNSUPERVISED TIMEPOINT DIFFERENCE ANALYSIS

MINEDXAI

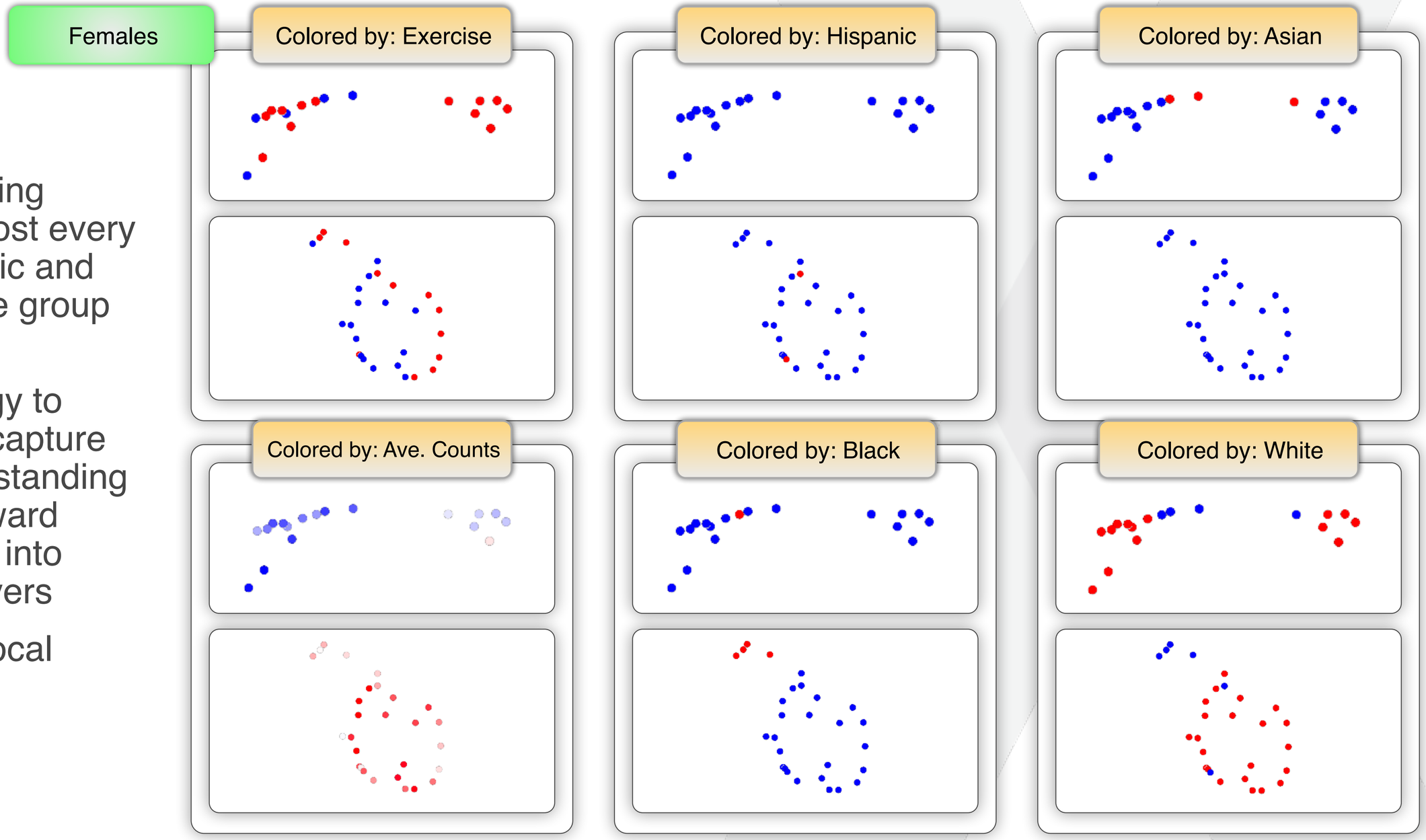




UNSUPERVISED TIMEPOINT DIFFERENCE ANALYSIS

MINEDXAI

- Complex understanding across almost every demographic and metafeature group
- Developed methodology to query and capture local understanding to feed-forward information into decoder layers
- Based on local enrichment



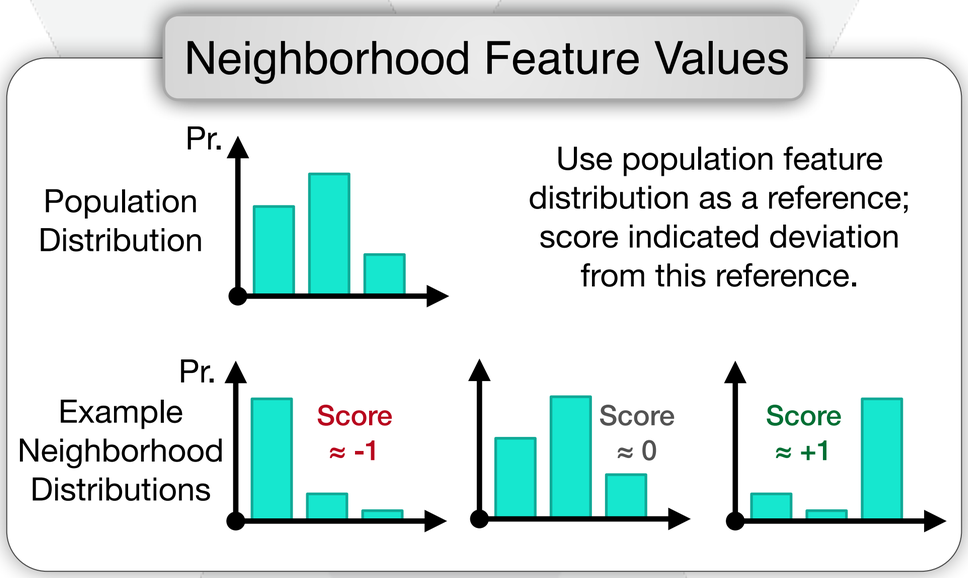
MEASURING ENRICHMENT OF LOCAL TRANSPOSES

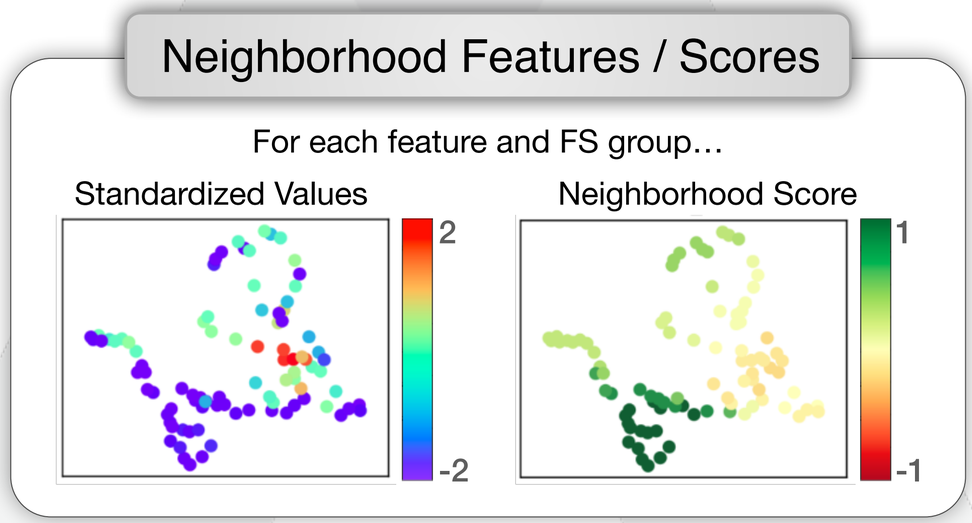
MINEDXAI

- Analyze each local transpose embedding w.r.t. the information it contains about each metadata feature (or composite metadata signature)
- Scores characterize the information content of the neighborhood around each point in the LT embedding space.
- We define the following ***enrichment score*** function for feature distributions P and Q, based on the Wasserstein-1 distance

$$S(P,Q) = \tanh\left(\beta \cdot \text{sign}(\mu_P - \mu_Q)\frac{W_1(P,Q)}{\sigma_Q}\right)$$

bounded between -1 (P deficient relative to Q) and 1 (P enriched relative to Q), where beta represents a contrast parameter. The reference Q is the population distribution.





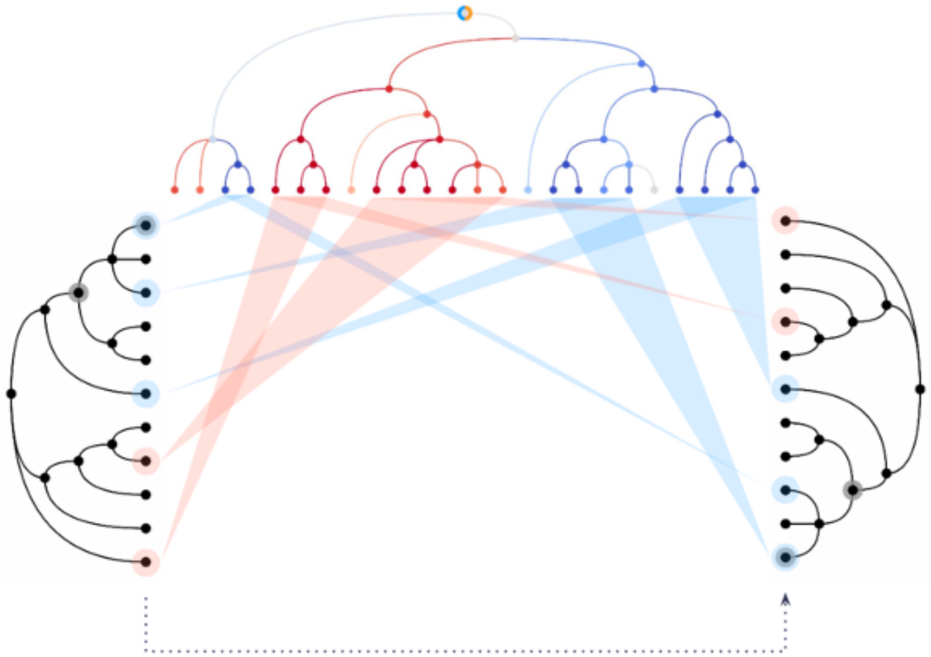
QUERYING

- We formulate reasoning and inference in our model as **querying**.
- Enrichment Scores provide a basis for constructing and answering queries across modalities, modulated by the underlying miRNA decompositions which act as an intermediary.
- Boolean logic enables composing a broad class of queries about the data as part of a query vector **q**., which we combine with the query kernel

$$\boldsymbol{S}_s^\top \boldsymbol{S}_t$$

- The query vector, in combination with enrichment scores, allows easily exposing relationships across modalities.

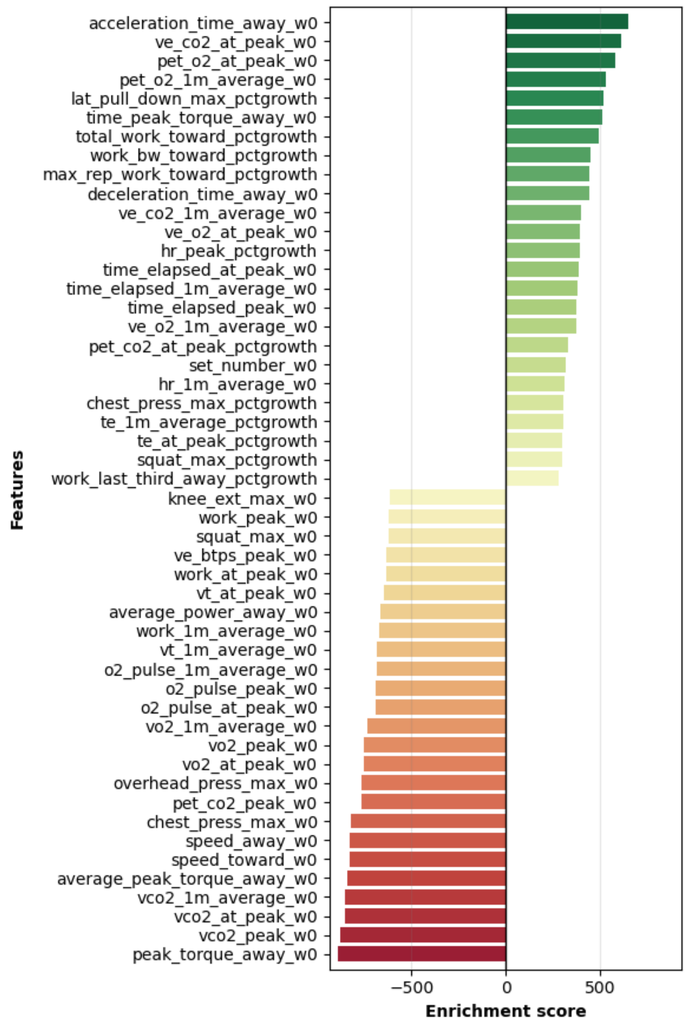
Querying Conceptual Diagram



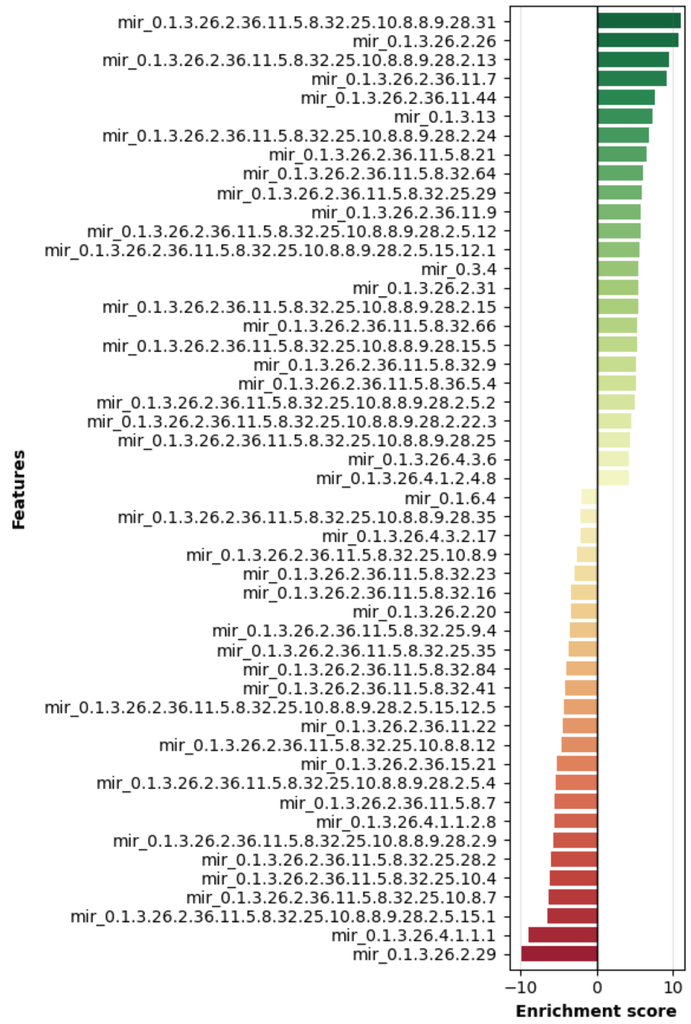
MULTIMODAL QUERYING THROUGH MIRNA

Example Query:
White
Female
HIIT

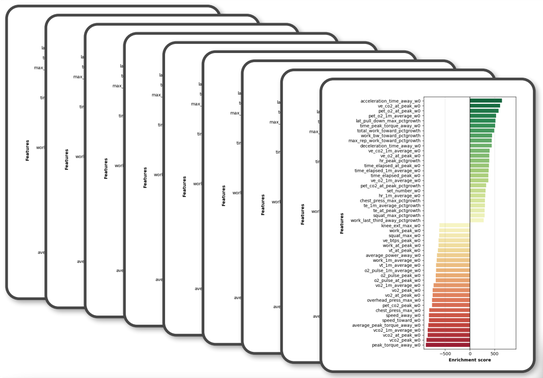
Querying into Metadata



Querying into miRNA Groupings

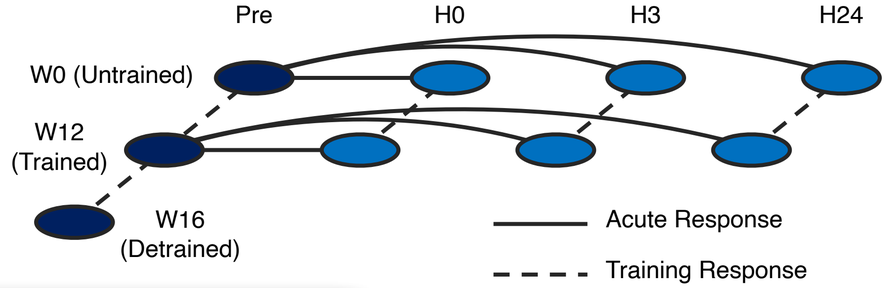


MULTIMODAL QUERYING THROUGH MIRNA

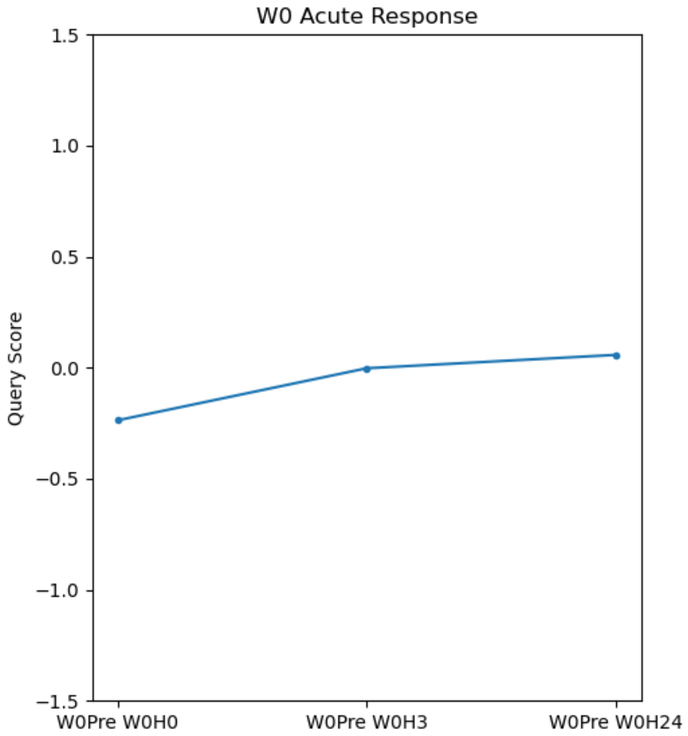


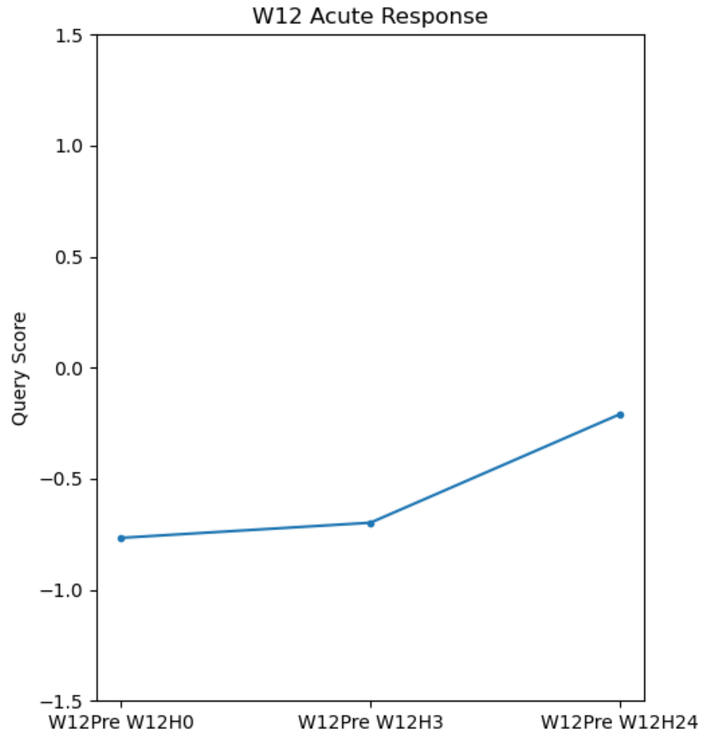
...

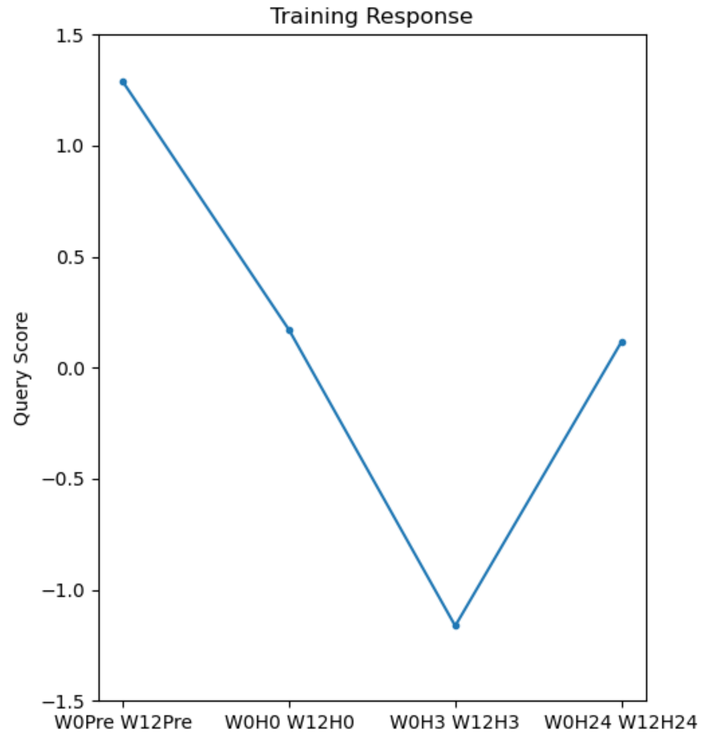
Collection of Metadata
Query Results
Across Time Pairs



Temporal Response w.r.t. Query Targets (e.g. BMI Growth)



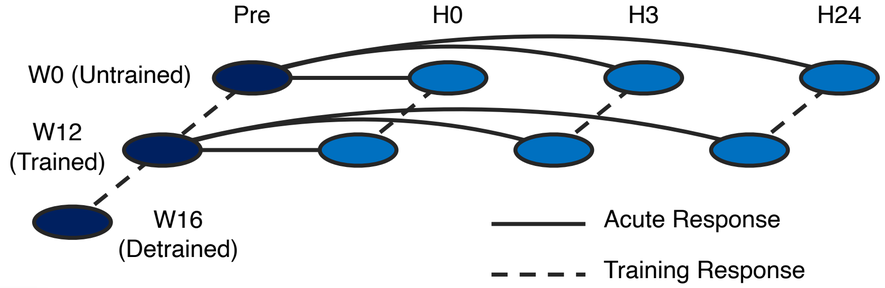




QUERYING MIR FROM FEATURES

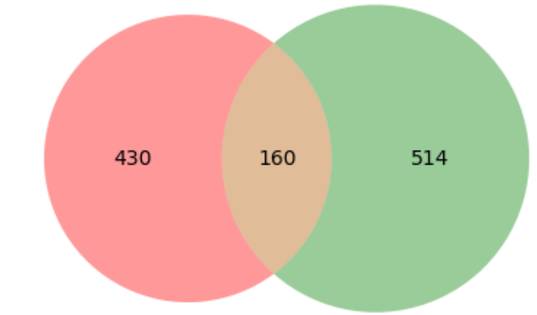
...

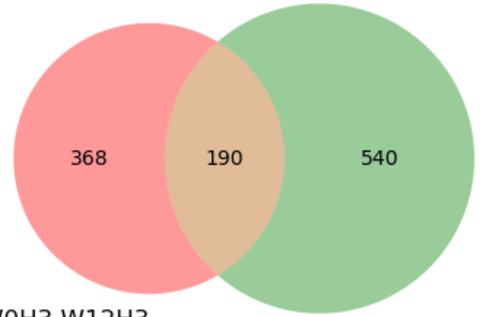
Collection of miRNA Query Results Across Time Pairs



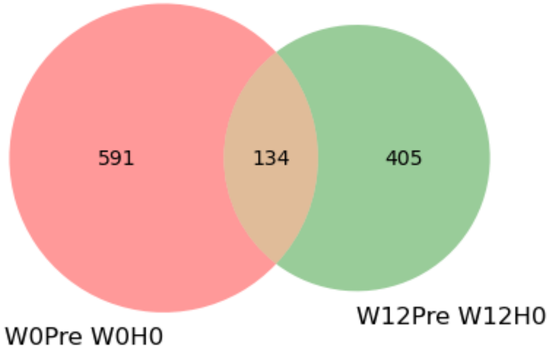
Top 20 FS Groups

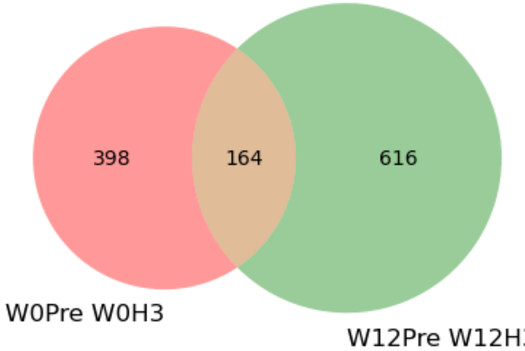
Acute Response Comparison

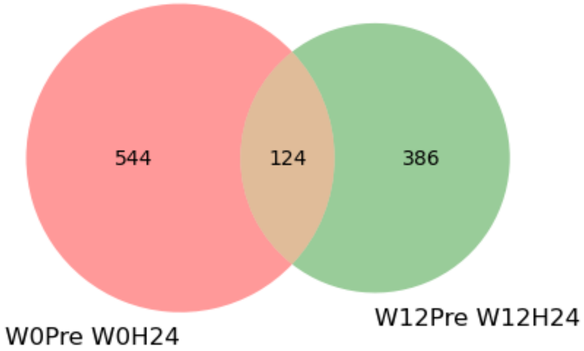




Training Response Comparison



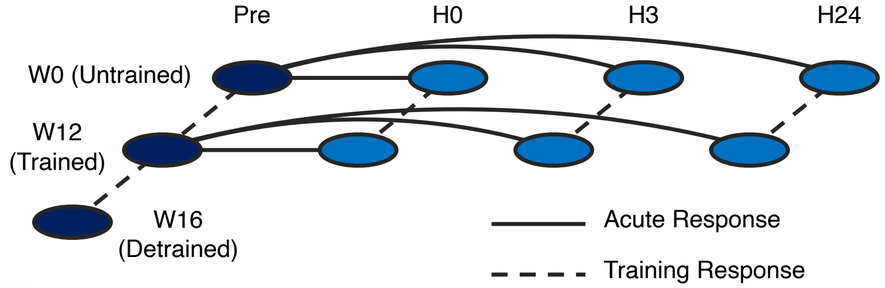




QUERYING MIR FROM FEATURES

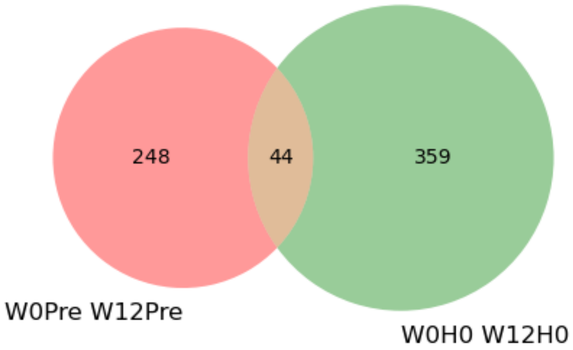
...

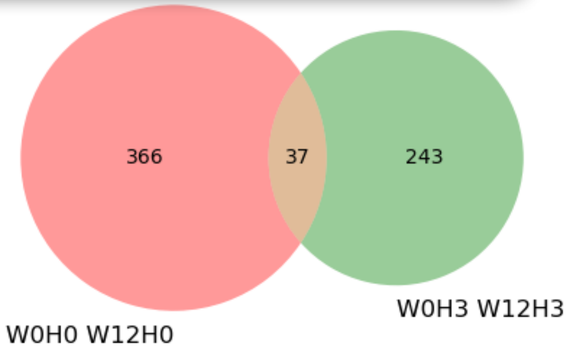
Collection of miRNA Query Results Across Time Pairs

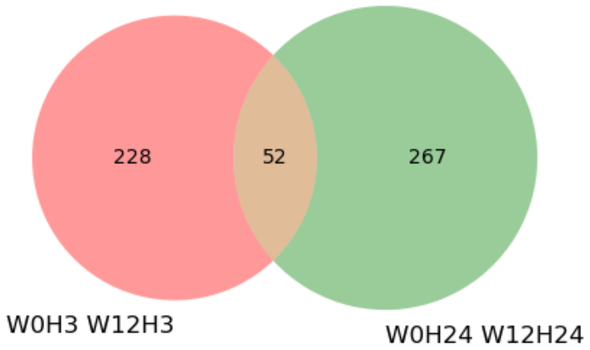


Top 10 FS Groups

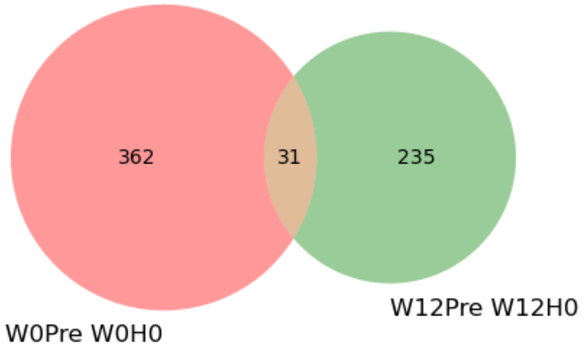
Acute Response Comparison

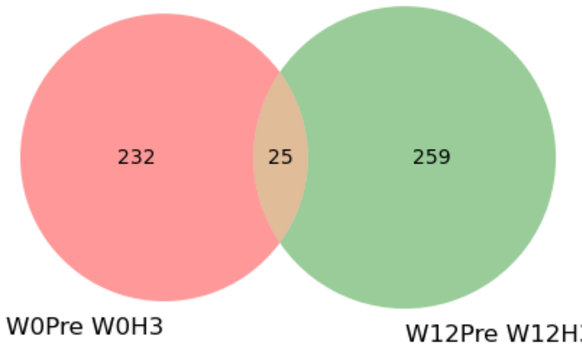


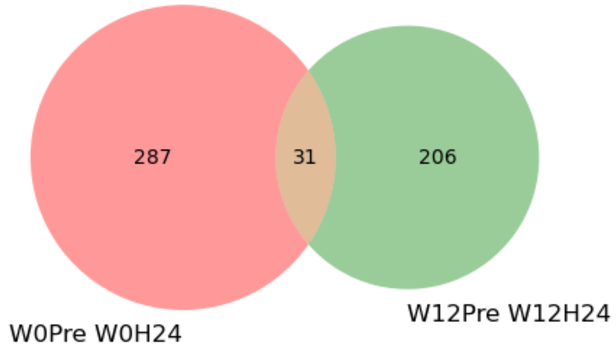




Training Response Comparison







QUERYING MIR FROM FEATURES

MINEDXAI

Common miRNA between Time Pairs

miRNA Identification Source		Count	Number of Common miRNA
W0Pre W12Pre	W0H0 W12H0	44	['7856-5p', '4252', '4666b', '548i', '1236-3p', '3085-5p', '3924', '6769a-3p', '5187-5p', '5698', '6742-3p', '651-5p', '**935**', '490-3p', '1909-3p', '6529-3p', '5589-3p', '597-5p', '6862-3p', '**195-3p**', '2113', '6849-5p', '6505-5p', '5047', '3151-5p', '8055', '4745-3p', '6772-5p', '4663', '3126-5p', '3682-3p', '**2278**', '3199', '1231', '6728-5p', '**517c-3p**', '3912-3p', '3132', '876-5p', '4327', '4483', '887-3p', '939-5p', '548bb-5p']
W0H0 W12H0	W0H3 W12H3	37	['571', '4665-5p', '4735-5p', '4788', '4494', '3124-5p', '5693', '1204', '3117-5p', '4789-3p', '6742-3p', '1302', '6726-5p', '507', '6857-3p', '4540', '**1294**', '4647', '**195-3p**', '**3689d**', '6885-5p', '1910-5p', '3945', '6799-5p', '4707-5p', '4802-5p', '**202-5p**', '371a-3p', '4778-5p', '299-3p', '4765', '6859-3p', '5587-3p', '3912-3p', '***627***-3p', '12120', '6760-3p']
W0H3 W12H3	W0H24 W12H24	52	['3143', '**589-3p**', '3200-5p', '6511a-5p', '4649-3p', '**1301**-5p', '1226-3p', '4507', '331-5p', '6726-3p', '1915-5p', '**1294**', '1248', '**636**', '6771-3p', '3074-5p', '6785-5p', '**221-5p**', '576-3p', '6079', '**548g**-5p', '6792-5p', '1200', '548o-5p', '3974', '148b-3p', '500a-3p', '4251', '4802-5p', '16-1-3p', '**548c**-5p', '**196b-5p**', '**183-3p**', '**7g**-3p', '548x-5p', '**15a-3p**', '451b', '1207-5p', '**181d-5p**', '548aj-5p', '4778-5p', '10396b-5p', '941', '**301a**-5p', '5010-3p', '1537-3p', '**20a-3p**', '4641', '7153-3p', '12120', '**106a-5p**', '548am-5p']
W0Pre W0H0	W12Pre W12H0	31	['4313', '**181c-3p**', '**181a-2-3p**', '365a-5p', '661', '6758-5p', '**4640-3p**', '6869-5p', '513c-3p', '4305', '3675-3p', '4539', '6508-3p', '**330-5p**', '3664-3p', '4297', '1307-5p', '3938', '4711-5p', '11181-5p', '449b-3p', '218-2-3p', '**455-3p**', '**503-5p**', '**7f-5p**', '6871-3p', '4740-5p', '6515-3p', '4317', '557', '4784']
W0Pre W0H3	W12Pre W12H3	25	['6847-5p', '7107-3p', '3168', '4730', '**4640-5p**', '**30d-3p**', '4449', '6808-3p', '10401-3p', '7973', '**542-5p**', '**29a-5p**', '3121-5p', '3157-5p', '6749-3p', '760', '4676-5p', '4752', '6134', '516a-5p', '6515-3p', '6503-5p', '548at-3p', '4787-5p', '**27a-5p**']
W0Pre W0H24	W12Pre W12H24	31	['320b', '17-3p', '7850-5p', '4703-5p', '6131', '30e-5p', '2114-3p', '425-5p', '**142-3p**', '641', '30a-5p', '3184-3p', '875-3p', '4452', '4680-5p', '320a-3p', '4491', '509-3-5p', '3689e', '595', '186-5p', '101-3p', '4664-5p', '423-5p', '548ar-5p', '25-3p', '4320', '891b', '15b-3p', '6868-3p', '106a-5p']

— Known connection to exercise
— Different end (and possibly function) of a miRNA with known connection
— Impact pathways that have been identified and connect to exercise

SUMMARY AND NEXT STEPS

MINEDXAI

- SUMMARY
  - Successfully built out full encoder architecture for miRNA and metadata queries
  - Query system provides the outputs of the encoder layer and inputs for the decoder layer (via transformer-like approaches that utilize attention mechanisms)
  - Can successfully query across any miRNA timepoints combinations, metafeature individually, metafeature distributions (from THDs)
- NEXT STEPS
  - Create visualization interfaces to allow non-expert users to fully interrogate experimental parameters
  - Utilize architecture outputs (local view) to integrate with large language models (world view) to provide natural language inputs and outputs across architecture
  - Evaluate generalization of modeling approach on other data types (MSKI)

Thank you to Dr. Pat Bradshaw for funding seedling effort

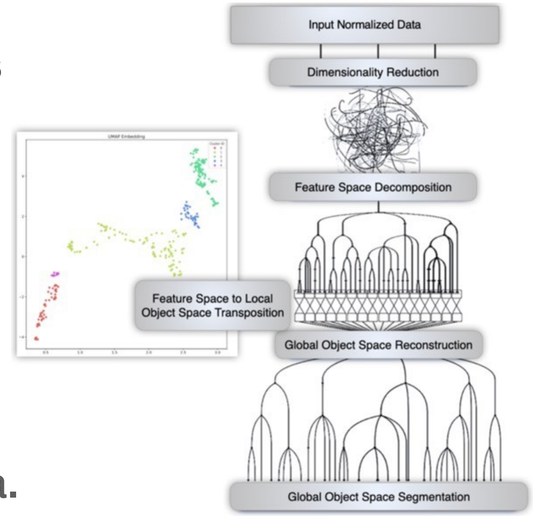
Topological Data Analysis (TDA) for Identification of miRNAs as Biomarkers for Human Performance (Mined XAI)

OBJECTIVES:

- Develop deep topological models for robust multi-variate understanding of human performance data.
- Identify miRNAs and other metadata features associated with performance outcomes.

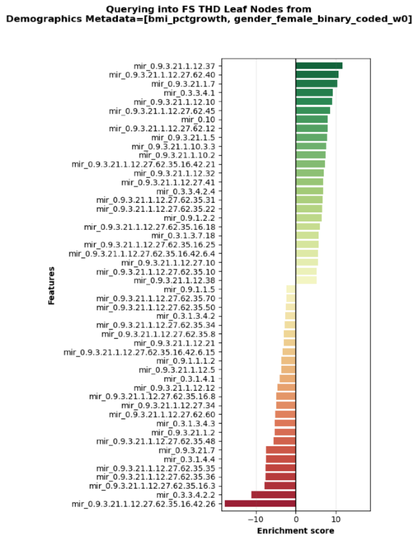
TECHNICAL APPROACH:

- Develop deep topological models of miRNA up/down-regulation across different temporal periods and of various metadata sources.
- Development of a multi-modal data fusion pipeline allowing querying relationship importance across both miRNA and metadata.



ACCOMPLISHMENT:

- Identify complex signatures of miRNA and other functional, demographic, and biological subject subsets.
- Links between metadata and metadata features are established based on statistics of the spatially-varying distribution of the metadata features, e.g. subject and miRNA subsets enriched in certain features.



DoD BENEFIT

- Development of methodology to integrate multi-modal data for human performance.
- Identification of biomarkers for adaption to physical and operational performance.
- Moving towards hyper-personalization of warfighter performance.

MINEDXAI

COLLABORATIONS/TRANSITIONS/PRESENTATIONS

- **AFRL/711 HPW**
  - Developing explainable AI model for micro-physiological systems (sub- contract awarded)
  - Integrating multimodal data for Integrated Cockpit Sensing (awaiting sub-contract award)
- **National Science Foundation (collaboration with Prof Singamaneni/Prof Raman)**
  - Convergence Accelerator track L (Phase 2) using XAI to identify chemical odors
- **Poster Presentations**
  - Poster presentation at AFRL Biotech Days (Feb 2024)

QUESTIONS?

CONTACT US

Rajesh Naik
Chief Operating Officer

rajesh@minedxai.com
O +1 937 550 6518
M +1 937 716 7941

Ryan Kramer
President, Founder

ryan@minedxai.com
O +1 937 550 6518
M +1 937 608 1835

Christopher Dean
Chief Technology Officer

christopher@minedxai.com
O +1 937 550 6518
M +1 614 804 5282

MINEDXAI