// ACADEMIC / 2026 / NOV

Copula and mixture prior variational autoencoders for small sample blasting fragmentation modelling

Paulo Lopes
https://linktr.ee/pftlopes

Abstract

Synthetic tabular data may support blasting research when observations are scarce, but distributional resemblance and predictive transfer answer different questions. This study evaluates an existing fragmentation benchmark containing 262 harmonised records from three literature sources and compares Gaussian, Student-t and Clayton copulas with vanilla, Kullback-Leibler annealed and Gaussian mixture prior tabular variational autoencoders. The reported protocol separates a real hold-out set before generator fitting and combines likelihood summaries, maximum mean discrepancy, an autoencoder embedding distance and train-on-synthetic-test-on-real utility. The mixture prior model attained the highest reported utility, with approximately 0.736 for R^2 and 1.698 for root mean squared error in fragmentation-index units; corresponding fidelity summaries were approximately 0.160 and 0.436. Annealing improved utility relative to the vanilla model, whereas copula-derived generator inputs generally reduced performance. These observations support multimodal latent modelling as a useful candidate for heterogeneous fragmentation tables. They do not establish an augmentation gain over a matched real-only predictor or transfer to an unseen mine. Interpretation is further limited by a discrepancy between the six-variable utility schema and the eight-dimensional parameter count implied by the copula summary, incomplete seed-level reporting and potentially shared blast records across literature sources. The contribution is a technically qualified comparison and a reproducible evaluation specification that distinguishes synthesis, post-blast inference and prospective engineering prediction.

Keywords: rock blasting; fragmentation; synthetic tabular data; copula; variational autoencoder; mixture prior; small samples; evaluation

1 Introduction

Fragmentation connects drill-and-blast design with loading, haulage, crushing and comminution. A useful data model must preserve the relationships between explosive loading, blast geometry, rock properties and measured fragmentation, rather than only reproduce plausible values in each column. Literature databases remain valuable in this setting, although their measurement conventions, sampling frames and levels of aggregation can differ. Neural-network and multivariate studies of fragment size demonstrate the relevance of these variables while also showing why source provenance matters [1,2]. More recent work combines clustering, hyperparameter optimisation and augmentation [3]; its existence motivates careful assessment of which component contributes useful information.

Generating more rows does not create more independent blast observations. A synthesizer estimates a distribution from a finite dataset and samples from that estimate. It may smooth an unstable empirical distribution, encode a useful prior or change how a downstream learner is regularised. It may also repeat observations, attenuate rare regimes or amplify errors introduced during data harmonisation. Reviews of tabular generation therefore distinguish the mechanism of synthesis from the task and evidence used to evaluate it [4]. The present question is whether synthetic records preserve dependencies relevant to the fragmentation-index task within a small heterogeneous benchmark.

Copulas and variational autoencoders provide contrasting inductive biases. A copula separates marginal behaviour from dependence and permits explicit inspection of a correlation or dependence parameter [5]. A variational autoencoder learns a shared latent representation, with its reconstruction model and prior jointly controlling the generated distribution [6]. The TVAE benchmark established a practical tabular implementation [7], while mixture-based latent modelling offers one route to representing multiple regimes [8]. Additional flexibility is attractive for heterogeneous sources, but it increases estimation and model-selection uncertainty when the real training set contains only a few hundred records.

Evaluation requires more than a single quality score. Marginal fit does not ensure preservation of conditional responses; predictive success does not prove support coverage or absence of memorisation. Kernel distribution distances [9], task transfer [10] and sample-level fidelity and coverage diagnostics [11] address different failure modes. Their interpretation also depends on the independence of the evaluation data. Leakage through preprocessing, repeated model selection or structurally related observations can produce apparently convincing results without genuine out-of-sample evidence [12-14].

2 Materials and methods

2.1 Data provenance and harmonisation

Eight candidate datasets were reviewed in the underlying project. Three source tables were retained for the fragmentation-index pool: hudaverdi_2010_full with 110 records, hudaverdi_2012 with 62 records and kulatilake_2010 with 90 records, giving 262 rows after vertical concatenation [15]. Source labels are identifiers of compiled tables, rather than evidence of three independent mine populations. A row key prevents index collisions but does not detect the reuse of a physical blast in different publications. This distinction is relevant because the cited fragmentation literature draws on previously compiled blast databases [1,2].

The harmonisation schema contains site, blast_id, S/B, H/B, B/D, T/B, powder factor, Young's modulus, the fragment-size field labelled X50, fragmentation index FI, group and mesh area. The reported utility experiments retain FI and five predictors with complete coverage: X50, PF, H/B, B/D and T/B. Table 1 defines the engineering nomenclature. Source records describe mean fragment size in parts of the underlying literature [1,2]; consequently, the field label X50 must not by itself be interpreted as proof of a mass-based median. Establishing whether each source reports a mean, a median or a fitted characteristic size is necessary before treating cross-source values as the same physical estimand.

Table 1. Engineering variables and units in the harmonisation schema. X50 is the retained database label; its percentile interpretation requires source-level confirmation. Ratios assume that numerator and denominator lengths have been converted to the same unit.

Symbol/name	Definition	Unit
S (spacing)	Centre-to-centre spacing between adjacent blastholes in a row	m
B (burden)	Distance from a blasthole to the free face, measured normal to the face	m
H (bench height)	Vertical height of the bench being blasted	m
T (stemming length)	Length of inert stemming material at the hole collar	m
D (hole diameter)	Drillhole diameter	m
A (mesh area)	Blast-pattern cell area	m^2
S/B	Spacing-to-burden ratio	-
H/B	Bench-height-to-burden ratio	-
T/B	Stemming-to-burden ratio	-
B/D	Burden-to-diameter ratio	-
X50	Fragment-size field labelled X50; source estimand requires confirmation	m
PF	Explosive mass per rock volume	$kg \cdot m^{-3}$
E	Young's modulus as defined by the source	GPa
FI	Source-reported fragmentation index	-
ssB, ss50	Specific surface measures used for back-calculation	m^{-1}

Where mesh area A and spacing-to-burden ratio r are available, burden and spacing follow from their geometric definitions:

$$B = \sqrt{A/r}, \qquad S = \sqrt{A \cdot r} \qquad (1)$$

The specific-surface conversions retained in the benchmark are:

$$\begin{aligned} \mathrm{ssB} &= 6/B, \quad \mathrm{ss50} = 6/X_{50}, \quad B = 6/\mathrm{ssB}, \\ S &= (S/B) \cdot B, \qquad X_{50} = 6/\mathrm{ss50} \end{aligned} \qquad (2)$$

Equation (1) is a geometric identity when A is the rectangular pattern-cell area and r is positive. Equation (2) requires the particular specific-surface convention used in the source; the coefficient 6 is not a universal conversion for irregular fragments. Both conversions preserve dimensions when lengths are expressed in metres and specific surfaces in inverse metres. Their use should be documented per source, together with rounding precision and any assumptions about representative particle shape.

The six-variable utility pool is reported as complete, so no imputation was required for that subset. Unreported fields in the wider schema remain missing. FI bands were encoded in this benchmark as FI < 4, 4 ≤ FI ≤ 8 and FI > 8. These endpoints are retained as the benchmark convention rather than silently substituted with a different grouping. Counts at exactly 4 and 8 are needed when comparing results with studies using other endpoint rules.

2.2 Exploratory structure and source dependence

The reported FI medians are 3.30, 5.11 and 3.59 for the three source tables, respectively. Mean values of the field labelled X50 are 0.299, 0.205 and 0.316 m [15]. These summaries indicate source-level differences, but not their cause. Geology, sampling, measurement definitions, blast design and data reuse are all possible contributors. Empirical cumulative distributions in Figure 1 provide a bin-independent view of this heterogeneity without assuming a parametric density.

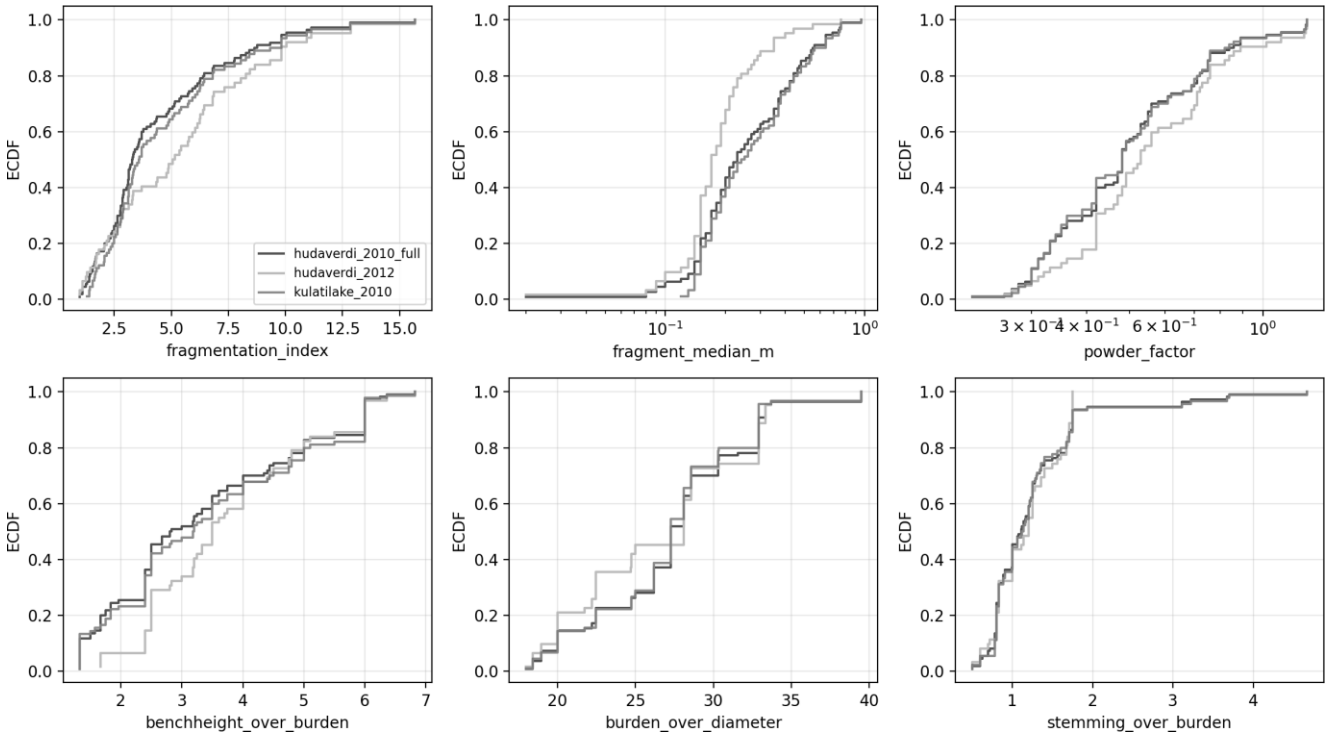


Figure 1. Source-specific empirical cumulative distributions for the six core utility variables. The three source traces and all original panels are retained in monochrome. Differences describe the compiled source tables and do not identify a causal geological effect.

Figure 2 contains an eight-variable exploratory analysis, including S/B and Young's modulus in addition to the utility variables. The difference between this diagnostic schema and the six-variable utility schema must be kept explicit. Correlations in the pooled table combine within-source and between-source variation. Accordingly, a stronger pooled association is not evidence that the corresponding relationship holds within every source, and weaker pairwise association does not exclude nonlinear conditional dependence.

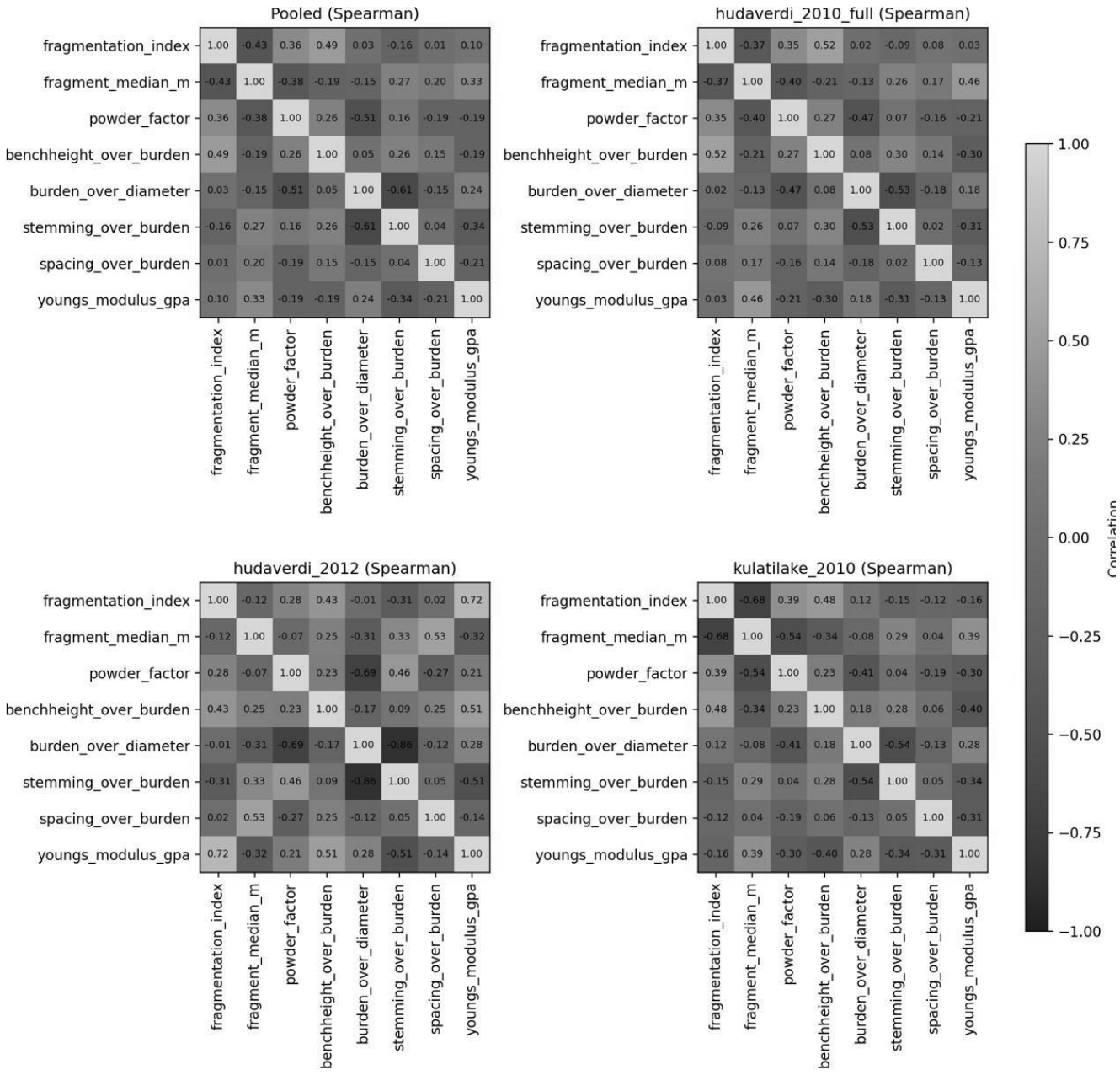


Figure 2. Original pooled and source-specific Spearman correlation matrices. The diagnostic uses eight variables; the reported utility task uses six. Cell annotations are retained and provide the numerical interpretation of the monochrome heatmaps.

2.3 Data partitioning and the scope of leakage control

The reported protocol uses a shuffled 80/20 real-data partition, corresponding to approximately 209-210 training records and 52-53 test records depending on the split implementation. Marginals, scalers, generator parameters and tuning choices are described as training-derived [15]. This ordering protects against direct test-to-training contamination. It does not establish independence if different rows represent the same blast, the same campaign or strongly overlapping source databases.

The random partition estimates interpolation within the mixture of represented sources. It is not a held-out-source or held-out-mine experiment. Lack of source stratification adds composition

variability; it does not automatically make the test more realistic. For prospective deployment, a split by physical blast, monitoring campaign or mine should reflect the intended unit of generalisation [13]. The existing split remains the reference for the reported results; the blocked protocol is a validation extension, not an experiment already performed.

Neural-generator inputs are standardised with real-training means and standard deviations and transformed back to engineering units after sampling. The downstream XGBoost regressor uses the physical-scale variables. No global target transformation or winsorisation is reported. An inner training-only validation layer is required to operationalise any hyperparameter search. The summary does not provide sufficient fold assignments and optimisation logs to independently verify this nesting, so leakage resistance is treated as a documented design property rather than a completed audit [12,14].

2.4 Copula models and parameter accounting

For continuous marginals, Sklar's representation separates the joint cumulative distribution from its dependence function [5]:

$$F(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big) \qquad (3)$$

Training observations are transformed through fitted marginal cumulative distributions:

$$U_j = \hat{F}_j\big(X_j\big) \in (0,1) \qquad (4)$$

The resulting values are clipped to the interval from ε to 1 - ε, with ε = 0.001. Clipping avoids infinite normal quantiles but also modifies the most extreme pseudo-observations. Rank ties, discretised blast-design ratios and the inverse-transform interpolation rule can therefore affect tail behaviour. Continuous-copula likelihoods are approximations when a nominally numerical variable has only a few distinct design levels.

The elliptical-copula correlation estimate uses the Kendall-tau relationship:

$$\hat{\rho}_{jk} = \sin\left(\frac{\pi}{2}\,\hat{\tau}_{jk}\right) \qquad (5)$$

A fixed shrinkage coefficient of 0.05 and, where needed, eigenvalue flooring stabilise the estimated matrix:

$$\rho \leftarrow (1-\lambda)\rho + \lambda I, \qquad \lambda = 0.05 \qquad (6)$$

For the Gaussian copula, the observation log-density is:

$$\log c_\rho(u) = -1/2 \log|\rho| - 1/2\, z^{\top}(\rho^{-1} - I)z \qquad (7)$$

Shrinkage improves conditioning, but likelihood values computed after rank estimation and matrix repair are not necessarily maximised full likelihoods. Equation (8) is the conventional AIC expression [16]; its use as a formal information criterion requires a common observation space, a common marginal treatment and a justified parameter count.

$$\mathrm{AIC}/n_{\mathrm{train}} = \frac{2k - 2\log\hat{L}_{\mathrm{train}}}{n_{\mathrm{train}}} \qquad (8)$$

The Student-t copula adds tail dependence through its degrees of freedom, while the Clayton family offers a parsimonious, asymmetric dependence form. A conventional positive-parameter multivariate Clayton model is exchangeable and positively associated. It cannot reproduce arbitrary mixtures of

positive and negative pairwise associations such as those in Figure 2. Lower-tail dependence concerns the joint behaviour of low marginal quantiles; it is not justified merely by right-skewed marginals [5]. Rotations or more flexible constructions would have to be specified explicitly to change these restrictions.

Synthetic uniform draws are returned to engineering space by the fitted inverse marginals:

$$x_j^{(s)} = \hat{F}_j^{-1}\left(u_j^{(s)}\right), \quad s = 1, \dots, m; \quad j = 1, \dots, d \qquad (9)$$

Figure 3 preserves the original workflow. Its AIC selection box describes the historical implementation. The reported Gaussian count k = 28 corresponds to 8 × 7 / 2 correlation parameters, whereas a six-dimensional Gaussian copula has 15. The summary does not resolve whether the copula comparison was fitted to eight variables and the utility comparison to six, or whether the parameter count was carried over incorrectly. The numerical values are therefore retained as reported fit summaries, with no newly asserted dimension-consistent AIC ranking.

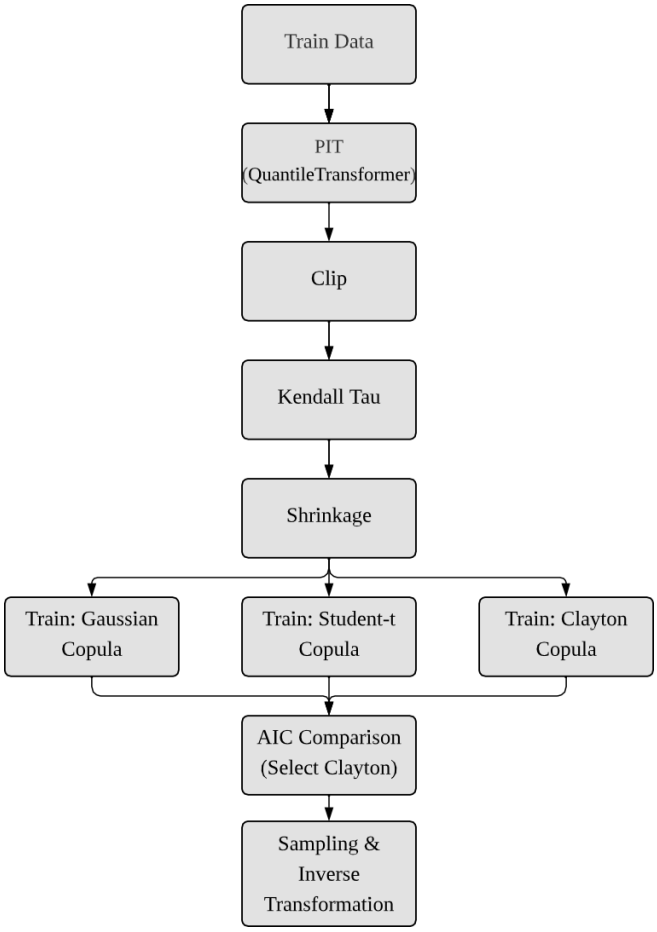


Figure 3. Original copula processing and sampling workflow. The diagram records the historical selection step; Section 2.4 explains the unresolved feature-dimension and parameter-count issue. All nodes, arrows and labels are preserved.

2.5 Variational autoencoders and generator-input regimes

The three neural alternatives are a vanilla tabular VAE with a standard-normal prior, an annealed variant and a model with a Gaussian mixture prior [6-8,17]. In this benchmark, the description emphasises standardised continuous inputs. It should therefore be distinguished from assuming

every preprocessing and likelihood choice of the packaged SDV TVAESynthesizer [18]. The exact implementation, decoder likelihood and feature transformation are part of the model specification.

For latent variable z and approximate posterior q, the evidence lower bound is:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right) \qquad (10)$$

The reparameterisation is:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \qquad (11)$$

The annealed loss varies the relative KL weight over training:

$$\mathcal{L}_\beta(x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] + \beta_t D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right) \qquad (12)$$

Reducing early KL pressure can allow the reconstruction term to become informative before strong prior matching is enforced [17]. Better downstream performance alone does not diagnose posterior collapse in the vanilla model; such a diagnosis would also require latent-usage measures, reconstruction and KL trajectories, or an equivalent analysis.

The mixture prior is represented by:

$$p(z) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(z; \mu_k, \Sigma_k) \qquad (13)$$

Mixture weights must be non-negative and sum to one, and covariance matrices must satisfy the chosen positive-definiteness constraints. The mixture KL is expressed as an expectation, estimated by Monte Carlo:

$$D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right) = \mathbb{E}_{q_\phi}\left[\log q_\phi(z \mid x) - \log p(z)\right] \qquad (14)$$

The mixture log-density is:

$$\log p(z) = \log \sum_{k=1}^{K} \pi_k \, \mathcal{N}(z; \mu_k, \Sigma_k) \qquad (15)$$

In implementation, Eq. (15) should be evaluated using a stable log-sum-exp of log mixture weights and component log-densities. A sample estimate of Eq. (14) is noisy; it is not an analytic Gaussian-to-Gaussian KL. The archived summary does not establish whether mixture parameters were fitted jointly, fitted after latent encoding or adapted through another procedure. These alternatives must not be treated as equivalent implementations of the VAE-GMM literature [8].

Figure 4 clarifies a consequential naming convention. Real-only, copula-only and mixed ratios describe the data used to fit the neural generator. They do not describe the data used to train the final FI predictor. A generator fitted to real-only inputs can still be evaluated through TSTR because the downstream predictor is trained on its generated rows. Mixed real:copula ratios of 1:1, 1:2, 1:4 and 1:8 therefore test the propagation of copula-derived inputs through the neural generation stage.

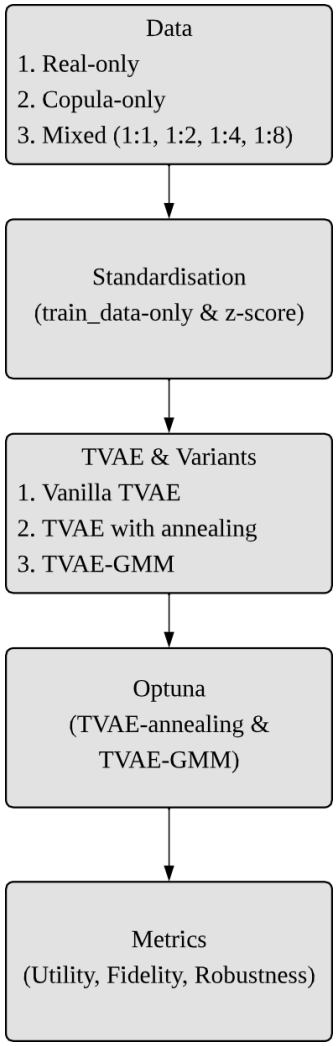


Figure 4. Original workflow for generator-input regimes, standardisation, TVAE variants, optimisation and evaluation. Real-only and mixed labels refer to generator fitting; TSTR refers to the subsequent predictor evaluation. All original steps are retained.

2.6 Utility and distributional metrics

In TSTR, an XGBoost predictor is trained on synthetic records and evaluated on held-out real records. In TRTS, a predictor fitted on real training records is evaluated on synthetic records generated without using the real test set [10]. TRTS diagnoses a different distribution and is not an external-validation substitute. FI is the target; X50 is among the predictors. Because fragment size is normally measured after blasting, this utility task concerns post-blast descriptor inference rather than a pre-blast fragmentation forecast.

The retained prediction metrics are:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \qquad (16)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \qquad (17)$$

RMSE is expressed in the dimensionless FI scale. A negative R^2 indicates squared error exceeding that of the evaluation-set mean in the denominator of Eq. (16); it does not imply negative correlation.

A matched train-on-real-test-on-real predictor evaluated on the same real test rows is needed to determine whether synthesis improves predictive accuracy over direct learning. That comparison is not recoverable from the principal table.

Distributional evaluation uses the unbiased squared maximum mean discrepancy estimator [9]:

$$\begin{aligned}\mathrm{MMD}^2(X,Y) = \frac{1}{n(n-1)}\sum_{i\neq i'} k\left(x_i, x_{i'}\right) \\ + \frac{1}{m(m-1)}\sum_{j\neq j'} k\left(y_j, y_{j'}\right) - \frac{2}{nm}\sum_{i,j} k\left(x_i, y_j\right)\end{aligned} \qquad (18)$$

The kernel is a mixture of radial basis functions:

$$k(a,b) = \sum_r \exp\left(-\frac{\|a-b\|^2}{2\sigma_r^2}\right) \qquad (19)$$

All scaling and bandwidth calibration should be fixed from real training data for comparable evaluation. The finite-sample unbiased estimate of MMD squared can be negative; any clipping, square root or alternative biased estimator must be disclosed. The historical summary labels its numerical values MMD without fully specifying this convention. Its values are retained as benchmark summaries, not interpreted on an absolute universal scale.

The Fréchet autoencoder distance uses the mean and covariance of autoencoder embeddings:

$$\mathrm{FAED} = \| \mu_r - \mu_s \|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_s - 2\left(\Sigma_r^{1/2}\Sigma_s\Sigma_r^{1/2}\right)^{1/2}\right) \qquad (20)$$

This adapts a Gaussian embedding-distance construction used in generative-model evaluation [19]. Training the embedding on real training records prevents direct test fitting, but its architecture, dimension and regularisation still determine which discrepancies it can detect. Mean and covariance agreement in that space does not establish agreement of all higher moments or engineering tails.

2.7 Repetition and reproducibility boundaries

For a metric evaluated over S seeds, the retained summary definitions are:

$$\overline{M} = \frac{1}{S}\sum_{s=1}^{S} M^{(s)}, \qquad \mathrm{SD}(M) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(M^{(s)} - \overline{M}\right)^2} \qquad (21)$$

Repeating synthetic draws while keeping the real split fixed measures conditional Monte Carlo and learner variability. Repeating generator fitting adds optimisation variability. Repeating the real partition addresses another source of uncertainty. These designs are not interchangeable. The principal summary provides approximate means but does not identify S, the complete seed hierarchy or the corresponding standard deviations; none are inferred from decimal precision.

3 Results

3.1 Reported copula fits

The historical comparison reports Clayton AIC per observation of 2.56 and log-likelihood per observation of -1.276, Student-t values of 6.44 and -3.080 with degrees of freedom 8, and Gaussian values of 8.19 and -3.962 with k = 28 [15]. Clayton was selected in the original workflow. These values establish the reported ordering in that implementation, but the schema and likelihood qualifications in Section 2.4 prevent interpreting them as a verified, dimension-matched model-

selection result. No replacement AIC values are calculated without the fitted likelihoods and feature lists.

3.2 Utility of the neural alternatives

With real-only generator inputs, the vanilla model attained TSTR R^2 of approximately 0.434 and RMSE of 2.489. Copula-only generator inputs produced approximately -0.613 and 4.203. For mixed real:copula inputs, reported R^2 values were 0.289, 0.385, 0.253 and -0.359 at ratios 1:1, 1:2, 1:4 and 1:8, respectively. Every mixed result is below the real-only vanilla result, but the sequence is not monotonically decreasing: the 1:2 result exceeds the 1:1 result. This distinction matters when describing a dose-dependent effect of synthetic dominance.

The annealed model attained approximately 0.646 for R^2, 1.967 for RMSE and 0.164 for MMD. The mixture prior model attained approximately 0.736, 1.698, 0.160 and 0.436 for R^2, RMSE, MMD and FAED, respectively. Table 2 preserves these central values and the copula summaries. Missing cells indicate unreported summaries, not zero errors or failed experiments.

Table 2. Principal reported fit and utility summaries. Real-only and copula-only identify generator inputs. Utility values are approximate historical summaries; the copula dimensionality remains unresolved. A dash denotes a value not reported in this table.

Model and regime	Reported fit or protocol	R^2	RMSE	Fidelity
Clayton copula	AIC/obs = 2.56; log L/obs = -1.276; k = 1	-	-	-
Student-t copula	AIC/obs = 6.44; log L/obs = -3.080; v = 8	-	-	-
Gaussian copula	AIC/obs = 8.19; log L/obs = -3.962; k = 28	-	-	-
Vanilla TVAE, real-only	TSTR	≈0.434	≈2.489	-
Vanilla TVAE, copula-only	TSTR	≈-0.613	≈4.203	-
Annealed TVAE, real-only	TSTR	≈0.646	≈1.967	MMD ≈0.164
TVAE-GMM, real-only	TSTR	≈0.736	≈1.698	MMD ≈0.160 FAED ≈0.436

Relative to the vanilla model, the mixture prior result corresponds to an arithmetic RMSE reduction of about 31.8%; relative to the annealed model, the reduction is about 13.7%. These percentages are calculations from the reported rounded values, not new experiments or confidence intervals. The smaller MMD value of 0.160 versus 0.164 is descriptive: without repeated-fit uncertainty and a fixed estimator convention, it does not demonstrate a meaningful fidelity improvement. FAED is reported for the mixture model only in the principal summary, so a numerical FAED advantage over every alternative cannot be established from that table.

3.3 Sampling scale and evidential scope

The project reports sampling 5,000 records from a generator trained on roughly 210 real rows [15]. That count describes the sampled dataset, not the number of independent blast events supporting the fitted distribution. It can reduce the Monte Carlo error of sampling from the estimated generator

while leaving uncertainty about the underlying physical population essentially governed by the real observations. The utility comparison therefore supports a candidate model within this benchmark, rather than a new 5,000-observation field database.

4 Discussion

4.1 Why a mixture prior may help

The performance ordering is compatible with a useful match between latent flexibility and source heterogeneity. A single Gaussian prior can make it difficult to represent separated or curved latent regions without changing the decoder substantially. A mixture can allocate probability to several regions while sharing statistical strength through the same network. This mechanism is consistent with the motivation of VAE-GMM generation [8], but the present results do not prove that individual mixture components correspond to mines, lithologies or fragmentation regimes. Components can overlap, exchange labels or capture measurement conventions rather than geology.

Annealing and mixture priors also change optimisation behaviour and effective regularisation. Differences in architecture, tuning budgets or stopping decisions could contribute to the reported advantage. A controlled comparison would hold these factors constant and contrast the latent-prior change with a parameter-matched alternative. Latent visualisation, per-source fidelity and component-occupancy diagnostics would help distinguish a representation effect from tuning or seed variation. These analyses are proposed tests of the interpretation rather than observations contained in the benchmark.

4.2 Parsimony does not establish dependence adequacy

The compact Clayton model may be less variable to estimate than a full correlation matrix, but parsimony can also impose substantial bias. The eight-variable correlation plot includes both positive and negative relationships. An unrotated positive-parameter Clayton model cannot reproduce that structure, and marginal inverse transforms cannot restore dependencies absent from the copula. A good reported fit summary therefore cannot stand in for a sign-pattern, conditional-dependence or tail-coincidence audit.

Copula-based augmentation has been useful in a geoscientific emulator setting where generated inputs were passed through an independent physical model to obtain outputs [20]. Here, by contrast, feature-target dependence is learned from the same small empirical table. No independent forward model supplies new response information. This difference limits direct transfer of augmentation claims between the two settings and helps explain why plausible marginal samples may offer weak utility for FI inference.

4.3 Synthesis utility and augmentation benefit are distinct

A TSTR R^2 near 0.736 demonstrates that a predictor trained on the generated table can recover part of the variation in the real test target under the reported split. It does not show that the predictor is more accurate than a matched real-only predictor. A generator can achieve useful TSTR by reproducing dominant associations already available to a direct model. Conversely, a synthetic distribution can be useful for software testing or controlled scenario exploration even if point-prediction accuracy does not improve.

The feature set further restricts the operational interpretation. X50 is a post-blast response, and FI may itself be defined through an in-situ-to-fragment-size relationship in some sources. Predicting FI from X50 is a valid retrospective descriptor task if its intended use is explicit. A pre-blast decision-support claim would require a feature set available before firing, a separately defined target and validation at the intended future mine or campaign. The present utility scores must not be transferred to that different task.

4.4 Source overlap and measurement semantics

Combining three named publications does not guarantee independent sampling. If a blast reappears in a later compilation, a random row split can place related representations on both sides of the evaluation boundary. Duplicate detection should include exact rows, unit-converted matches, near-matches within source precision and provenance-based links. Numerical similarity alone is not definitive evidence of duplication, because mines can reuse design settings. Where record identity remains uncertain, grouping plausible matches before splitting provides a conservative sensitivity analysis.

Fragment-size definitions deserve equal attention. The arithmetic mean, mass median and a distribution's characteristic size are distinct quantities. Renaming all of them X50 can introduce a source-specific shift that a flexible generator learns very successfully. Harmonisation should retain original variable definitions and measurement methods, then either derive a justified common estimand or analyse incompatible measurements separately. Without this step, improved fit can reflect better reproduction of reporting differences rather than a more accurate description of fragmentation physics.

4.5 Fidelity, tails and memorisation

MMD is sensitive to kernel bandwidth, sample size and feature scaling. FAED is sensitive to representation learning and compresses distributions to first and second moments in its embedding. Neither metric ensures preservation of rare high-FI cases, physically coherent design ratios or conditional extremes at a particular powder factor. A useful extension would report source-specific quantile errors, joint exceedance frequencies and conditional residuals, while calibrating all diagnostics without adapting them to the real test outcomes.

Nearest-neighbour distances, exact-match counts and real-hold-out reference distances should accompany a small-data generator evaluation. High coverage does not establish originality; high resemblance may coexist with copying [11]. Synthetic release also provides no automatic confidentiality guarantee, as demonstrated in the privacy literature [21]. The open-data origin of this benchmark reduces the role of access restrictions in the comparison but does not remove the scientific need to distinguish new combinations from memorised records.

4.6 Validation needed for engineering use

A stronger experiment would lock an outer split by blast and, where possible, by source or mine; run harmonisation rules, scaling, generator fitting and downstream tuning inside that boundary; and compare identical predictors under real-only, synthetic-only and hybrid training. An inner loop should select generator complexity, sampling multiplier and learner settings without observing outer-test metrics [12-14]. Repeated outer splits would then provide paired performance differences and uncertainty at the real-observation level.

The benchmark should also expose model architecture, latent dimension, mixture size and covariance structure, optimisation settings, annealing schedule, downstream hyperparameters, stopping rules, software versions and seed-level outputs. An auditable eight-variable copula analysis could coexist with a six-variable utility task, provided the transition is explicit and all comparisons use the appropriate dimension. Until that mapping is reconstructed, the safest model conclusion is that the mixture prior is the leading reported neural candidate, rather than a universally preferred fragmentation synthesizer.

5 Conclusions

Within the documented benchmark, the mixture prior VAE attained the highest reported FI transfer utility, while KL annealing improved the vanilla alternative. Copula-derived generator inputs did not improve the reported vanilla result, and their mixed-ratio sequence did not support a monotonic trend. The findings make multimodal latent modelling a defensible candidate for further investigation on heterogeneous fragmentation tables.

The evidence remains conditional on the original split, variable definitions and implementation. The six-versus-eight-variable discrepancy prevents a verified copula AIC comparison, and absent real-only utility and uncertainty summaries prevent a demonstrated augmentation advantage. Future validation should prioritise physical-blast provenance, consistent fragment-size semantics, a predictor feature set appropriate to the intended decision and independent real observations. Synthetic sampling can support that workflow, but cannot substitute for new measurements of geological and operational regimes absent from the training data.

References

[1] Kulatilake, P.H.S.W.; Qiong, W.; Hudaverdi, T.; Kuzu, C. (2010). Mean particle size prediction in rock blast fragmentation using neural networks. Engineering Geology, 114(3-4), 298-311. https://doi.org/10.1016/j.enggeo.2010.05.008

[2] Hudaverdi, T.; Kuzu, C.; Fisne, A. (2012). Investigation of the blast fragmentation using the mean fragment size and fragmentation index. International Journal of Rock Mechanics and Mining Sciences, 56, 136-145. https://doi.org/10.1016/j.ijrmms.2012.07.028

[3] Krop, I.; Sasaoka, T.; Shimada, H.; Hamanaka, A. (2024). Optimizing mean fragment size prediction in rock blasting: A synergistic approach combining clustering, hyperparameter tuning, and data augmentation. Eng, 5(3), 1905-1936. https://doi.org/10.3390/eng5030102

[4] Fonseca, J.; Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. Journal of Big Data, 10, 115. https://doi.org/10.1186/s40537-023-00792-7

[5] Nelsen, R.B. (2006). An Introduction to Copulas. 2nd edition. Springer. https://doi.org/10.1007/0-387-28678-0

[6] Kingma, D.P.; Welling, M. (2014). Auto-Encoding Variational Bayes. International Conference on Learning Representations. https://arxiv.org/abs/1312.6114

[7] Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[8] Apellaniz, P.A.; Parras, J.; Zazo, S. (2024). An improved tabular data generator with VAE-GMM integration. 32nd European Signal Processing Conference, 1886-1890. https://arxiv.org/abs/2404.08434

[9] Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Scholkopf, B.; Smola, A. (2012). A kernel two-sample test. Journal of Machine Learning Research, 13, 723-773. https://www.jmlr.org/papers/v13/gretton12a.html

[10] Esteban, C.; Hyland, S.L.; Ratsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633. https://doi.org/10.48550/arXiv.1706.02633

[11] Alaa, A.; Van Breugel, B.; Saveliev, E.S.; van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. Proceedings of Machine Learning Research, 162, 290-306. https://proceedings.mlr.press/v162/alaa22a.html

[12] Kapoor, S.; Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[13] Roberts, D.R.; Bahn, V.; Ciuti, S.; et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881

[14] Cawley, G.C.; Talbot, N.L.C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079-2107. https://jmlr.org/papers/v11/cawley10a.html

[15] Zhao, X. (2025). MSc Independent Research Project report. Imperial College London. Project identifier irp-xz2821. Experimental source identified in the accompanying manuscript. Repository recorded as https://github.com/ese-ada-lovelace-2024/irp-xz2821

[16] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

[17] Bowman, S.R.; Vilnis, L.; Vinyals, O.; Dai, A.M.; Jozefowicz, R.; Bengio, S. (2016). Generating sentences from a continuous space. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, 10-21. https://doi.org/10.18653/v1/K16-1002

[18] DataCebo. TVAESynthesizer. Synthetic Data Vault documentation. https://docs.sdv.dev/sdv/modeling/single-table-synthesizers/tvaesynthesizer (accessed 28 October 2026).

[19] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.08500

[20] Meyer, D.; Nagler, T.; Hogan, R.J. (2021). Copula-based synthetic data augmentation for machine-learning emulators. Geoscientific Model Development, 14, 5205-5215. https://doi.org/10.5194/gmd-14-5205-2021

[21] Stadler, T.; Oprisanu, B.; Troncoso, C. (2022). Synthetic data - Anonymisation groundhog day. 31st USENIX Security Symposium, 1451-1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler