Modeling Healthcare Lab Panels with VAEs, Evaluating Reconstruction and Anomaly detection for Quality control

Kamakoti Jagannath Chivukula

Abstract: Electronic health record (EHR) laboratory data are often incomplete, despite strong physiological relationships among analytes within standard laboratory panels. We present an unsupervised framework for modeling routine hospital laboratory panels: Complete Blood Count (CBC), Hepatic Function Panel (HFP), and Comprehensive Metabolic Panel (CMP), using large-scale de-identified records from Weill Cornell Medicine. Laboratory values are z-score normalized and paired with binary observation masks to distinguish observed from missing analytes. We train β-Variational Autoencoders (β-VAEs) that represent missing analytes using learnable per-feature embeddings rather than conventional zero imputation. The proposed approach achieves strong reconstruction performance and improves masked leave-one-feature-out prediction for multiple physiologically correlated analytes. To evaluate unsupervised anomaly detection, we simulate laboratory Quality Control scenarios by corrupting highly predictable analytes and identifying abnormal samples using reconstruction error. Results demonstrate that masked β-VAEs capture clinically meaningful structure in routine laboratory data, enabling more effective Quality Control and robust modeling of incomplete EHR laboratory records.

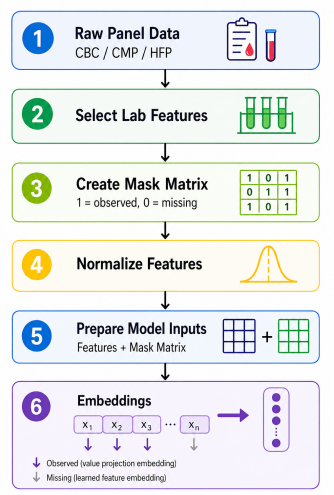
Introduction: Quality control (QC) in the clinical laboratory involves monitoring analytic performance to ensure that reported patient results remain accurate, precise, and clinically reliable. The central challenge is detecting when a measurement, or a batch of measurements, deviates from expected behavior before erroneous values propagate into the patient record. Such deviations may arise from instrument drift, reagent degradation, calibration failure, or sample mix-ups, all of which can produce laboratory values inconsistent with stable operating conditions. This problem is both clinically and operationally significant, as undetected laboratory errors can affect diagnosis and treatment decisions while also increasing costs through unnecessary repeat testing.

Historically, laboratory QC has relied on statistical process-control methods that flag observations exceeding fixed mean ± (k) standard deviation thresholds. Although these approaches remain the foundation of routine QC practice, they were designed primarily for univariate monitoring of individual analytes measured on control specimens. In contrast, modern electronic health record (EHR) laboratory data consist of multivariate panels in which physiologically correlated analytes are reported together. In this setting, abnormality may arise not from a single extreme value, but from inconsistency within the overall analyte profile.

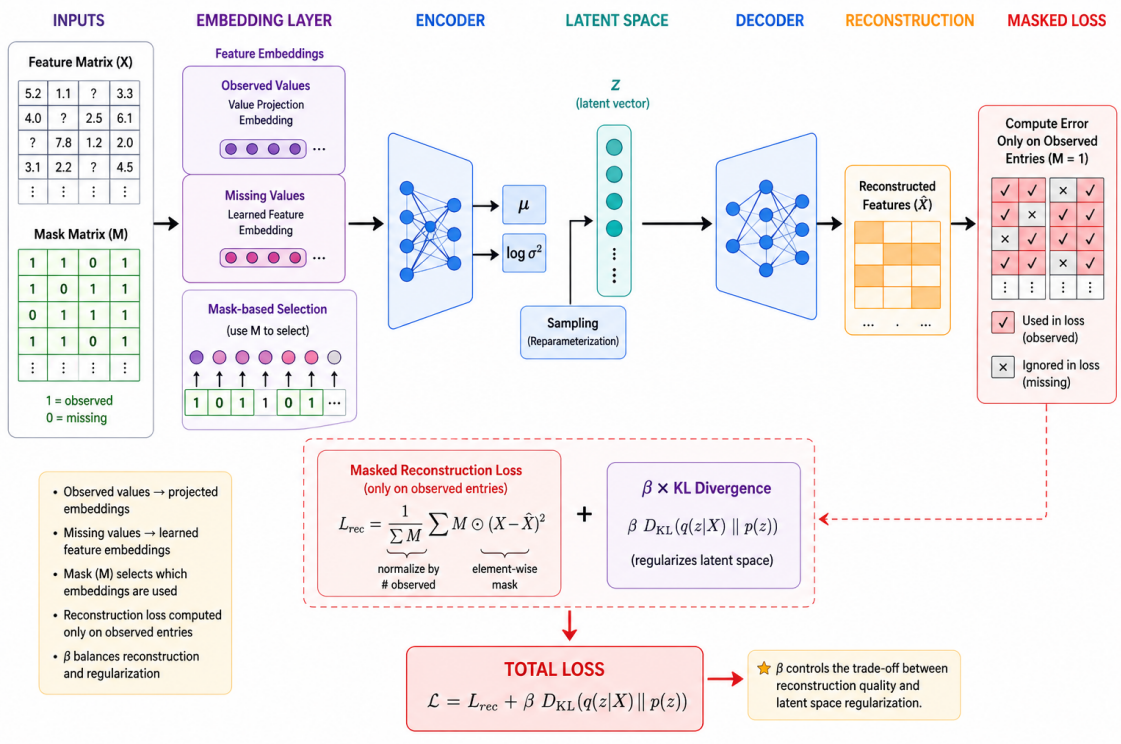
Consequently, there is growing interest in multivariate statistical and machine learning approaches that can complement traditional QC methods by modeling joint analyte relationships using patient or pooled laboratory data. Such methods may improve detection of subtle panel-level failures that univariate rules fail to identify, while remaining interpretable and compatible with established laboratory quality-assurance workflows.

Methodology:

Data Preprocessing: After normalization, the feature matrix and mask are used to construct inputs for the VAE. Observed values use projected embeddings, while missing values use learned feature embeddings selected via the mask. Embeddings are learned during training.

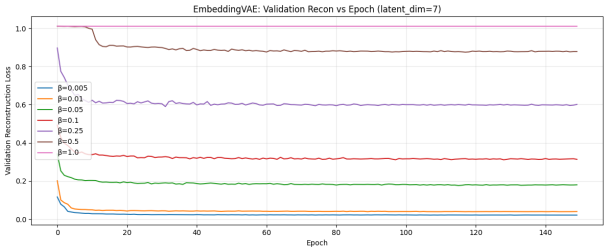


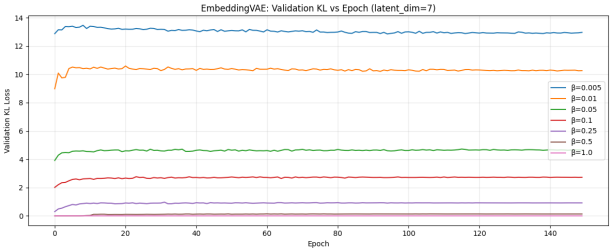
Training: The embeddings from input layer are encoded into a low-dimensional latent space and decoded to reconstruct the original features, optimizing a masked reconstruction loss (on observed entries only) plus β × KL, where β controls the reconstruction–regularization trade-off.

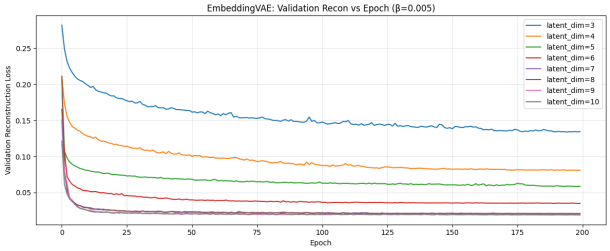


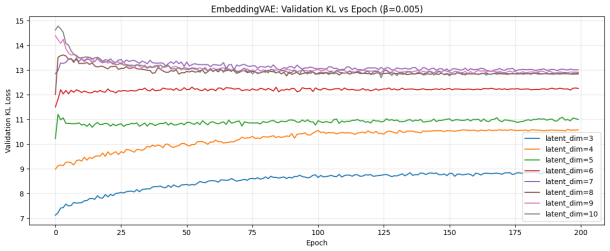
Training results:

Complete Blood Count (CBC):

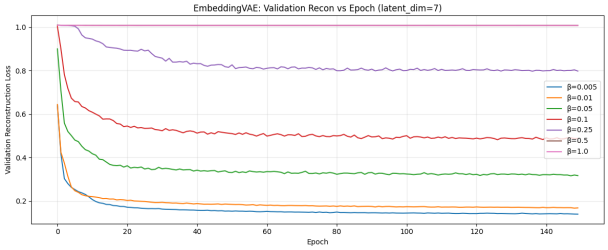


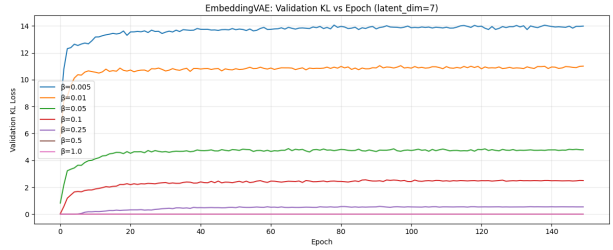


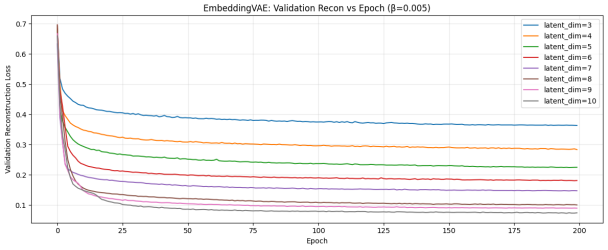


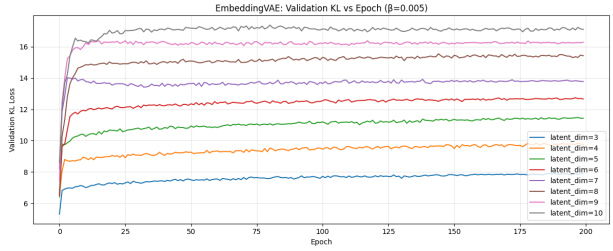


Comprehensive Metabolic Panel (CMP):

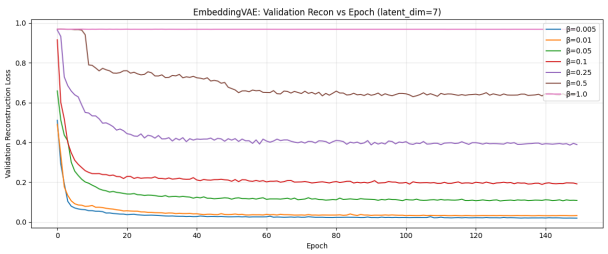


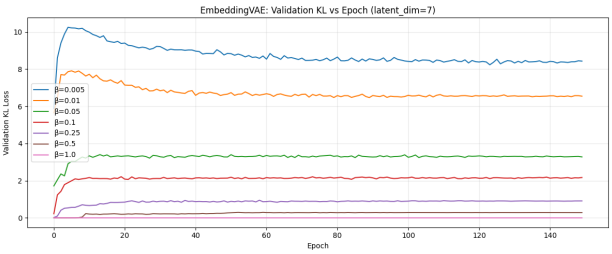


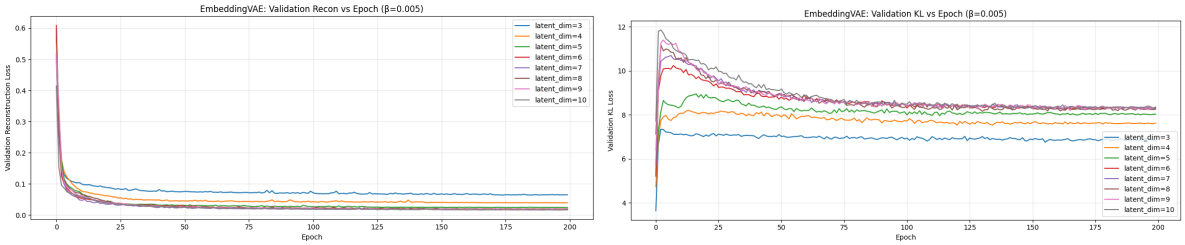




Hepatitic Function Panel (HFP):





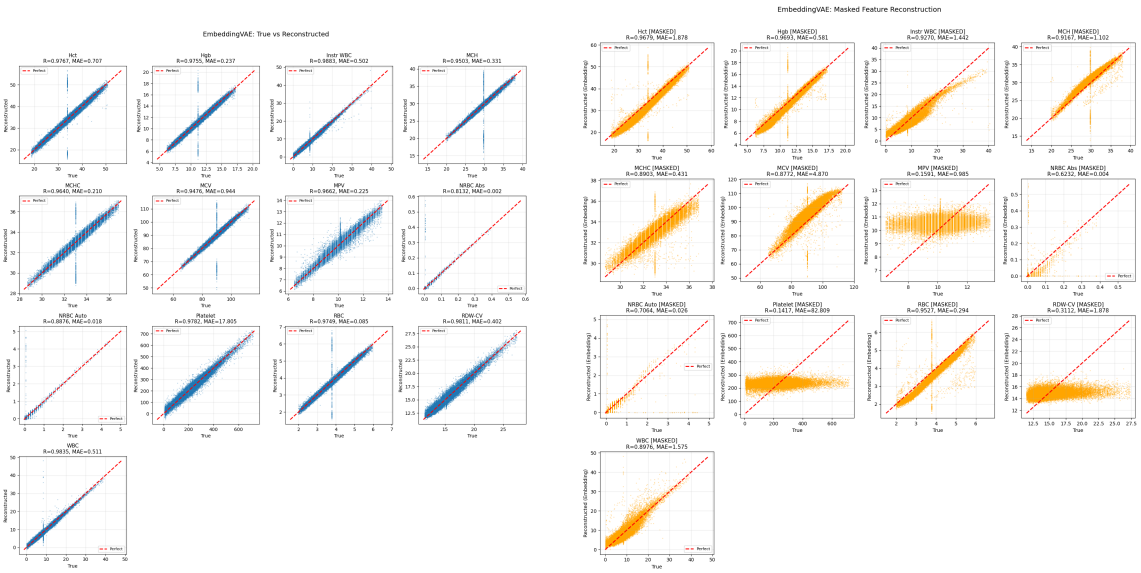


With $\beta = 1$, the KL term dominates the VAE objective and strongly regularizes the approximate posterior toward the prior $p(z) = \mathcal{N}(0, I)$. As a result, the encoder learns latent representations that contain little or no information about the input, leading to posterior collapse. The decoder therefore relies primarily on its unconditional bias and produces nearly constant predictions, yielding a masked MSE close to 1, which is comparable to predicting the global mean after normalization.

This behavior appears consistently across CBC, HFP, and CMP panels because all three use the same masked reconstruction objective and identical β-weighted ELBO formulation. Reducing β(for example, to 0.005) weakens the KL regularization, allowing the latent variables to encode informative structure from the input and thereby improving reconstruction accuracy.

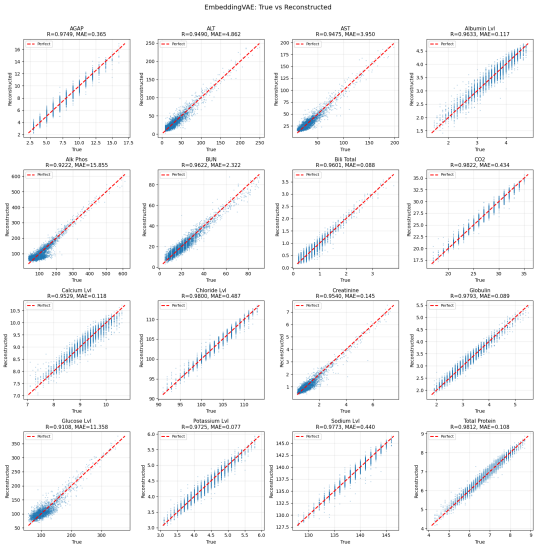
Evaluation: During evaluation, the trained model is tested in two ways: (1) normal reconstruction using all inputs, and (2) masking-based inference, where one feature is removed and predicted from the others. Metrics are computed only where the true value is available, using Pearson's r to measure how well each feature is reconstructed. A high masked r means the feature can be reliably inferred from others (high dependency), while a drop in r after masking indicates the feature carries more unique information. This comparison highlights which lab features are redundant versus uniquely informative.

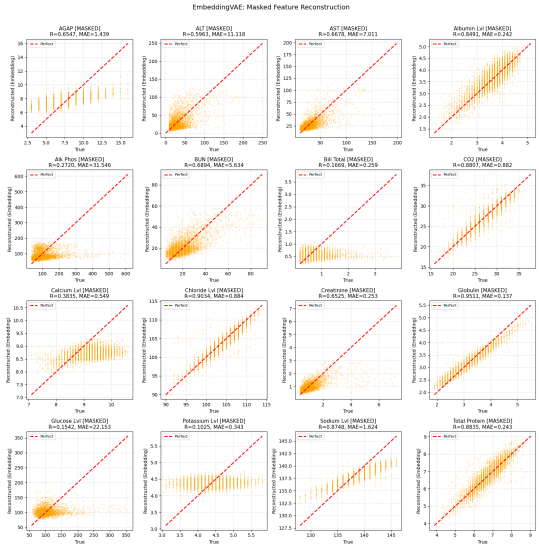
CBC:



Hct	0.976680	0.967867	0.008813
Hgb	0.975458	0.969169	0.006289
Instr WBC	0.988338	0.927985	0.060353
MCH	0.950262	0.915380	0.034882
MCHC	0.963981	0.888093	0.075888
MCV	0.947608	0.877163	0.070446
MPV	0.966231	0.159027	0.807203
NRBC Abs	0.813181	0.626038	0.187144
NRBC Auto	0.887649	0.705989	0.181659
Platelet	0.978153	0.157779	0.820374
RBC	0.974950	0.952527	0.022423
RDW-CV	0.981069	0.308095	0.672974
WBC	0.983464	0.897596	0.085868

CMP:

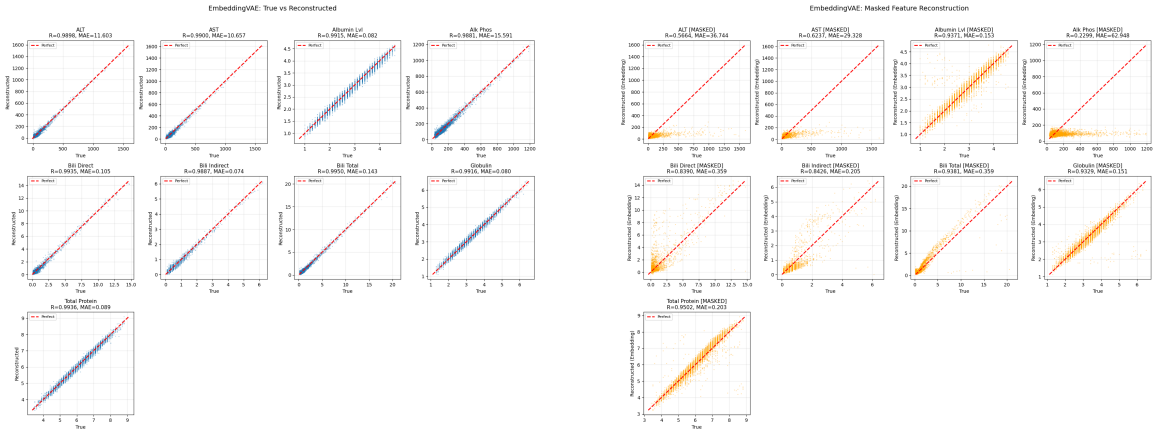




AGAP	0.976098	0.659636	0.316462
ALT	0.951570	0.592301	0.359269
AST	0.945920	0.663841	0.282079
Albumin Lvl	0.963405	0.849171	0.114234
Alk Phos	0.921224	0.271251	0.649972
BUN	0.961968	0.684579	0.277388
Bili Total	0.961457	0.167600	0.793857
CO2	0.982482	0.878985	0.103497
Calcium Lvl	0.953302	0.380840	0.572462
Chloride Lvl	0.980153	0.904695	0.075459
Creatinine	0.953184	0.652511	0.300673
Globulin	0.978571	0.952257	0.026313
Glucose Lvl	0.909979	0.149204	0.760775
Potassium Lvl	0.971632	0.093408	0.878224
Sodium Lvl	0.976834	0.872165	0.104669
Total Protein	0.980683	0.885507	0.095176

HFP:

ALT	0.989811	0.565608	0.424203
AST	0.990007	0.625459	0.364548
Albumin Lvl	0.991541	0.934639	0.056902
Alk Phos	0.988147	0.243816	0.744331
Bili Direct	0.993499	0.829407	0.164093
Bili Indirect	0.988679	0.844194	0.144485
Bili Total	0.995022	0.937419	0.057602
Globulin	0.991582	0.932634	0.058948
Total Protein	0.993554	0.951288	0.042266

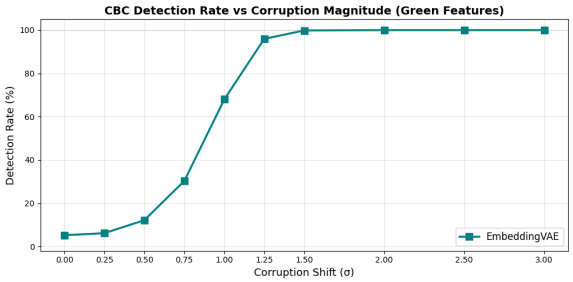


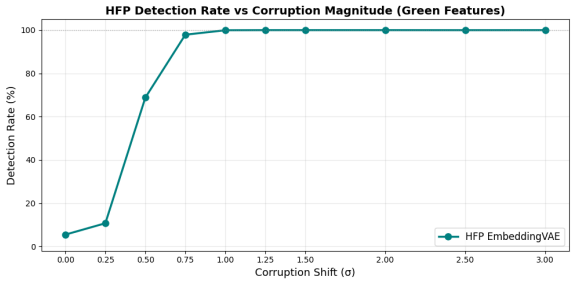
The first column represents evaluation using (1) and the second column using (2) and the last column is the difference i.e.; (1) – (2). The blue plot represents reconstructed vs true using (1) and orange plot using (2).

Anomaly Detection for Quality Control: The method focuses on "green features," defined as features that the model reconstructs most reliably under masking-based inference. Synthetic anomalies are generated by perturbing only these features in otherwise clean validation samples, allowing evaluation of the model's sensitivity to deviations in highly predictable laboratory measurements.

For each sample, the reconstruction error is computed exclusively over the selected green-feature subset. An anomaly threshold is then determined from the distribution of reconstruction errors on normal validation data, using the 95th percentile as the cutoff. Samples whose reconstruction error exceeds this threshold are classified as anomalous for Quality Control.

Because the selected green features are normally reconstructed with high accuracy, perturbations in these features produce disproportionately larger reconstruction errors, making them effective targets for Quality Control.





Conclusion: The proposed Quality Control framework leverages "green features," defined as laboratory variables that are reconstructed most reliably under masking-based inference, indicating strong learnable dependencies within the panel. By focusing on highly predictable features, the framework amplifies deviations from expected inter-feature relationships, enabling effective detection of abnormal laboratory patterns through reconstruction-based scoring.