Public Synthetic-Financial & AML Datasets / Simulators (with Download Links)

All of these datasets are directly applicable to AML, fraud, and transaction-monitoring scenarios — the perfect starting point to experiment with synthetic data and accelerate innovation.

Dataset / Repo	What / Why / Download link
SAML-D (Synthetic Transaction Monitoring Dataset)	Large synthetic AML dataset with ~9.5 million transactions covering 28 typologies (normal + suspicious). Good for transaction-monitoring / AML modeling. Available on Kaggle. (Kaggle)
PaySim / Synthetic Financial Datasets for Fraud Detection	A classic mobile-money / payments simulator generating synthetic transactions + injected fraud. Useful for fraud detection, anomaly detection and model prototyping. (GitHub)
AMLSim (Multi-Agent AML Simulator)	A simulator for AML / banking-transaction simulation. Generates synthetic transaction data (transfers, payments, etc.), enabling design/testing of AML detection models. (GitHub)
IBM AML-Data	Synthetic dataset modelling bank transfers, purchases, credit-card transactions, checks, etc., with a laundering tag for some transactions. Useful for general AML / fraud experiments. (GitHub)
FiFAR Dataset — Fraud Detection + Human-AI "Learning to Defer" Dataset	Synthetic bank-fraud dataset with synthetic "analyst predictions" + fraud labels — useful if you want to prototype hybrid human-AI workflows (fraud alert triage, review, capacity constraints). (arXiv)
MoMTSim Mobile-Money Fraud Dataset	Newly released (2025) synthetic dataset simulating mobile-money ecosystem — deposits, withdrawals, transfers, payments, and explicit fraud labels (account takeover, refund fraud, fake-credential fraud, etc.). Good for mobile money / emerging-markets fraud research. (ScienceDirect)

Key Academic / Industry Papers Using Synthetic Data for Fraud / AML / Transaction Monitoring

Academic and industry papers (2023–2025 and some earlier foundational ones) that use synthetic data for financial-crime, fraud, AML or transaction-monitoring research.

Paper / Study	What it Does / Key Contributions
A synthetic data set to benchmark anti-money laundering methods (2023)	Presents **SynthAML**, a large synthetic dataset with > 16 million transactions and ~20 000 AML alerts — built using the Synthetic Data Vault (SDV) calibrated on real-world bank data. Enables benchmarking of statistical and ML-based AML detection methods. (Nature)
Realistic Synthetic Financial Transactions for Anti-Money Laundering Models (2023)	Introduces an **agent-based synthetic transaction generator** (with publicly released AML datasets) aiming to closely mimic real banking transaction patterns. Demonstrates how synthetic data can support development and comparison of AML detection models, with full ground-truth labeling. (arXiv)
Development of a Synthetic Transaction Monitoring Dataset (2023)	Describes the design and use of **SAML-D**, a synthetic AML transaction dataset — useful for research, statistical analysis and ML experiments on money-laundering detection, without exposing real personal data. (Bournemouth University Research Online)
Improvement of Bank Fraud Detection Through Synthetic Data Generation with Gaussian Noise (2025)	Demonstrates a machine-learning approach to fraud detection using **synthetic data augmentation** (via Gaussian noise) to address class imbalance (rare fraud events). Shows that classical ML models (XGBoost, CNN) outperform naive oversampling (e.g. SMOTE) on fraud classification tasks. (MDPI)
Synthetic Data Applications in Finance (2023)	Provides a broad overview of synthetic-data generation methods in finance (tabular, time-series, event-based), discusses application areas including fraud detection, credit scoring, privacy, regulatory compliance; helps situate AML/fraud use-cases in the wider synthetic-data ecosystem. (arXiv)

Paper / Study	What it Does / Key Contributions
SynDEc: A Synthetic Data Sharing Ecosystem for Financial Institutions (2025)	Explores a synthetic-data sharing ecosystem designed for financial institutions (especially smaller banks) — facilitating collaboration while preserving privacy. Relevant if you contemplate cross-institution data sharing or joint model development under privacy / regulatory constraints. (SpringerLink)

What These Papers Show and Why They Matter for Your Work in Financial Crime

- **Synthetic data + public benchmarking**: Papers like the SynthAML and SAML-D releases solve a major pain point: they allow researchers/practitioners to benchmark AML/fraud models on realistic but non-sensitive data. Useful for comparing different detection techniques or conducting "what-if" experiments without regulatory/ privacy friction.
- **Ground-truth & label completeness**: Real-world suspicious transactions are often **under-detected** (you don't know all the "true positives"). Synthetic datasets let you generate full ground-truth labels — essential for supervised ML (fraud/AML) and for evaluating model performance realistically.
- **Use of ML / advanced techniques**: The Gaussian-noise augmentation study shows synthetic data can improve model robustness in class-imbalanced situations. The "Applications in Finance" paper shows how synthetic data supports not only fraud/AML, but other regulated-finance workflows (credit, scoring, compliance analytics).
- **Enabling smaller institutions / emerging markets**: The synthetic-data sharing framework described in SynDEc suggests a viable path for smaller banks/fintechs (perhaps in MENA region or emerging markets) to collaborate or build AML/fraud-detection capabilities even with limited real data.
- **Flexibility for scenario-based and stress-testing**: Synthetic data generators (agent-based or rule-driven) enable simulation of different typologies, evolving laundering tactics, class imbalance — useful for consulting / advisory scenarios where you want to test detection logic under hypothetical or extreme conditions.