Breakout Session 4: Track B

Optimizing Diagnostic and Prognostic Biomarkers of CASH using Machine Learning

Dr. Diana Vera Cruz
Bioinformatician, University of Chicago

Dr. Romuald Girard
Assistant Professor, University of Chicago

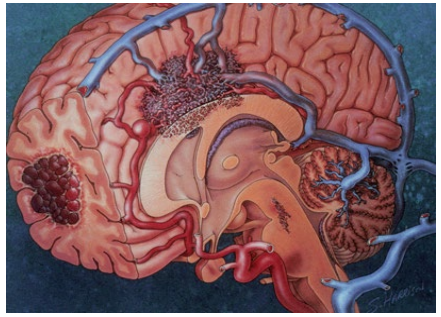
Optimizing Diagnostic and Prognostic Biomarkers of CASH using Machine Learning

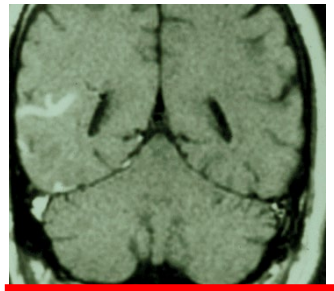
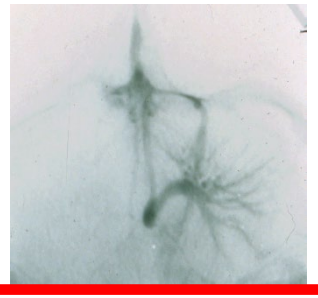
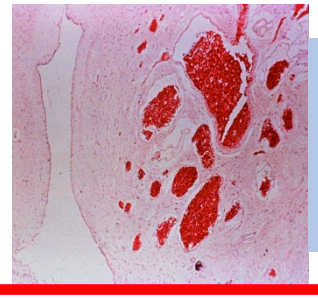
Optimizing and Sharing Data for Machine Learning [ML] Analyses of Multiomic Biomarkers of Cavernous Angiomas with Symptomatic Hemorrhage [CASH]

National Institute of Neurological Disorders and Stroke

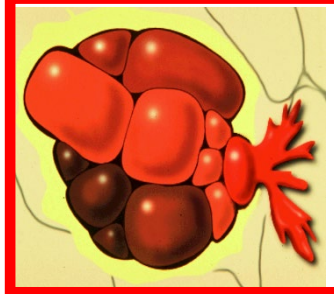
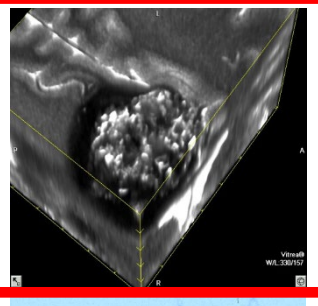
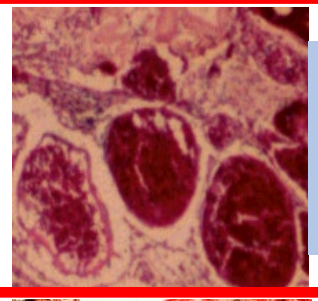
THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

Cavernous Angiomas (CAs) are fairly common cerebrovascular anomalies

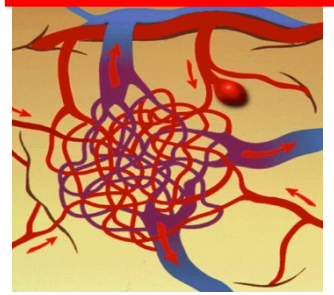
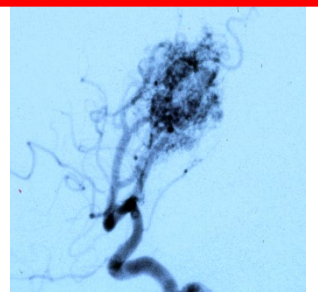


VENOUS ANGIOMA
Venous developmental anomaly
Regional venous dysmorphism

CAVERNOUS MALFORMATION
Hemorrhagic proliferative dysangiogenesis


ARTERIOVENOUS MALFORMATION
Arteriovenous shunting

Cerebral CAs behavior is unpredictable

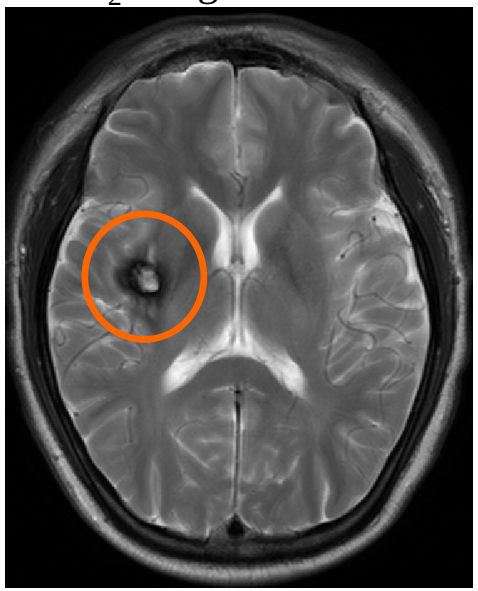
- CAs are abnormal clusters of enlarged capillary vessels embedded in normal brain or spinal cord tissue

- 2 forms : sporadic/solitary or familial/multifocal

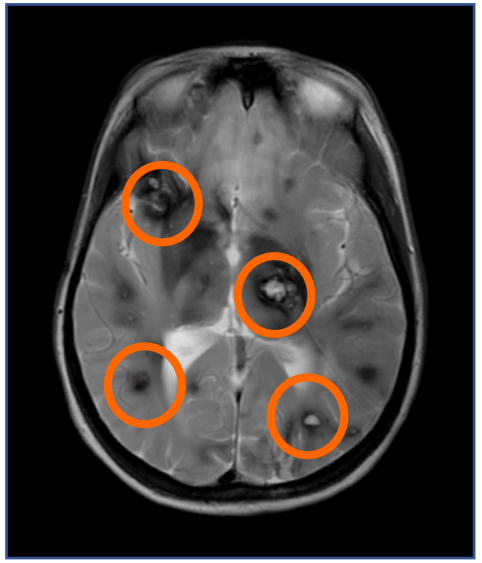
- CA without prior symptomatic hemorrhage (SH)
  - ➢ Low initial risk of SH (0.4 to 2.4% per year)

- CA with recent SH
  - ➢ High risk of rebleeding after initial SH (*Al-Shahi et al., 2012*)
    - 10-fold increase
    - 3.8 to 29.5% per year

T_2-weighted MRI

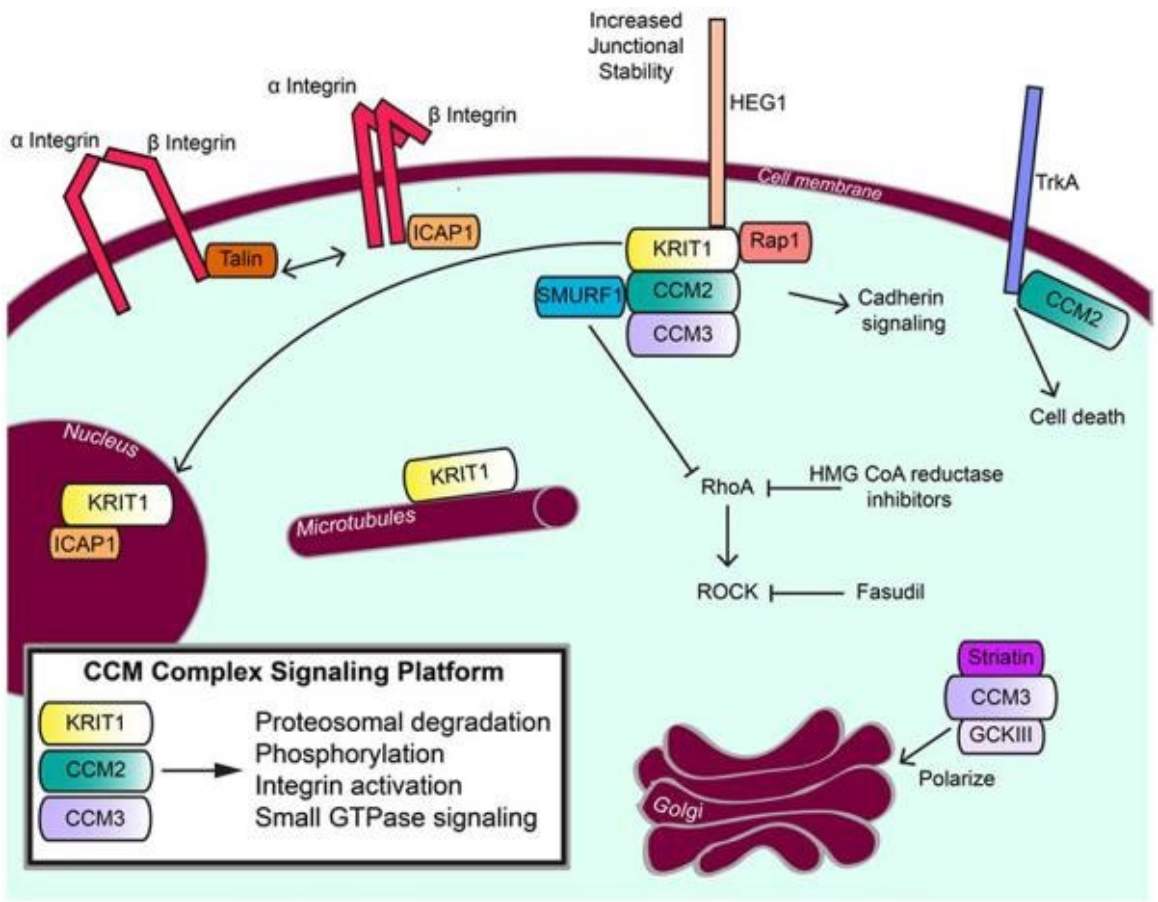


in sporadic patient



in familial patient

A complex interplay of angiogenesis and inflammatory processes



Fisher & Boggon, 2014

Double hit mutations on one of the CCM genes (stochastic and/or inherited)

↓

Signaling aberration

↓

Disruption of endothelial cell junctions

↓

Vascular hyper-permeability

↓

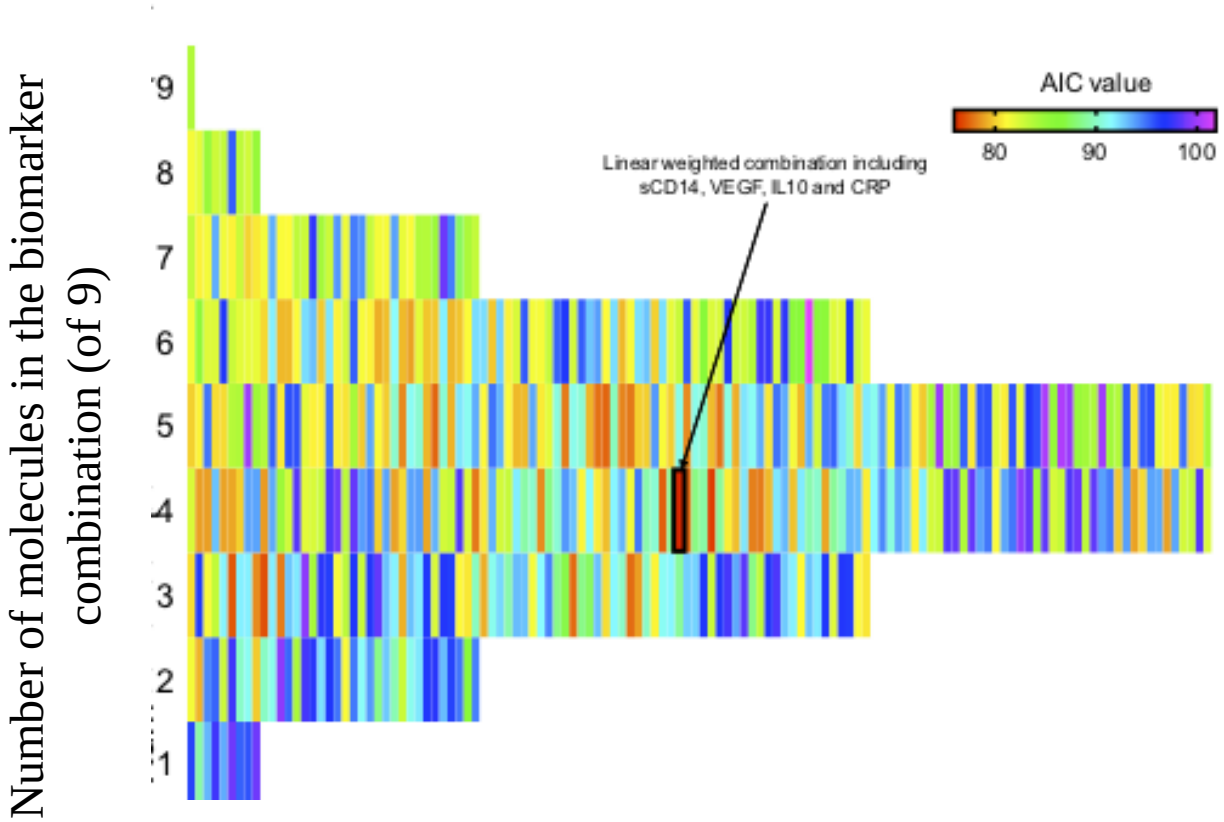
Hemorrhage and iron deposition

4 categories of biomarkers defined by the FDA-NIH Biomarker Working Group

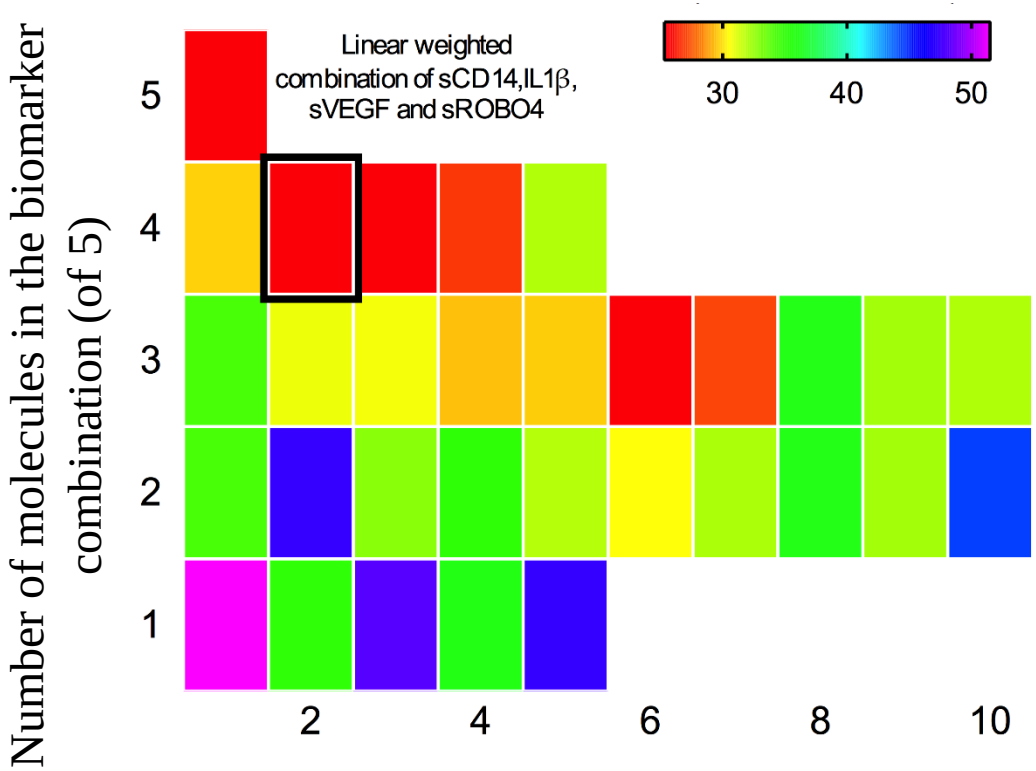
- A relevant biomarker may reflect chronic disease over the patient's lifetime, recent acute clinical activity or predict future events (Amur et al., 2015).

- 4 categories of biomarkers:
  - ✓ **Diagnostic** distinguish patients with a particular disease.
  - ✓ **Prognostic** provide information on the likely course of disease in an untreated individual.
  - ✓ **Predictive** provide a forecast of the potential responses (favorable or unfavorable) to one or more specific treatments.
  - ✓ **Response** are dynamic assessments of a biological response after a therapeutic intervention, include:
    - ➢ **Safety** indicating biological adverse effects in response to treatment.
    - ➢ **Pharmacodynamic** indicating the intended activity of the drug.
    - ➢ **Efficacy-response or surrogate endpoints** predicting a specific disease-related clinical outcome.

Plasma molecules effectively combine into a diagnostic and prognostic biomarker of hemorrhagic activity of CCM



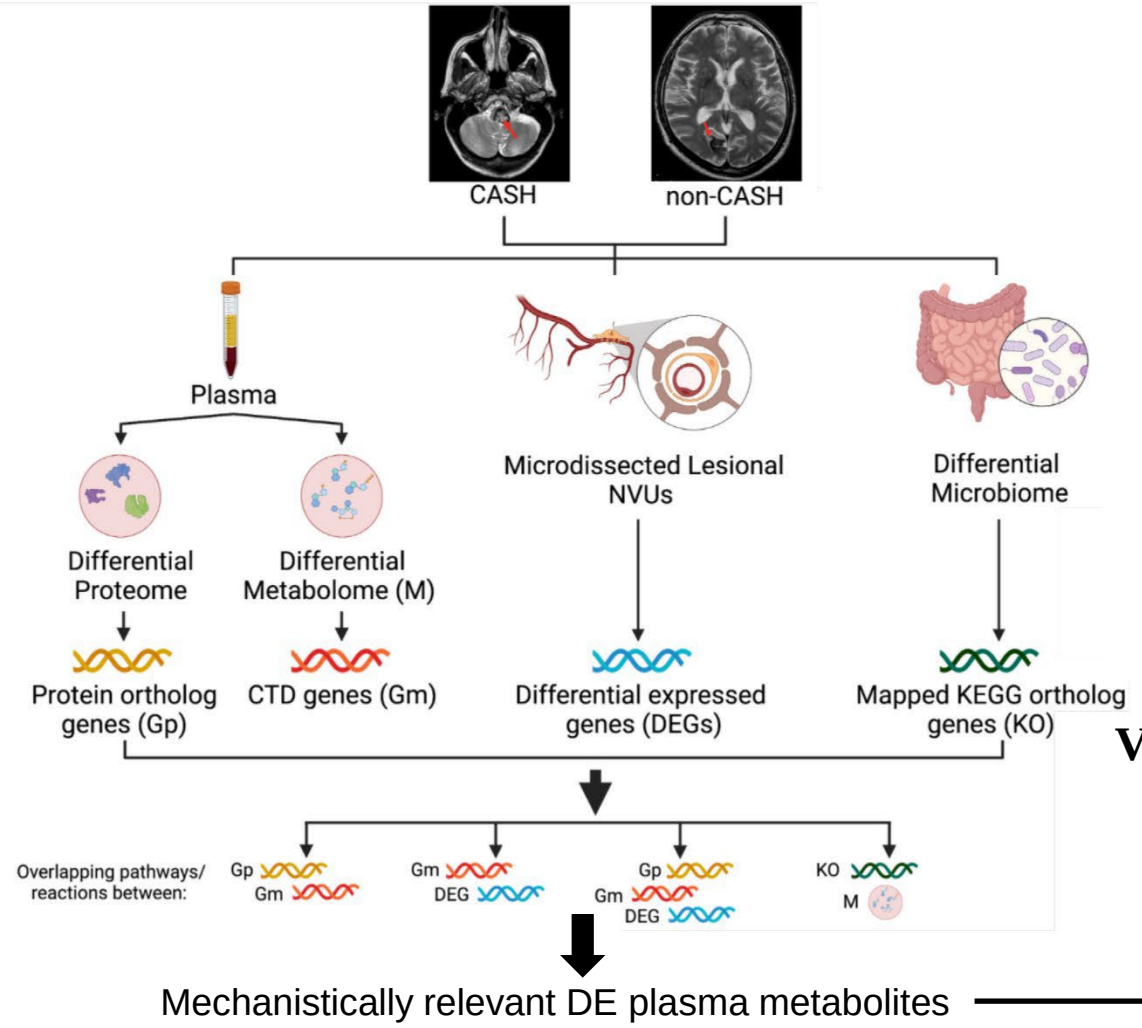
Lyne et al., 2019



Girard et al., 2018

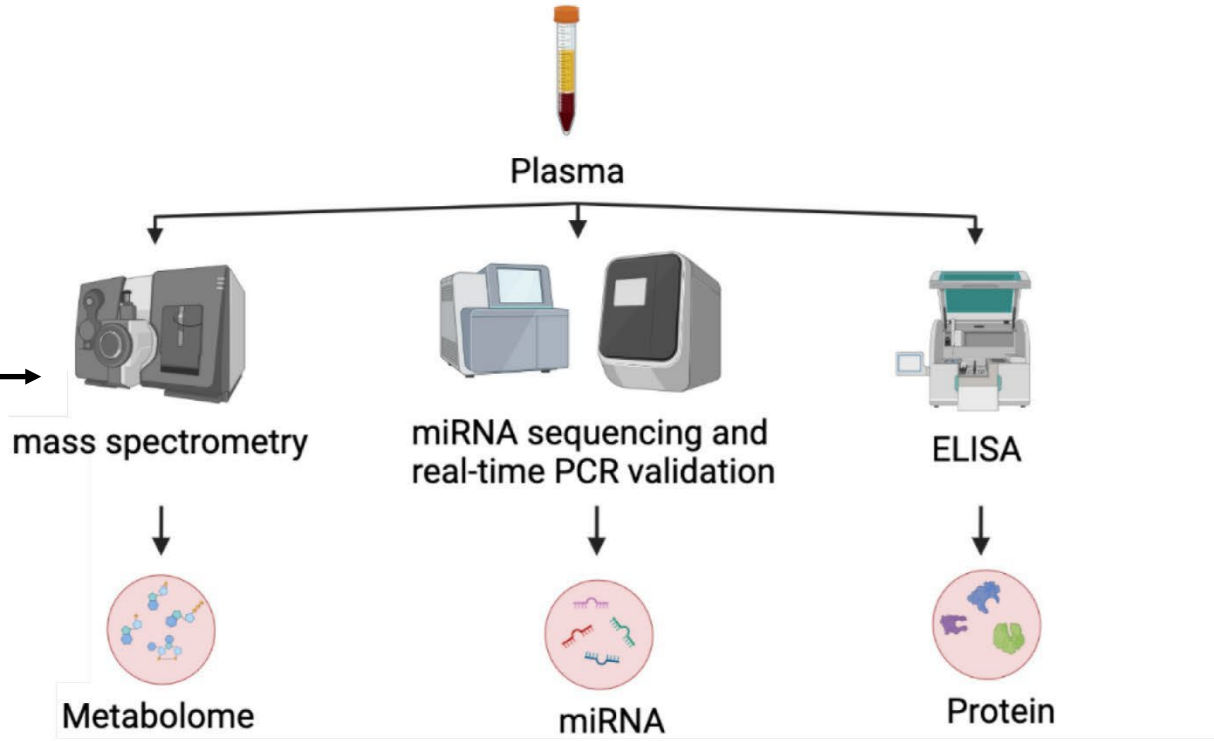
Methodology to Identify Candidate Biomarkers

Discovery cohort



Validation Steps

Independent validation cohort of CASH vs non-CASH patients (n=20/20), propensity matched for age, sex, phenotype, and brainstem lesion location

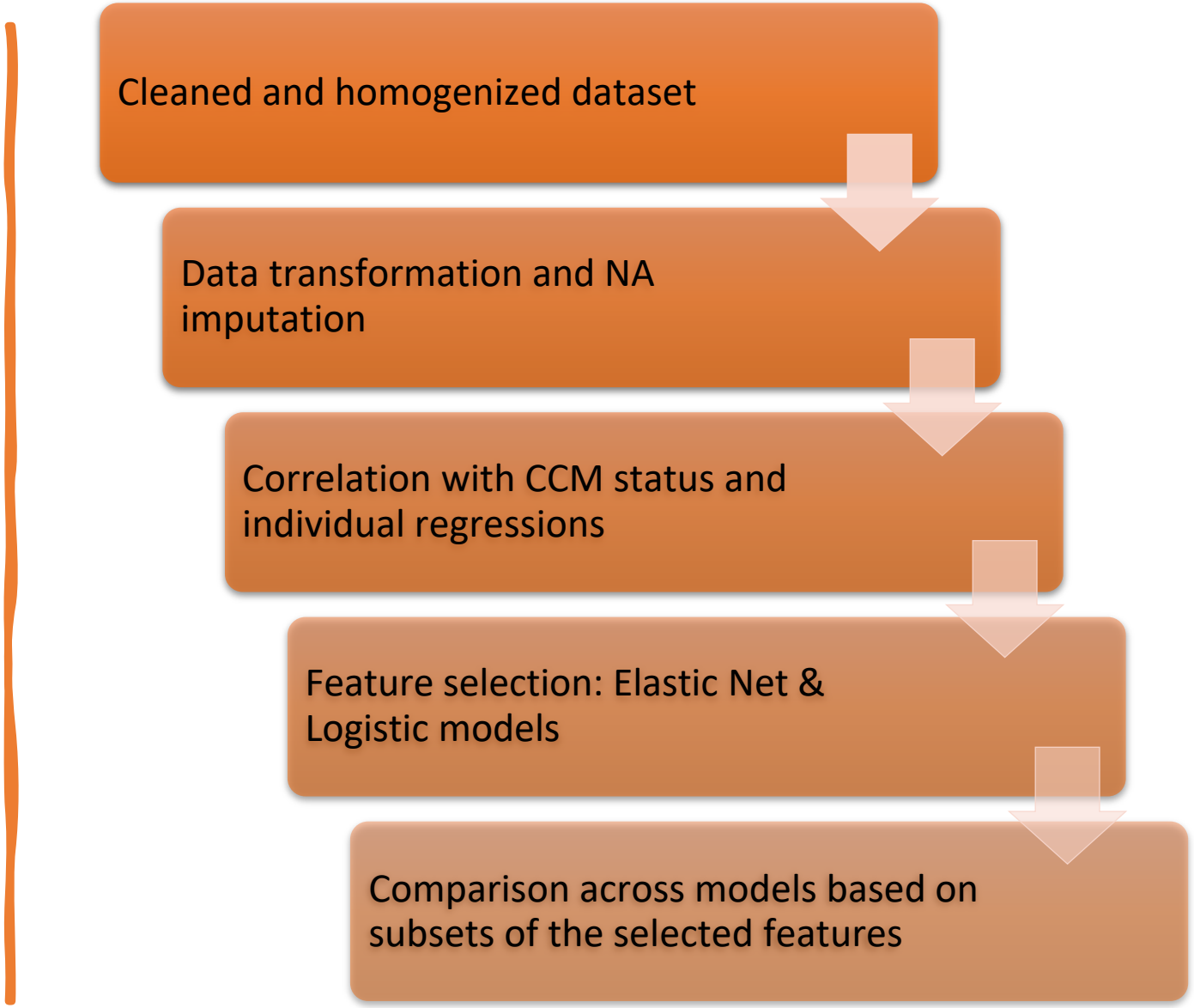


Srinath et al., 2023

Multi-Omic Datasets in Diagnostic and Prognostic Discovery Cohorts

	Assay	Diagnostic	Prognostic
Metabolites	LC-MS/MS	11	11
Proteins	ELISA	16	16
miRNA	ddPCR	5	-
Patients		**20/20**	**15/15**

Pilot cohorts: General Workflow



Pilot cohorts: General Workflow

Cleaned and homogenized dataset

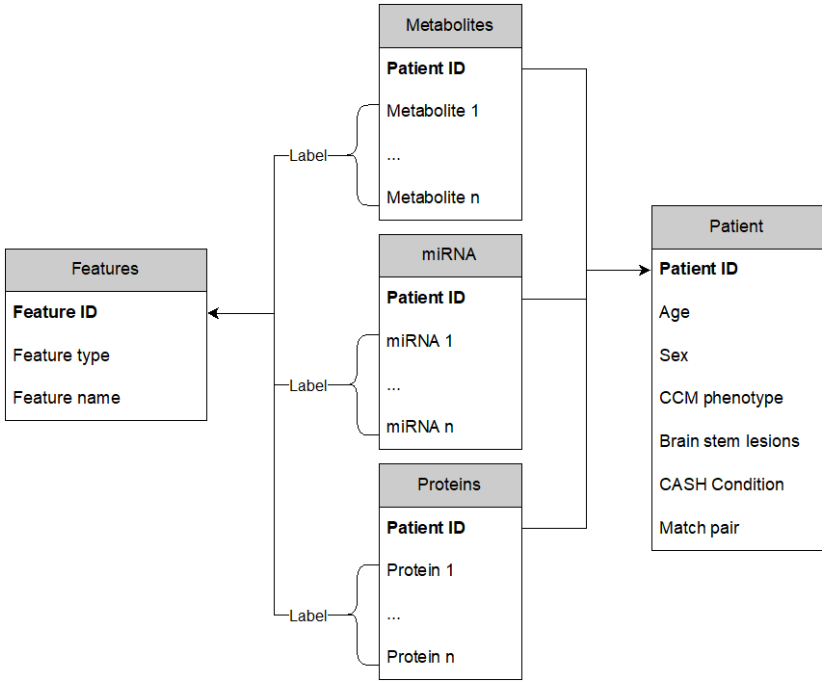
Data transformation and NA imputation

Correlation with CCM status and individual regressions

Feature selection: Elastic Net & Logistic models

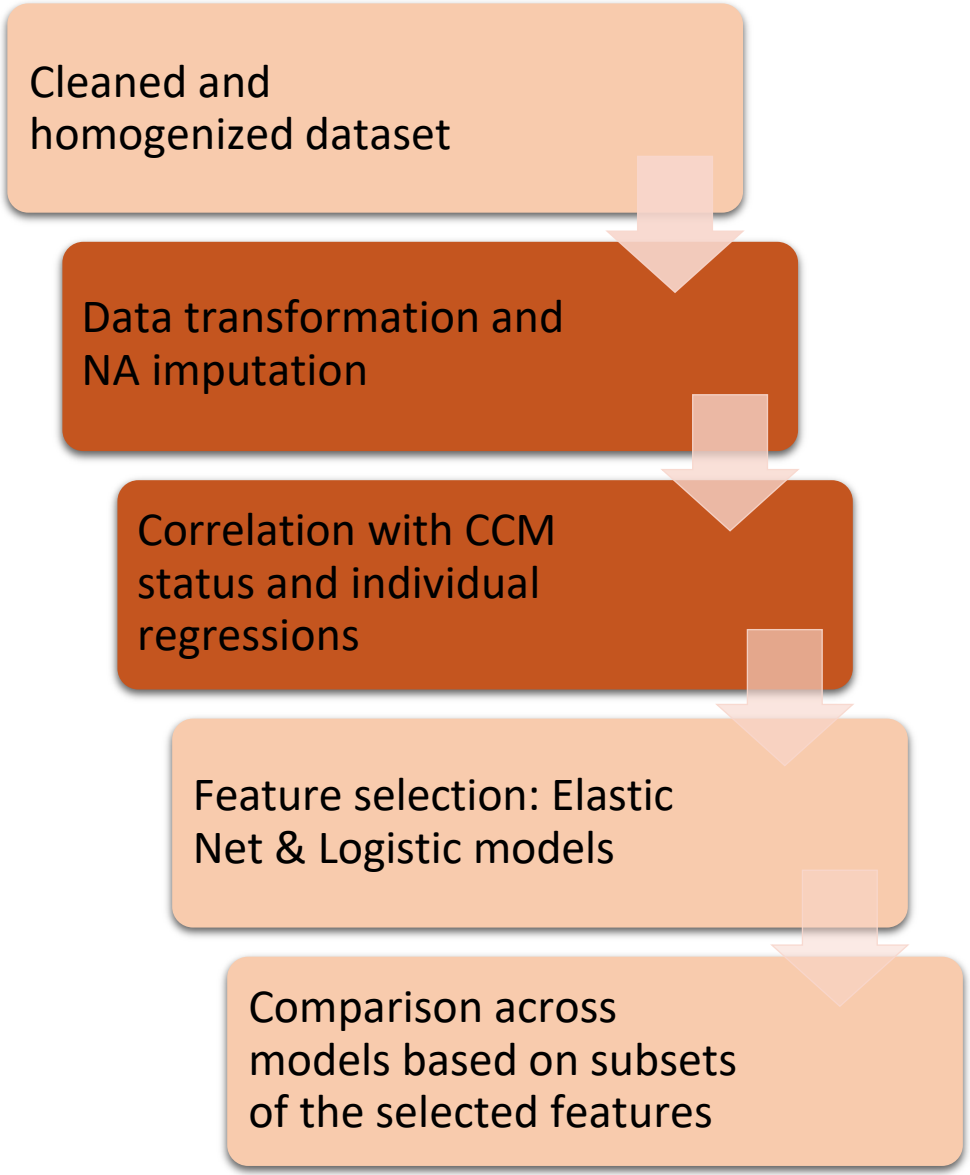
Comparison across models based on subsets of the selected features

Database structure



- Homogeneity across tables: Universal patient ID.
- Tidy format for data analysis and repository sharing.

Pilot cohorts: General Workflow



Data transformation

Test of normality (Shapiro-Wilk test)

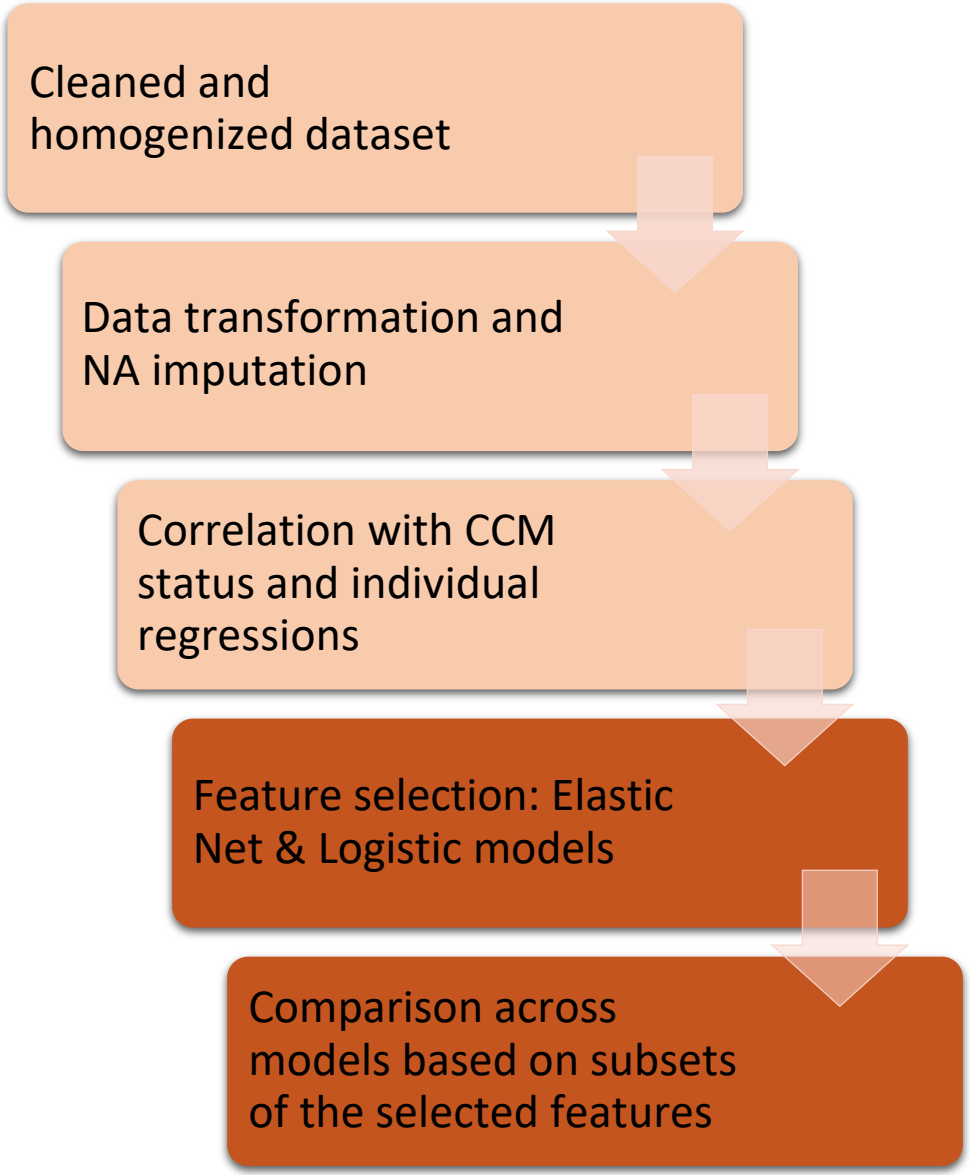
- Metabolites: linear
- Proteins: $\log_2$
- miRNA $\log_2$

NA imputation

- Model-based imputation method
- Hot-Deck initialized

Individual logistic regression

Pilot cohorts: General Workflow



Feature selection

- **Elastic Net** optimized for **accuracy** and **repeated k-fold** cross-validation.
- Logistic regression over the complete set and conditional logistic regression to evaluate the performance of propensity-match.

Reduced models

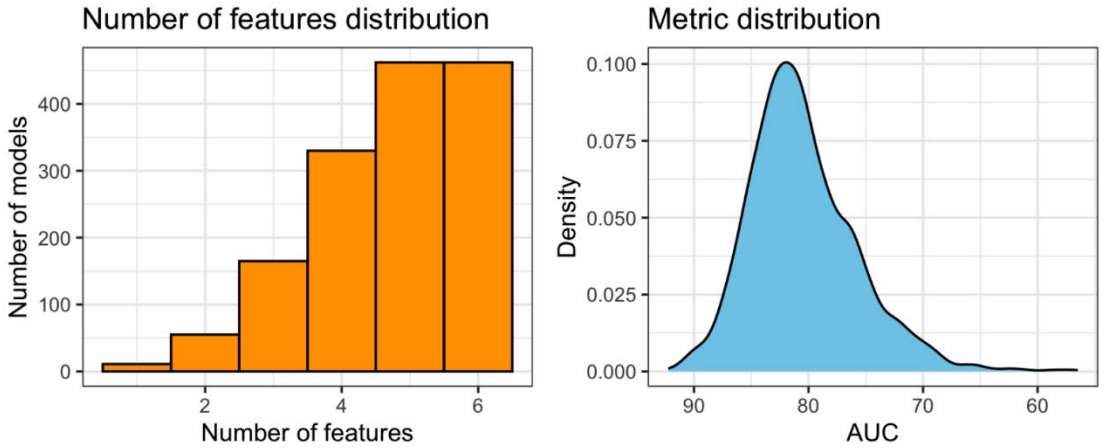
- Subset combinations of n elements, arranged by highest AUC.

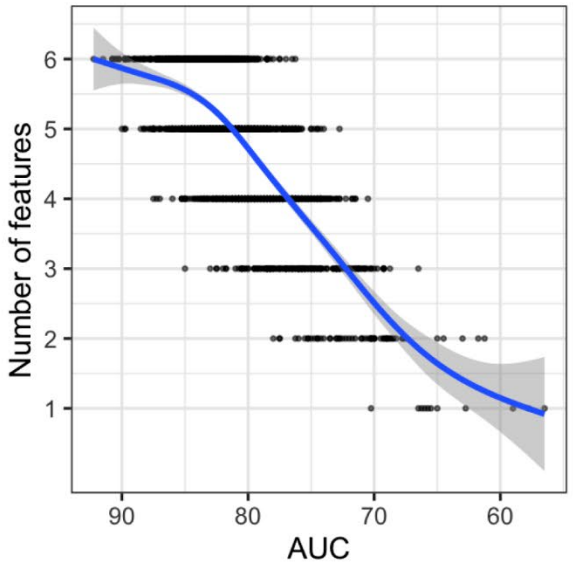
Best models criteria

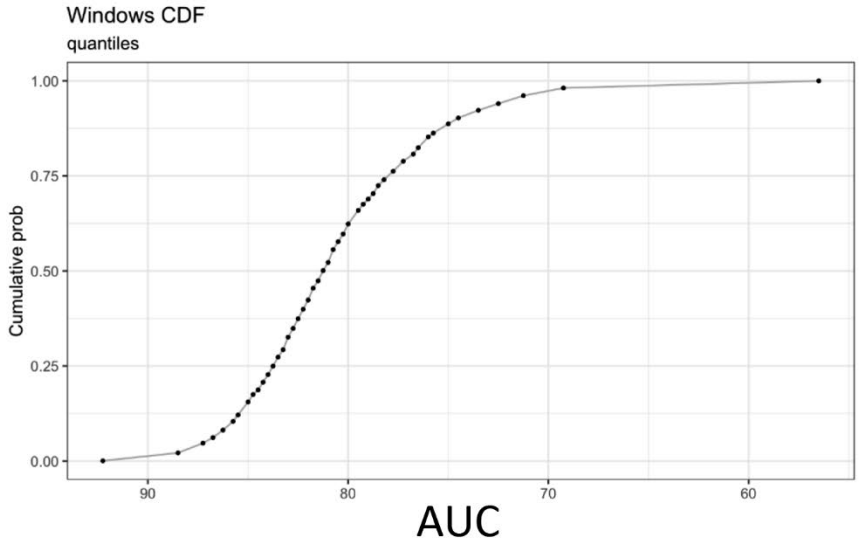
- Highest AUCs for a given number of features and lowest number of unique molecule types.

Subset models - AUC Comparison

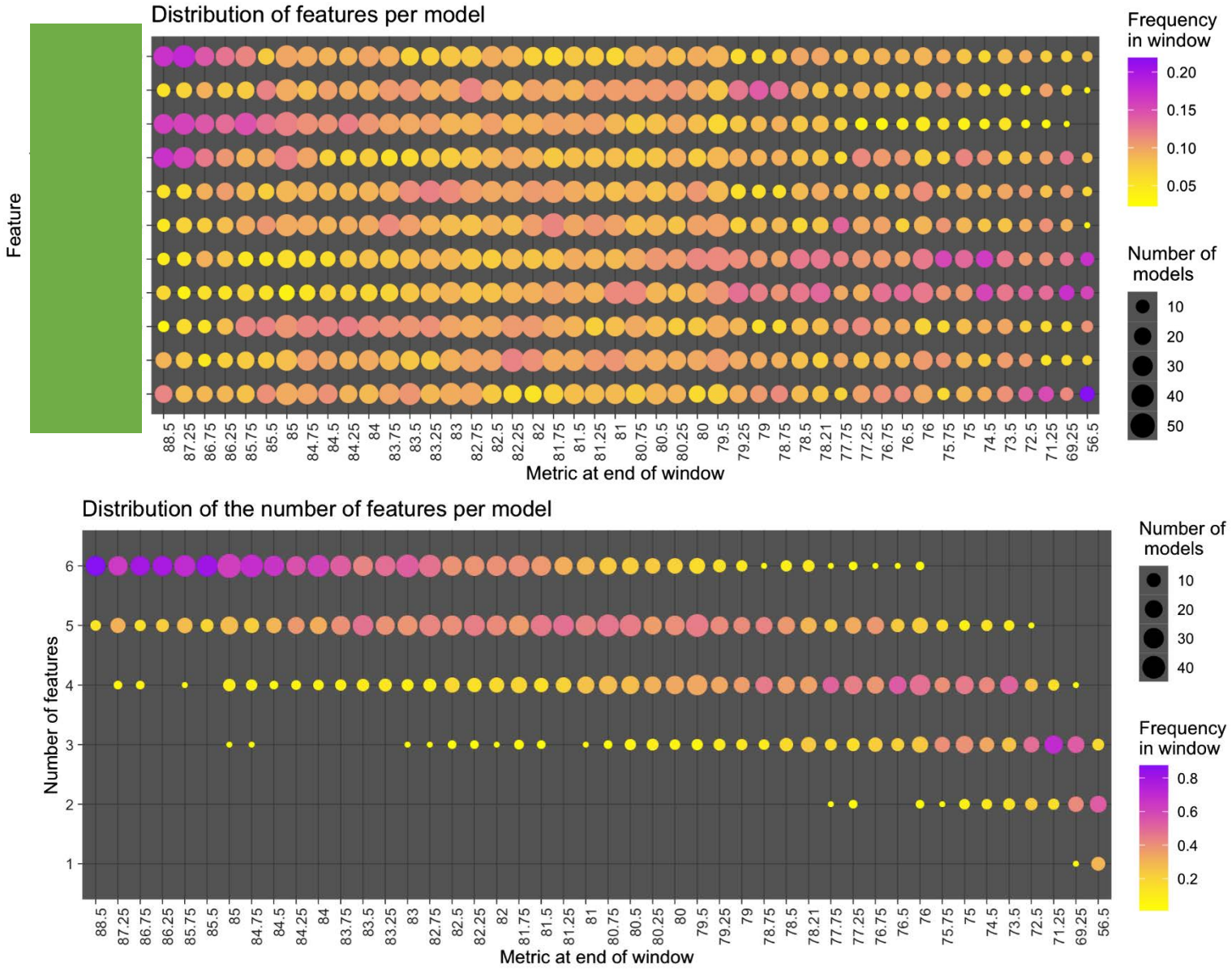
Total model considered: 9948 (Combinations from 1 to 6 elements max)







Subset models - AUC Comparison



Future work and Perspective

- Identification of the best diagnostic and prognostic models.
- Best models evaluation in testing cohorts: n > 260 patients.
- Model performance in relation of the known confounders of CCM clinical activity (e.g., CCM phenotype, lesion localization, gender and age).