MILK COAGULATION · MIR SPECTROSCOPY

Predicting milk coagulation with prior-fitted transformers

Honest benchmarking of PLS, AutoGluon and TabPFN for milk-coagulation prediction in Holstein cattle.

Tin Josip Curik, Dario Domović, Nataša Mikulec, Boris Lukić, Marija Špehar, Ino Curik

Maastricht University · University of Zagreb · J. J. Strossmayer University of Osijek · Croatian Agency for Agriculture and Food · Hungarian University of Agriculture and Life Sciences (MATE)

Efficient selection for cheese depends on low-cost phenotyping

Milk coagulation properties are measured with the Formagraph — up to 60 minutes per 10 samples.

Far too slow to phenotype a breeding population.

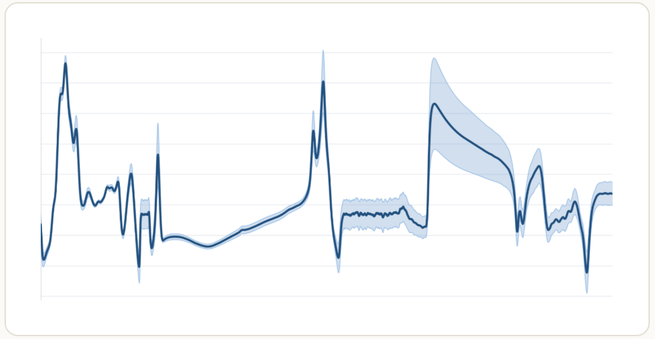
OBJECTIVE

Develop robust and accurate ML models to predict milk coagulation properties from MIR spectra *(routinely collected)*.

THE PROBLEM

Reduced to a regression: 1,000 numbers in, six traits out

MIR spectrum



$x = [\ a_1,\ a_2,\ \dots\ ,\ a_{1000}\ ]$

1,000 wavenumbers

→

supervised regression

→

6 coagulation traits

RCT	min
k20	min
a30	mm
a60	mm
amax	mm
a2r	mm

DATA

1,608 Holstein cows · 3 farms · 12 batches

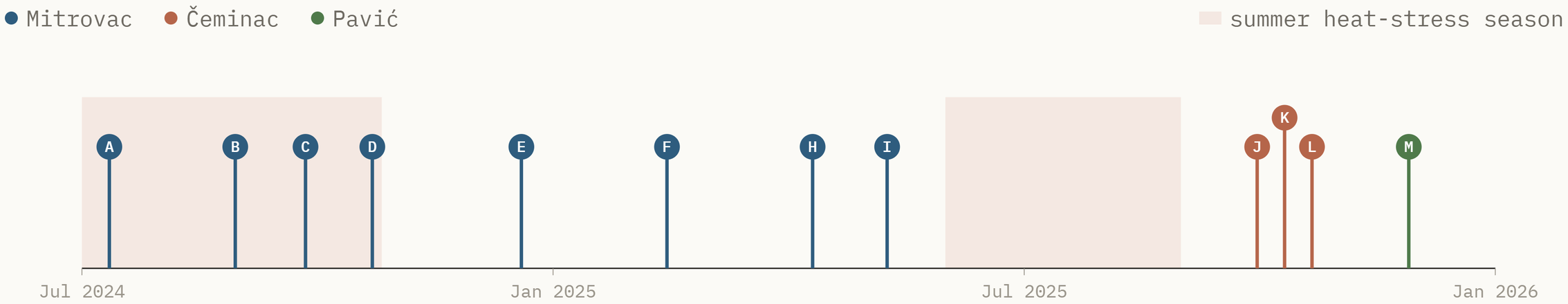
1,608

Holstein cows · after QC
all milk from morning milking

Mitrovac	8 batches	1,212
Čeminac	3 batches	284
Pavić	1 batch	112

DATA

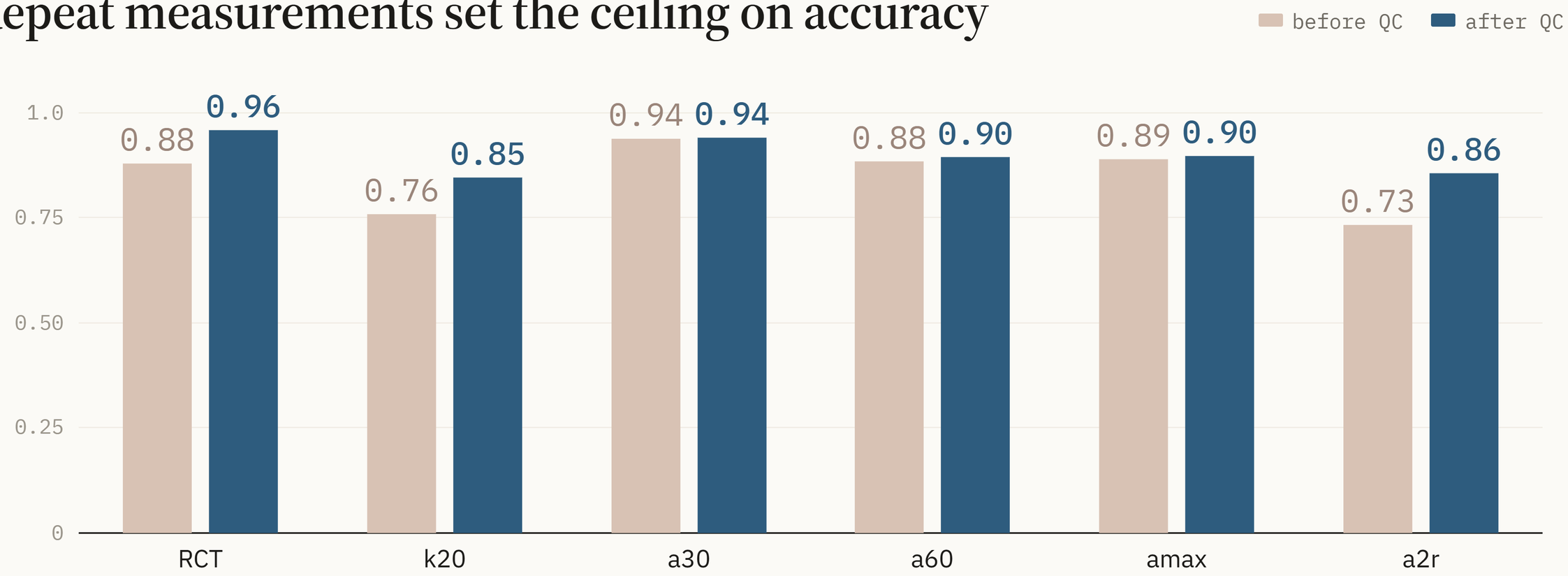
Every batch is a different moment in time



A batch encodes date · season · spectrometer state — the seed of the confounding

METHOD · MEASUREMENT CEILING

Repeat measurements set the ceiling on accuracy



> R^2_{max} = the highest R^2 any model could reach. Quality control excludes the **5.7% of samples** (n = 111) whose duplicate runs disagree most — multivariate SED > 0.8 — which lifts the ceiling.

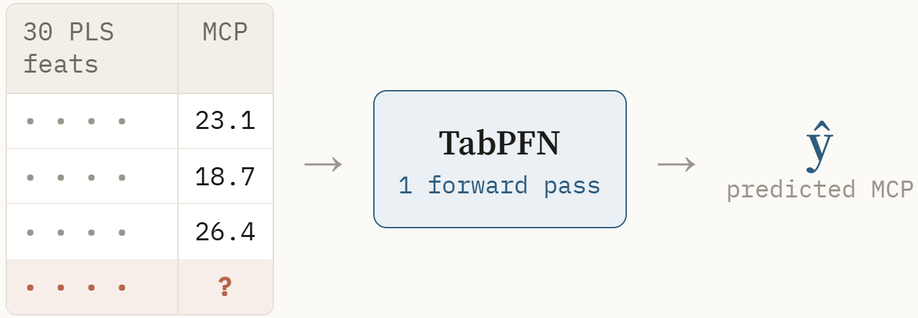
MODEL · 01 TABPFN

A transformer pre-trained to predict — not trained on your data

① PRE-TRAINED ONCE · OFFLINE

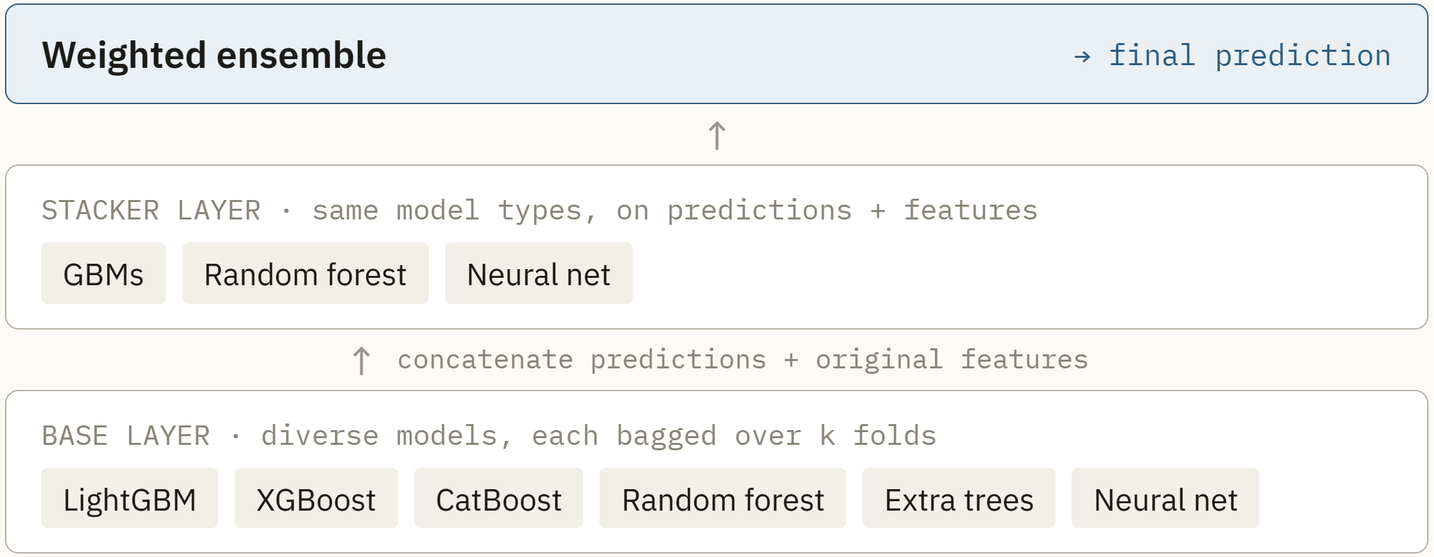
Pre-trained on **millions of synthetic tables** drawn from random causal structures — it learns how tabular prediction behaves in general. Trained once; never retrained.

② PREDICTS IN A SINGLE PASS



Every training cow — features + measured MCP — and the new sample go in together as **context**. One forward pass returns the answer: **no gradient updates, no tuning**.

Many models, stacked into one



An AutoML ensemble — not one model, but **a stack of many.**

- Trains many models at **preset configs** — tuning is optional, not the point.
- Accuracy comes from **bagging + multi-layer stacking**.

> Retrained **from scratch for every task**.

METHOD · DIMENSIONALITY

More features than cows — so PLS goes first

RAW SPECTRUM

2,051 features

2,051 > 1,608 cows

Highly collinear → models overfit.

prune uninformative regions + SNV →

PREPROCESSED

1,000 bands

Scatter-corrected — but **still highly collinear**.

PLS · supervised reduction →

PLS LATENT VARS

30 components

Decorrelated · **aligned to MCP**.

> Unlike PCA, PLS is supervised — its components maximise covariance with the measured MCP, so the 30 inputs to TabPFN & AutoGluon keep exactly the signal that matters.

Wold et al. 2001 (PLS) · Barnes et al. 1989 (SNV)

RESULTS · 01 RANDOM 10-FOLD

Random CV

All values are R^2 · best per trait in blue

Trait	PLS	PLS+AutoGluon	PLS+TabPFN
RCT	0.67	0.68	**0.70**
k20	0.55	0.56	**0.60**
a30	0.62	0.66	**0.68**
a60	0.47	0.48	**0.51**
Amax	0.60	0.60	**0.63**
A2R	0.68	0.66	**0.69**

RESULTS · 02 LEAVE-ONE-BATCH-OUT

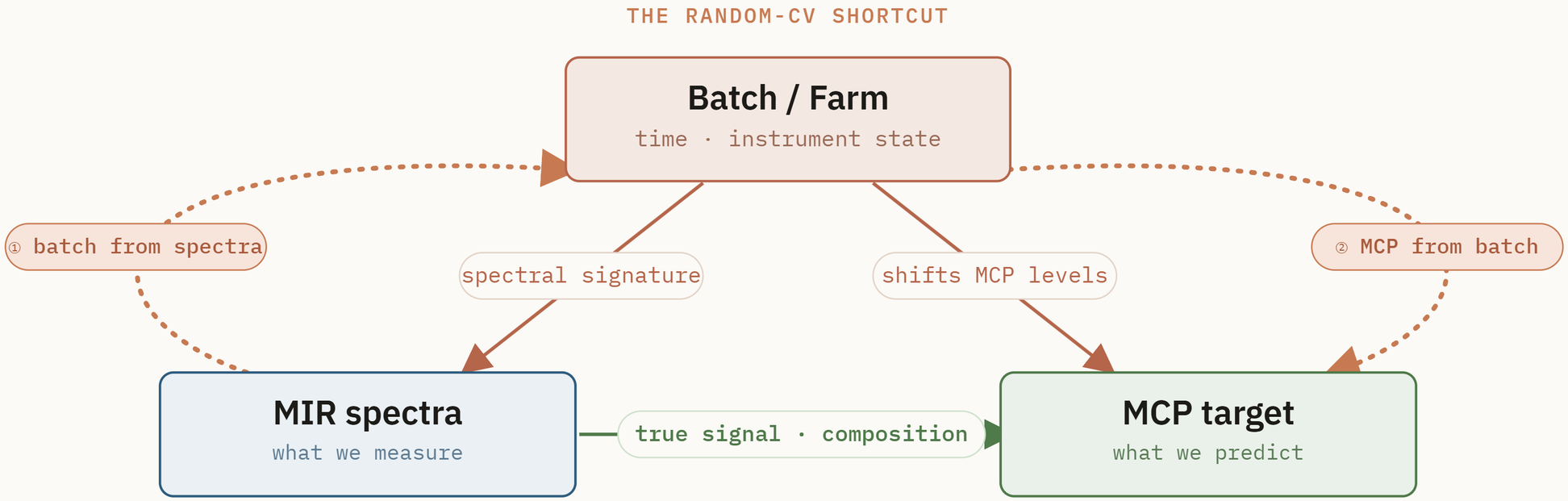
Honest LOBO: accuracy drops across the board

All values are R^2 · best per trait in blue

Trait	PLS	PLS+AutoGluon	PLS+TabPFN
RCT	0.65 (−3%)	0.64 (−6%)	**0.65** (−7%)
k20	0.53 (−4%)	0.56 (0%)	**0.58** (−3%)
a30	0.60 (−3%)	0.62 (−6%)	**0.64** (−6%)
a60	0.37 (−21%)	0.38 (−21%)	**0.42** (−18%)
Amax	0.54 (−10%)	0.55 (−8%)	**0.57** (−10%)
A2R	0.64 (−6%)	0.63 (−5%)	**0.65** (−6%)

WHY

Batch is the confounder



RANDOM CV
Train & test share batches → the shortcut works. **Inflated.**

LOBO
Test batch unseen → the shortcut is blocked. **Honest.**

WHY

Our pipeline reads the batch off the spectra

PLS + TabPFN · BATCH FROM SPECTRA

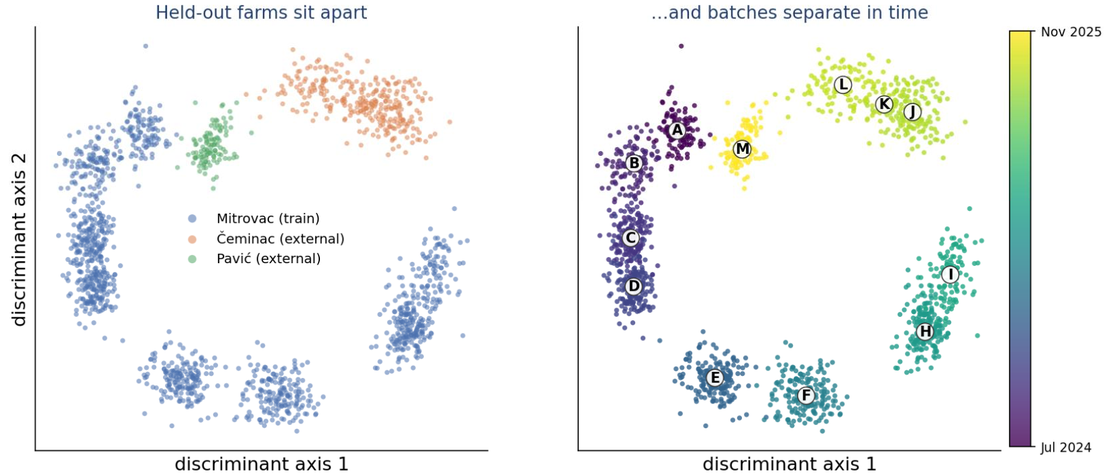
98%

balanced accuracy recovering the **analytical batch** — 1 of 12 · macro AUC ≈ 1.00.

84%
PLS-DA baseline

8%
chance level

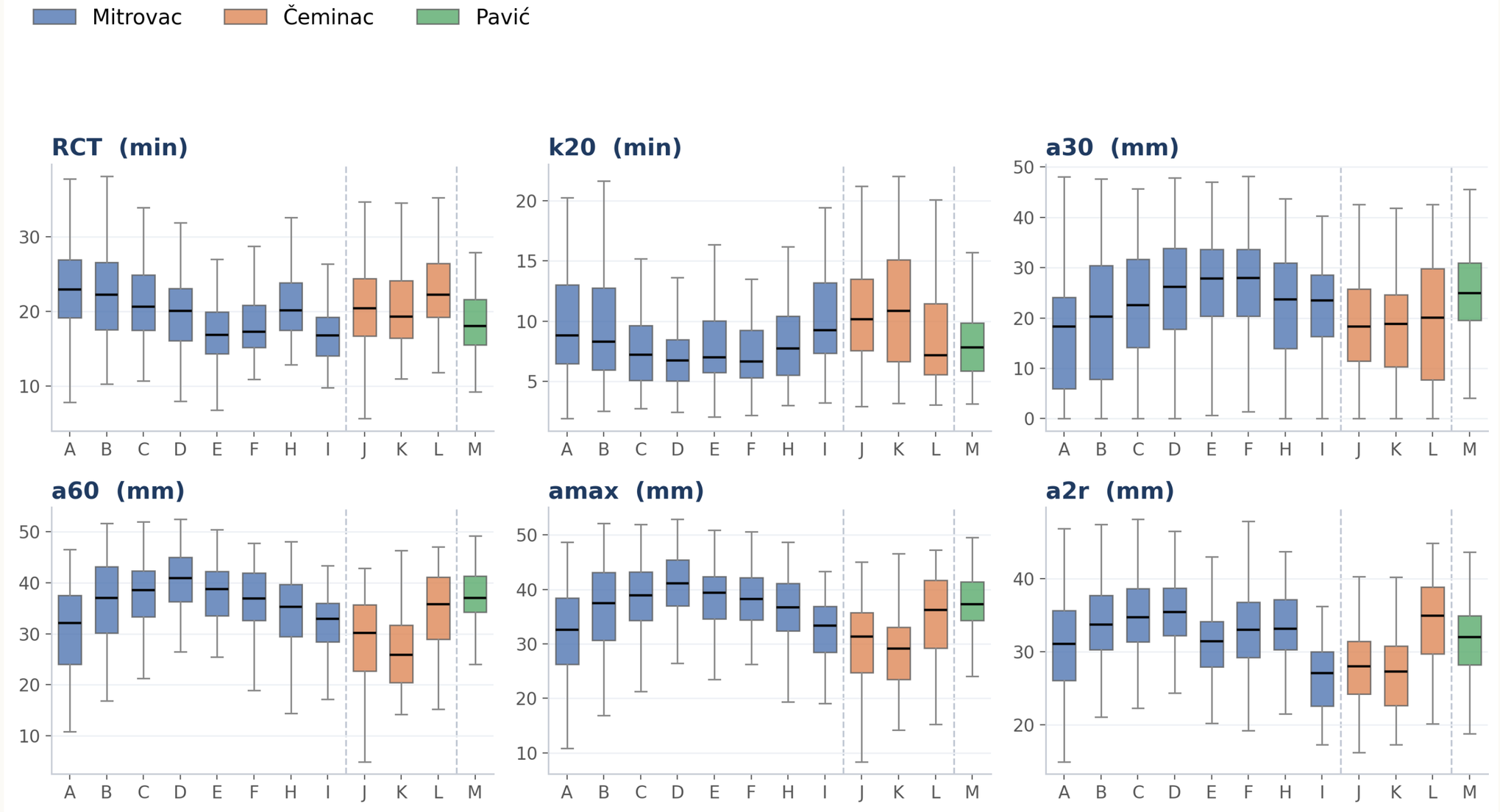
> The spectra carry a **strong batch fingerprint**.



MIR spectra · UMAP coloured by batch

PERMANOVA: batch ≈ 15.5–15.8% of total spectral variance (P < 0.001)

Milk-coagulation traits shift across the 12 analytical batches



Each batch carries its own MCP distribution — with the spectral fingerprint, this is what random CV exploits

CONCLUSIONS

Validate honestly; the model is the easy part

01

MODEL

PLS + TabPFN gives the best honest accuracy — no tuning, in seconds. A first for dairy.

02

VALIDATION

Random CV overstates accuracy — the analytical batch is a hidden confounder.

03

THE LIMIT

The ceiling is data quality, not model power.

OUTLOOK

Better information and better targets — not bigger models

01 ADD ORTHOGONAL DATA

Genomics · NIR · fluorescence

Independent signal — e.g. PLS on SNP markers — to close the **34–47% gap** to the measurement ceiling.

02 AIM AT BETTER TARGETS

Cheese yield · fat & protein yields

Predict what processors **actually pay for**.

> Not more cows, not bigger models — accuracy has already plateaued.

ACKNOWLEDGEMENTS

Thank you for your attention

COLLABORATING INSTITUTIONS

Financed by the European Union – NextGenerationEU within the project “Next Generation Animal Production” (NPOO.C3.2.R3-I1.04.0141), and by the Croatian Science Foundation under project HRZZ IP-2022-10-6914.