# BSTabDiff

## Block-Subunit Diffusion Priors for High-Dimensional Tabular Data Generation

**Al Zadid Sultan Bin Habib**[1], Md Younus Ahamed[1]
Prashnna Kumar Gyawali[1], Gianfranco Doretto[2]
Donald A. Adjeroh[1]

[1]West Virginia University [2]University of Utah

zadidhabib.com/bstabdiff.html

# Thousands of features, very few samples

**High-Dimensional, Low-Sample Size data**

$$n \ll m$$

Example: omics tables with thousands of measurements and few subjects.

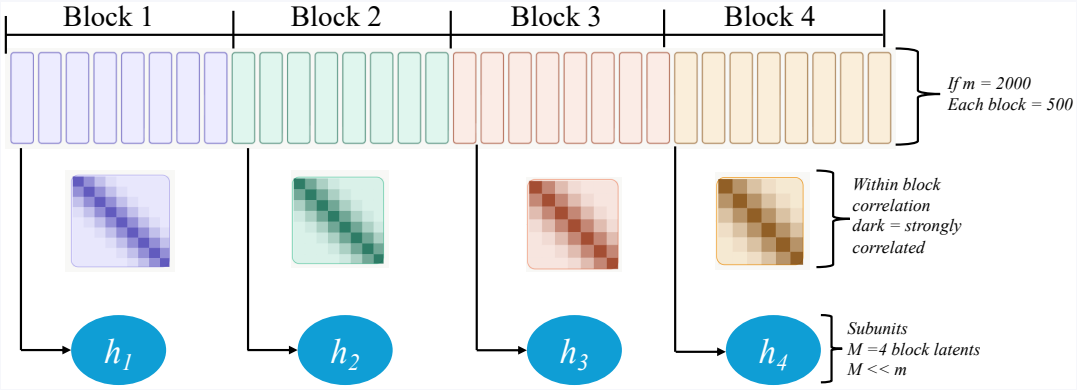
HDLSS geometry: [1].

**What a generator must capture**

- Strong dependence within feature groups
- Dependence between groups
- Non-Gaussian marginals and missingness

**Goal:** useful synthetic tables without learning dense dependence directly in $\mathbb{R}^m$.

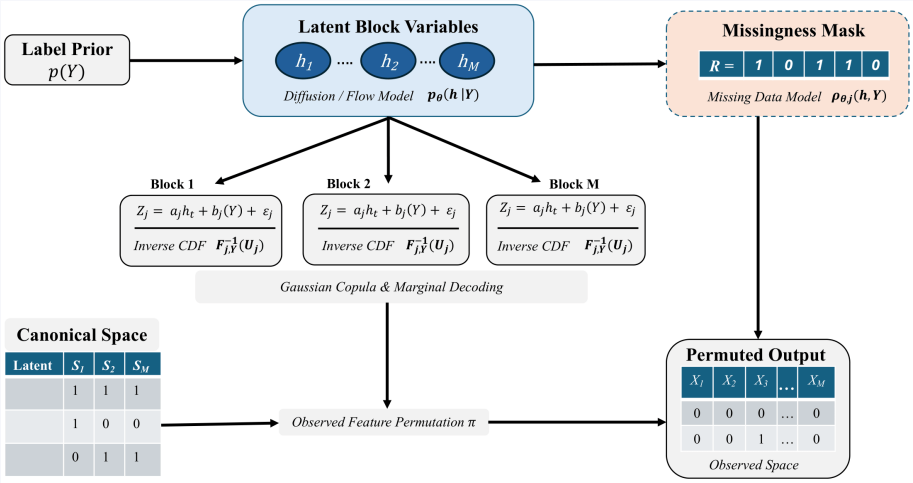
# A shared subunit for each feature block




**2,000 features** Four blocks of 500 features

**4 latent subunits** A joint prior across blocks

# Compact latent priors, full-dimensional synthesis



# GO-BS aligns blocks with feature dependence

**Graph-guided Ordering with Block Segmentation**

Build a feature-affinity graph.

Place related features together and estimate contiguous block boundaries.

Keep the learned permutation fixed at generation time.

**Structural objective**

Reward within-block affinity.

Penalize cross-block leakage.

Regularize block boundaries.

**NP-hard objective** Sparse approximate solver.

**Optional structure:** GO-BS, GO-BS-FC, or the original feature order.

# Higher utility than the strongest competing baseline

**Eight core HDLSS datasets** TSTR accuracy (%), mean over $5 \times 5$ CV

| Dataset | Best non-BSTabDiff | Baseline | **BSTabDiff** | **Gain (pp)** |
|---|---|---|---|---|
| COL | TabDiff | 79.72 | **83.26** | +3.54 |
| LNG | TabDiff | 88.64 | **95.96** | +7.32 |
| GLI | TVAE | 79.23 | **84.94** | +5.71 |
| SMK | SMOTE | 68.44 | **72.08** | +3.64 |
| AML | SMOTE | 93.22 | **95.56** | +2.34 |
| PRS | SMOTE | 87.16 | **90.76** | +3.60 |
| ARC | SMOTE | 79.08 | **86.50** | +7.42 |
| TOX | TabDiff | 86.32 | **89.36** | +3.04 |

Table 2. Best default/tuned BSTabDiff vs. strongest other synthetic method; real data TRTR excluded.

**Nine additional datasets:** mean AUC **89.21%** vs. **85.16%** (SMOTE) **+4.05 pp**

Table 4. BSTabDiff-no-p. TSTR: train on synthetic, test on real.

# A small latent prior keeps generation practical

**Where the work happens**

Global dependence learning operates on $M$ block latents.

Feature-wise inference and decoding still process the full $n \times m$ table.

**Energy and $CO_2$e reporting** complements synthetic utility.

**4.36–9.34 seconds**

reported runtime per run

**0.030–0.209 GiB**

peak GPU memory

Eight core datasets; Vulcan.

Single runs, no permutation or tuning.

Paper Table 1. Measurements are configuration and hardware-specific.

Green AI context: [2]. Ordering and hyperparameter search incur additional cost.

# BSTabDiff: learn compactly, generate all features

**Block latents carry global dependence.**
Feature-wise emissions recover the full table.

**Explore the paper and implementation**

**Project** zadidhabib.com/bstabdiff.html

**Code** github.com/zadid6pretam/BSTabDiff

**Hugging Face: BSTabDiff Studio**

https://huggingface.co/spaces/zadid6pretam/BSTabDiff-Studio

```
pip install bstabdiff
```

ah00069@mix.wvu.edu

**Try the live demo**

## Thank you!

Supported in part by US National Science Foundation (NSF) Awards #1920920, #2125872, #2223793, and #2318708.

# Selected references




Peter Hall, James Stephen Marron, and Amnon Neeman.
Geometric Representation of High Dimension, Low Sample Size Data.
*Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):427–444, 2005.



Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni.
Green AI.
*Communications of the ACM*, 63(12):54–63, 2020.