Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FASA: Feature-Agnostic Stacked Autoencoders for Accurate Adverse Drug Reaction Prediction

Martin Gustavo Perez Bonany Torrealva

martin.perezbonany@utec.edu.pe

Universidad de Ingeniería y Tecnología

Edward Jorge Yuri Cayllahua Cahuina

Universidad Católica San Pablo

Rensso Victor Hugo Mora Colque

Universidad de Ingeniería y Tecnología

Research Article

Keywords: Adverse Drug Reactions prediction, Feature Agnostic, Autoencoder, Optimistic Bias

Posted Date: April 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8952090/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FASA: Feature-Agnostic Stacked Autoencoders for Accurate Adverse Drug Reaction Prediction

Martin G. Perez-Bonany Torrealva[1,*], Edward Jorge Y. Cayllahua Cahuina[2,+], and Rensso V. H. Mora-Colque[1,+]

[1]Data Science, UTEC, Medrano Silva 165, Lima, 15063, Peru
[2]Universidad Católica San Pablo, Urb. Campiña Paisajista, Arequipa, 04001, Peru
[*]Corresponding author: martin.perezbonany@utec.edu.pe
[+]Contributing author: ejcayllahua@ucsp.edu.pe
[+]Contributing author: rmora@utec.edu.pe

[1]These authors contributed equally to this work.

Abstract

Purpose: Adverse drug reactions (ADRs) remain a major obstacle to drug safety, yet many computational predictors depend on molecular or biological features that are often unavailable for newly designed compounds. Several published models also report inflated performance due to biased preprocessing or unsuitable evaluation metrics. This work introduces a feature-free deep learning framework that predicts ADRs using only drug–ADR incidence matrices, which allows early-stage assessment even when auxiliary features are missing.

Methods: FASA (Feature-Agnostic Stacked Autoencoders) was developed and trained solely on binary drug-ADR incidence matrices. FASA includes a cardinality-preserving regularization term that constrains reconstructed ADR vectors to follow realistic label-count distributions, preventing degenerate solutions and encouraging the model to learn meaningful structure from sparse data. Performance was evaluated via cross-validation, and the area under the precision-recall curve was reported, as it is well suited to extremely sparse pharmacovigilance data.

Results: On the harmonized WPLMF dataset (1,177 drugs and 4,247 ADRs), the method achieves an AUPR of 0.7150, surpassing all baseline models reported in the original study, including MCS-MKL, FGRMF, IDSE-HE, Galeano, LogitMF and WPLMF, which obtains 0.6553 under the same five-fold protocol. On the raw SIDER benchmark, FASA reaches an AUPR of 0.6456, again outperforming previously published results on the unmodified matrix.

Conclusion: These findings show that carefully regularized deep architectures can recover meaningful pharmacological structure directly from sparse incidence data. FASA offers a straightforward and competitive approach for large-scale ADR prediction using only drug-ADR incidence matrices, without

requiring chemical, biological, or phenotypic features, and generalizes across datasets with varying levels of curation.

Keywords: Adverse Drug Reactions prediction, Feature Agnostic, Autoencoder, Optimistic Bias

1 Background

Adverse drug reactions (ADRs) persist as a formidable obstacle to public health safety worldwide. From an economic standpoint, the strain on healthcare infrastructures is immense: recent analyses estimate that national annual expenditures exceed $30 billion in the United States, $13 billion in Canada, £2.2 billion in the United Kingdom, $460 million in Australia, and €1.3 billion in France [1]. On an individual scale, hospital admissions linked to ADRs in France incur an average cost of approximately €5,974 per patient [2], a figure primarily driven by prolonged hospitalization, which typically extends the length of stay by more than three days. While pre-market clinical trials are designed to identify such risks, they inherently lack the statistical power to capture rare events; detecting a reaction with a prevalence of merely 0.1–0.3% would necessitate cohorts of thousands of participants per arm, rendering exhaustive experimental detection practically impossible [3].

Given these limitations, computational models have become essential complements to experimental pharmacovigilance. Early approaches framed ADR prediction as matrix completion, using linear latent factor models to infer missing associations [4, 5]. Subsequent work extended this paradigm by incorporating auxiliary chemical information through graph regularization [6] or kernel-assisted factorization [7], or multi-kernel learning with soft pseudo-labels [8]. Notable among these is the WPLMF framework [9], which advanced the field by combining matrix factorization with iterative pseudo-labelling and biological graph embeddings to mitigate sparsity. The field has since shifted toward non-linear architectures [10], including Convolutional Neural Networks like DeepSide [11] and Graph Neural Networks that explicitly encode molecular and bipartite structures [12, 13]. Most recently, complex multi-modal frameworks such as HMMF [14], MGPRED [15], and NRFSE [16] have attempted to fuse text, graphs, and attribute similarities to predict not just the existence but the frequency of side effects. While these feature-rich models demonstrate strong predictive capabilities, they introduce significant complexity and dependency on external metadata.

A persistent challenge across these methodologies is the long-tailed distribution of ADRs, shown in Fig. 1a. Similar to drug–drug interactions [17], ADR datasets contain a small set of frequent "head" reactions and a large number of rare but clinically significant "tail" events. Figure 1b shows that 13% of side effects account for 80% of observed drug–ADR associations, while approximately 97% of all possible associations remain unobserved [4]. This extreme imbalance makes ROC-AUC an unreliable metric, as it can mask poor performance on the minority class due to the large number of True Negatives. Precision–recall metrics are therefore preferred, as they more accurately reflect performance on the minority positive class. Polypharmacy further exacerbates the complexity of the task: many dangerous ADRs emerge only when multiple drugs are co-administered, and experimentally evaluating the roughly 692,000 pairwise combinations among SIDER's 1,177 drugs is infeasible.

Many supervised ADR prediction systems including the multi-modal frameworks mentioned above rely heavily on auxiliary drug features such as molecular fingerprints, protein targets, transporters, enzyme

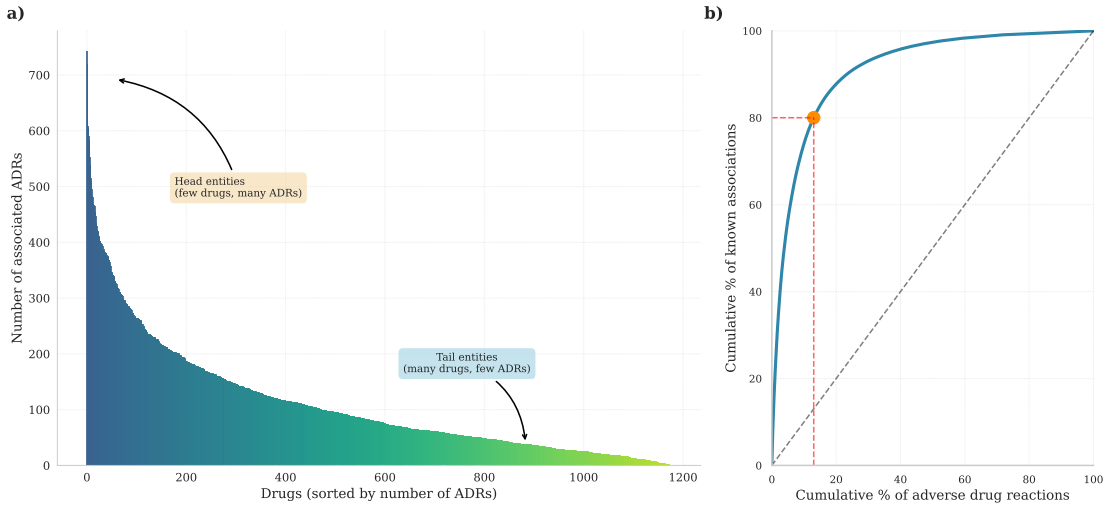


Figure 1: **Long-tailed distribution of ADRs.** (a) Imbalance between frequent "head" reactions and many rare "tail" events. (b) Lorenz curve of ADR frequencies. Roughly 13% of ADRs account for 80% of known drug-ADR associations (orange marker).

profiles, or pathway information. These descriptors are often incomplete, unavailable for newer drugs, and heterogeneous across drug classes. Prior work shows that model performance can degrade substantially when such information is missing [18], and that predictive accuracy typically increases only when increasingly complex feature sets are incorporated [19]. This dependency motivates the evaluation of models that operate solely on the drug–ADR incidence matrix, without requiring any chemical or biological side information.

Motivated by these issues, we restrict our comparisons to models that (i) use publicly available datasets, (ii) follow transparent and reproducible evaluation protocols, and (iii) avoid data leakage or synthetic oversampling. Under these criteria, the WPLMF method provides a well-established and controlled baseline, achieving AUPR scores of 0.65 on its curated dataset and 0.60 on SIDER. Building on this foundation, we introduce a new autoencoder-based model that improves predictive performance without relying on auxiliary biological or chemical features.

2 Materials & Methods

2.1 Dataset

The dataset used by Chen et al., [9] for WPLMF was built by aligning Side Effect Resource (SIDER[1]) database ADR reports with DrugBank[2] [20] identifiers, retaining only drugs present in both sources. ADRs were mapped to MedDRA Preferred Terms, and entries were set to 1 when a drug–ADR association appeared in SIDER. The resulting matrix comprises 1,177 drugs and 4,247 ADRs, with a positive rate of

Table 1: Summary statistics for the benchmark datasets used in this work.

Dataset	Drugs	ADRs	Associations	Known-Associations
SIDER4 ∩ DrugBank	1177	4247	4,998,719	2.7%
SIDER4	1080	5579	6,025,320	2.0%

2.7%. As reported in Fig. 1, ADR frequencies follow a heavy-tailed distribution, where a small number of drugs contribute a disproportionate fraction of associations. The SIDER4 [21] dataset was used in its original form, reporting ADRs at the MedDRA Lowest Level Term. This produces a larger ADR vocabulary of 5,579 terms and an estimated positive rate of approximately 2%, resulting in a sparser and more challenging matrix-completion task.

To ensure that performance reflects structure learned directly from incidence patterns, no DrugBank-derived metadata (e.g., targets, enzymes, transporters, pathways, indications, carriers) were used. These fields are reported only in the Supplementary Materials for completeness. Table 1 summarizes the dataset dimensions used throughout the experiments.

2.2 Pseudo-labels

Extreme sparsity in drug–ADR matrices complicates learning because unobserved entries correspond to unknown rather than true negatives. Treating all missing entries as negatives introduces bias, while synthetic balancing distorts the naturally heavy-tailed distribution. To mitigate these issues, we employ a lightweight pseudo-labeling mechanism inspired by self-training [22] and positive–unlabeled learning [23]. The goal is not to densify the matrix substantially, but to provide a small number of high-confidence positives that stabilize early training.

Unlike standard self-training, which risks confirmation bias by adding too many predicted labels, our approach uses a conservative selection strategy. At each round, only the most confident predictions among previously unobserved entries are promoted, and a shallow retraining step updates the model accordingly. This restrained design aligns with PU-learning principles by ensuring that only the most reliable pseudo-positives influence the latent space.

[1] SIDER database: https://sideeffects.embl.de/
[2] DrugBank database: https://go.drugbank.com

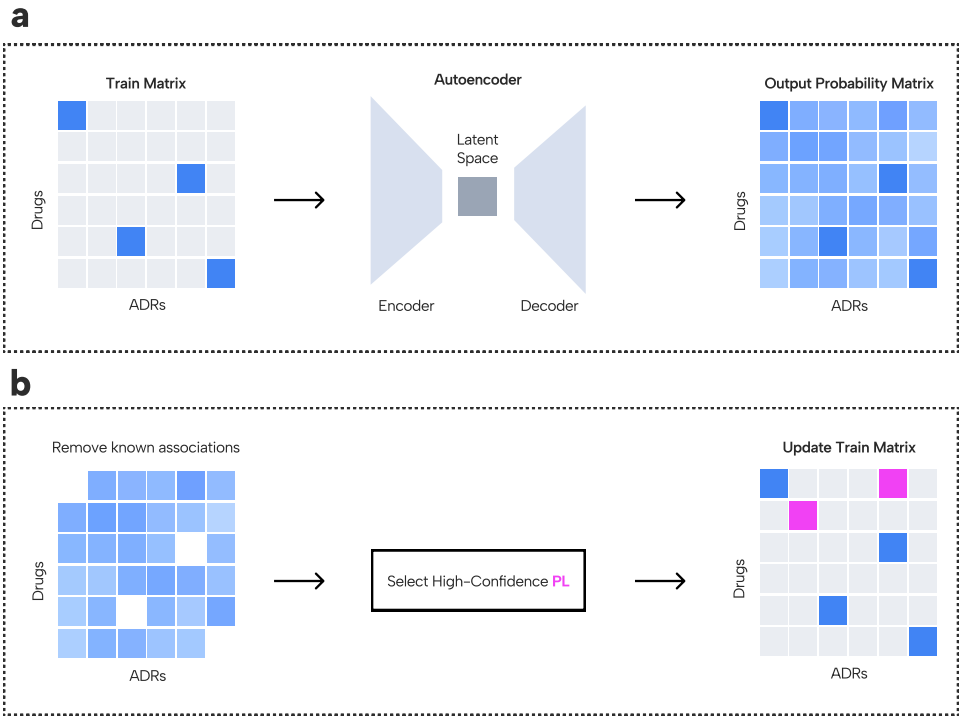


Figure 2: **Pseudo labeling framework.** (a) Illustrates the initial prediction stage where the autoencoder receives the sparse drug–ADR incidence matrix (blue indicates known positives, grey indicates zeros) and outputs a probability matrix with graded intensities of blue. (b) Shows the pseudo label extraction step where known associations are removed, high confidence predictions are selected, and the augmented training matrix is formed by injecting pseudo labels (purple) into previously unobserved entries.

A schematic of the workflow is shown in Fig. 2. At iteration t, the model is trained on the current augmented matrix and produces reconstruction scores for all unobserved entries. A small subset of top-scoring candidates is added as positives for the next iteration, after which the model is retrained. The loop stops once performance on a held-out set no longer improves, ensuring that augmentation remains minimal and does not accumulate noise.

Full algorithmic details, thresholding rules, percentile schedules, and exact stopping logic are provided in the Supplementary Material.

2.3 Proposed Model

The proposed model adopts a deep autoencoder architecture (Fig. 3) to reconstruct drug–ADR incidence vectors $M_i. \in \{0,1\}^{n_{ADRs}}$ from highly sparse binary inputs. Unlike shallow matrix factorization approaches that impose linear low-rank structure, the autoencoder captures non-linear pharmacological

relationships and higher-order co-occurrence patterns. The model transforms the input M_i into a latent representation $z_i \in \mathbb{R}^{d_{Latent}}$, from which it generates a reconstructed vector of probabilities $\hat{M}_i. \in [0,1]^{n_{ADRs}}$.

The encoder consists of two fully connected layers with 2048 and 1536 units, respectively, each followed by batch normalization, LeakyReLU activation ($\alpha = 0.1$), and dropout (rate 0.2). During early experimentation, the first encoder layer was configured with two consecutive LeakyReLU activations ($\alpha = 0.1$) separated by dropout. While initially unintentional, we observed that this configuration consistently improved reconstruction performance compared to a standard single activation. Mathematically, two consecutive LeakyReLU activations ($\alpha = 0.1$), with dropout applied between them, yield an effective negative slope of $\alpha^2 \approx 0.01$ for negative inputs (0.1×0.1) while preserving unit slope for positive inputs. We hypothesize that this aggressive suppression of negative activations is beneficial for learning from high-dimensional, sparse binary inputs M_i, as it strictly limits the propagation of noise in the inactive regions of the latent space. Consequently, this configuration was retained. Layer widths were selected based on a preliminary ablation study: smaller layers ($512 - 1024$ units) consistently underfit the long-tailed ADR label distributions, whereas larger layers yielded negligible performance gains at substantially higher computational cost. All layers used He-normal initialization to improve optimization stability under sparse activations.

The encoder compresses the M_i into a latent vector z_i of dimension d_{latent}. A systematic search over $d_{\text{latent}} \in \{64, 128, 256, 384, 512, 1024, 2048, 4096\}$ revealed that performance increases steadily up to $d_{latent} \approx 1024$, after which it stabilises. There was no meaningful improvement between 1024 and 2048, indicating that the ADR incidence matrix possesses substantial intrinsic dimensionality and cannot be captured by the low-rank embeddings typically assumed by matrix-factorization models. The high-capacity latent space allows the autoencoder to model complex, nonlinear incidence structure while remaining compact enough for the decoder to reconstruct the full ADR profile.

The decoder mirrors the encoder with layers of 1536 and 2048 units, preserving architectural symmetry and facilitating stable optimization. The output layer contains n_{ADRs} sigmoid units that produce reconstruction probabilities $\hat{M}_i. \in [0,1]^{n_{\text{ADRs}}}$, where each element $\hat{M}_i j$ represents the probability of the j-th ADR being associated with drug i. The output layer biases are initialized as described in Sect. 2.4. The model is trained with the composite loss described in Sect. 2.5, combining binary cross-entropy with a cardinality-preserving penalty. This objective regularizes the autoencoder toward ADR vectors whose total number of predicted reactions matches empirical distributions, reducing the tendency of purely BCE-trained models to collapse toward uniformly sparse or dense reconstructions.

2.4 Output-layer bias initialization

The extreme sparsity of drug–ADR incidence matrices poses a challenge for stable optimization in binary reconstruction models. To ensure that the autoencoder begins training with realistic output magnitudes, the bias of each sigmoid unit in the decoder was initialized to match the global ADR prevalence. Specifically,

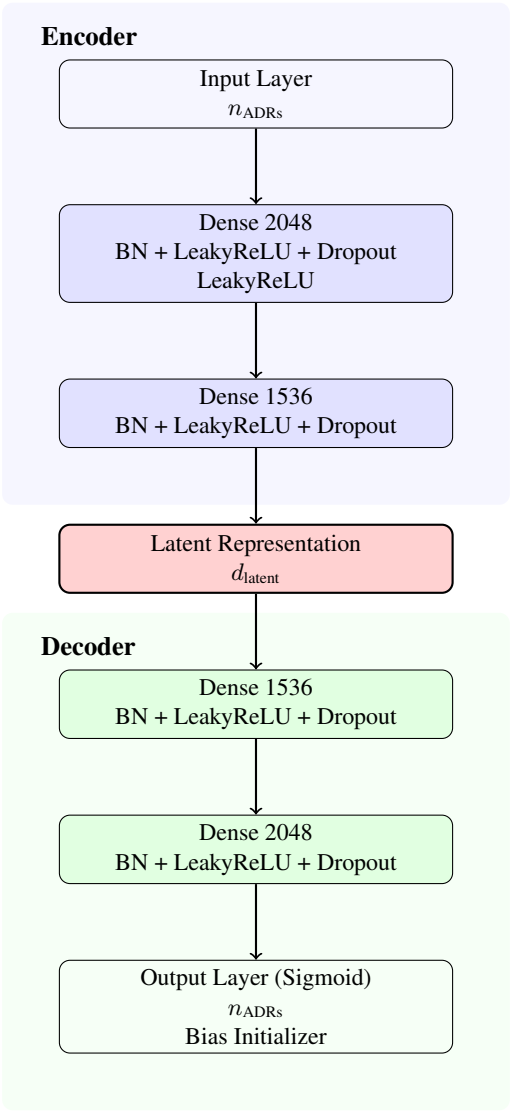


Figure 3: Architecture of the proposed deep autoencoder for drug–ADR matrix reconstruction. Hidden layers use LeakyReLU activation with $\alpha = 0.1$, batch normalization for stable optimization, and dropout (0.2) for regularization. The latent layer compresses the ADR incidence vector into a dense, low-dimensional pharmacological representation.

let p denote the proportion of positive entries in the incidence matrix. Setting the output bias to

$$b_0 = \log\left(\frac{p}{1-p}\right)$$

yields initial predictions $\sigma(b_0) = p$, aligning the model's prior with the empirical sparsity level. In our dataset, $p = 0.026$, resulting in a bias value of $b_0 = -3.61$. This initialization prevents the network from producing implausibly dense ADR profiles at the start of training, improves gradient flow for rare positive entries, and accelerates convergence under highly imbalanced conditions. We found empirically that this bias strategy consistently improved AUPR and reduced early-training instability compared to zero-initialized outputs.

2.5 Proposed Loss: Structure-Aware Reconstruction Objective

Reconstructing a drug–ADR matrix requires capturing not only entry-wise correctness but also the global structure of adverse reaction profiles. Standard binary cross-entropy (BCE) optimizes each matrix entry independently and therefore ignores the fact that drugs exhibit characteristics such as the number of associated ADRs. Because ADR counts follow a highly skewed long-tailed distribution as seen in Fig. 1, an objective based solely on BCE is under-determined: many prediction matrices achieve similar element-wise likelihoods while differing substantially in total ADR counts per drug. Empirically, this leads autoencoders to over-predict ADRs for rare drugs and under-predict for common ones, even when the local reconstructions are accurate.

To address this, we introduce a composite objective which augments BCE with a cardinality-matching term. This connects directly to label cardinality regularization in multi-label learning, where matching empirical label counts is known to reduce degenerate solutions and improve calibration. Let $M_i.$ denote a ground-truth ADR incidence vector and $\hat{M}_i$ its reconstruction. Define

$$C_{\text{true}} = \sum_{j=1}^{n_{ADRs}} M_{ij}, \qquad C_{\text{pred}} = \sum_{j=1}^{n_{ADRs}} \hat{M}_{ij}.$$

The cardinality term penalizes deviations between C_{true} and C_{pred} via,

$$\mathcal{L}_{\text{card}} = (C_{\text{true}} - C_{\text{pred}})^2.$$

The complete loss is

$$\text{Loss} = \mathcal{L}_{\text{BCE}} + \lambda\, \mathcal{L}_{\text{card}},$$

with λ controlling the strength of the constraint. After combining the two components, the objective for a single drug–ADR vector becomes:

$$\text{Loss} = -\sum_{j=1}^{n_{ADRs}} \left[M_{ij} log \hat{M}_{ij} + (1 - M_{ij}) \log(1 - \hat{M}_{ij}) \right] + \lambda \left(\sum_{j=1}^{n_{ADRs}} M_{ij} - \sum_{j=1}^{n_{ADRs}} \hat{M}_{ij} \right)^2.$$

Conceptually, this loss imposes a weak moment-matching condition: while BCE governs individual

entries, $\mathcal{L}_{\text{card}}$ encourages the predicted cardinality distribution to approximate the empirical one, preventing the model from converging to uniformly sparse or uniformly dense regimes. This is particularly relevant under long-tailed ADR frequencies, where the reconstruction problem admits many BCE-equivalent solutions. From a distributional perspective, the loss reduces the divergence between the predicted and observed cardinality histograms; a quantitative comparison and corresponding histograms illustrating the effect of this loss on predicted ADR cardinalities are presented in Sect. 3.

2.6 Stacked Ensemble & Meta Learning

Autoencoders with different latent dimensionalities capture complementary structure: smaller bottlenecks emphasize global, low-rank patterns, whereas larger ones preserve fine-grained local variations. Because ADR incidence matrices exhibit both broad co-occurrence trends and drug or class specific subtleties, combining models across this compression spectrum is beneficial. Regularization strength further modulates these representations, producing a diverse set of reconstructions even under identical training splits.

To aggregate these heterogeneous predictors, we adopt a two-stage stacked ensemble summarized in Fig. 4. In the first stage, each autoencoder variant is trained under the outer five-fold split and produces out-of-fold predictions, ensuring that all meta-features are strictly out-of-sample. These predictions form a low-dimensional table in which each column corresponds to a model's reconstruction scores.

Ridge regression is used as the meta-learner because (i) the meta-features are highly collinear-autoencoders share architecture and data-and ridge's ℓ_2 penalty stabilises weight estimation; (ii) its linear form makes the relative contribution of each latent size interpretable; and (iii) it avoids overfitting in settings where the number of meta-features is small but correlated. A second layer of five-fold cross-validation is applied on the meta-features to isolate the stability of the combination rule from the variability of the base models. After cross-validation, the meta-learner is retrained on all meta-features and produces the final ensemble predictions for each fold. This preserves the same cross-validation structure as the base models and is compatible with threshold-free metrics such as AUPR.

Empirically, we observed that while individual autoencoders exhibited performance fluctuations depending on initialization and latent dimensionality, the stacked ensemble consistently yielded higher AUPR scores than any single constituent model. By aggregating predictions across the compression spectrum, the meta-learner effectively smooths out model-specific noise. Importantly, the learned ridge coefficients did not collapse onto a single model; instead, weights were distributed across autoencoders with different latent sizes (e.g., 1024 and 2048), indicating that the ensemble benefits from capturing structure at multiple levels of abstraction rather than relying on a single dominant representation.

2.7 Model Evaluation

The prediction task is treated as a binary classification problem on a sparse drug–ADR incidence matrix, where the goal is to determine whether a drug produces a particular adverse reaction. Because the proportion of positive entries is extremely small, evaluation focuses on metrics that are informative under

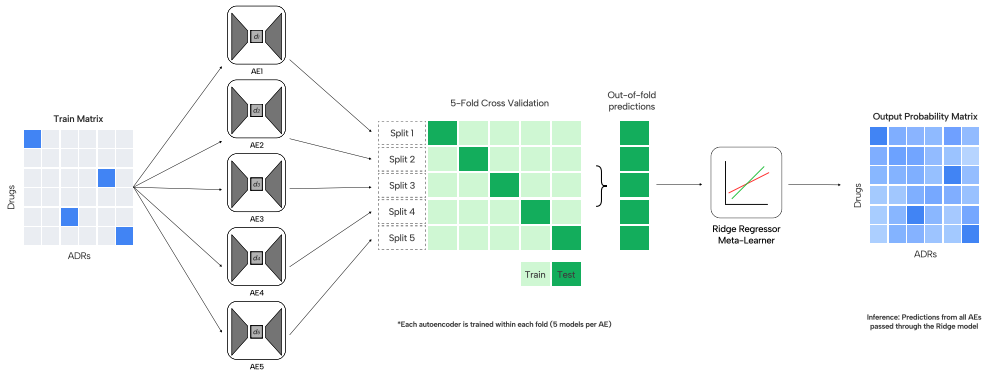


Figure 4: Workflow of the stacked ensemble. Multiple autoencoders with different latent dimensions generate out-of-fold (OOF) predictions under 5-fold cross-validation. The OOF outputs are stacked to form a meta-feature matrix, which is used to train a Ridge meta-learner. The resulting model combines base predictions into a final ensemble reconstruction.

heavy class imbalance. The primary metric is the Area Under the Precision–Recall Curve (AUPR). Let $P(r)$ and $R(r)$ denote precision and recall at threshold r. AUPR is defined as

$$\mathrm{AUPR} = \int_0^1 P(r)\, dR(r),$$

which reflects the ability of the model to recover true ADRs while minimizing false positives across relevant thresholds.

AUPR is preferred over ROC-AUC for highly imbalanced drug–ADR data. In sparse settings, the number of true negatives is orders of magnitude larger than the number of positives, causing ROC curves to remain artificially high even for trivial classifiers. For example, predicting all zeros can yield a ROC-AUC close to 0.90 on matrices where positive entries represent less than 1% of observations. ROC-AUC therefore primarily measures the model's ability to rank negatives which is uninformative in long-tailed ADR prediction. In contrast, precision–recall curves are sensitive to the minority positive class and directly quantify the tradeoff between false discoveries and missing true ADRs. As a result, AUPR provides a more faithful assessment of real-world predictive performance.

Several secondary metrics are also reported. (i) F1-score measures the harmonic mean of precision and recall and summarizes classification performance at a chosen threshold. (ii) Precision quantifies the proportion of predicted positive ADRs that are correct, while Recall measures the proportion of true ADRs that are successfully recovered. (iii) Precision@15 and Recall@15 assess the correctness of the top 15 predicted ADRs per drug and are useful for practical ranking scenarios.

3 Results

3.1 Experimental Setup

All experiments were conducted in a controlled computational environment to ensure full reproducibility and fair comparison between models. Training and evaluation were performed on a workstation running Windows Subsystem for Linux (WSL2) with Python 3.10 and TensorFlow with CUDA support. The system was equipped with an NVIDIA RTX 3060 GPU with 12 GB of VRAM, an AMD Ryzen 7 3600X CPU, and 32 GB of RAM. GPU acceleration was used for all deep learning experiments.

Model evaluation followed a five-fold cross-validation protocol. For each fold, the training matrix was partitioned into disjoint train and validation subsets, after which every autoencoder in the ensemble was trained on the corresponding training split and produced out-of-fold predictions for the held-out split. The collection of out-of-fold predictions across all five folds formed the input to the meta-learning stage. The regularization coefficient of the cardinality-aware loss function was determined through a grid search. The optimal value was $\lambda = 10^{-6}$, which yielded the best balance between stability and reconstruction quality. The decoder bias was initialized according to the method described in Section 2.4. The final ensemble consisted of four autoencoders with latent dimensions of 2048, 1024, 2048, and 1024, respectively. The dropout rate was set to 0.2, which consistently provided superior generalization, and the negative slope of the LeakyReLU activations was fixed at 0.1. Gradients were clipped at a maximum norm of 1.0, which improved training stability and resulted in higher precision–recall performance. The Adam optimizer with learning rate 10^{-3} was used in all experiments, together with a batch size of 512 and a training horizon of 50 epochs.

All model components, including the loss function, dropout configuration, activation parameters, optimizer, bias initialization, gradient clipping strategy, and training schedule, were kept fixed across the ensemble. Only the latent dimension varied between autoencoders, following the four selected configurations. For completeness, Table 2 summarises the full set of hyperparameters used for both the autoencoders and the baseline methods.

3.2 Performance Analysis and Comparison

To contextualize the predictive behavior of the proposed autoencoder, we compare its performance with the full suite of baseline methods introduced in the WPLMF study. These methods rely on auxiliary molecular descriptors, graph-based drug representations or curated similarity matrices, whereas the proposed model operates exclusively on the binary drug–ADR incidence matrix. A brief summary of each baseline is provided below.

(i) FGRMF [6] incorporates drug features into a graph-regularised matrix factorisation model. A similarity graph is constructed from molecular descriptors, and the factorisation is constrained to preserve this graph structure, allowing the method to exploit external chemical information.

(ii) IDSE-HE [12] employs a graph neural network to encode drug representations from multiple

Table 2: Detailed settings for the baseline methods (adapted from [9]).

Method	**Features**	**Setting**
MSC-MKL	CS, DAA	$\mu = 2e-3, \quad v = 2e-5, \quad k = 17$
FGRMF	CS, DAA, DPI	$\mu = 8, \quad \lambda = 4$
idse-HE	CS, DAA	$\text{lr} = 1e-3, \quad \text{dropout} = 0.1, \quad \alpha = 0.02, \quad \lambda = 5e-3$
Galeano's	DAA	$\text{lr} = 1e-2, \quad \lambda = 15$
LogitMF	CS, DAA	$\text{lr} = 1e-2, \quad \lambda = 1e-4, \quad \alpha = 1, \quad \beta = 0.2$
WPLMF	DAA, DPI	$\text{lr} = 1e-2, \quad p = [0.5, 0.7, 1, 1.5, 2], \quad q = [2, 1.5, 1, 0.7, 0.5], \quad \alpha = 0.9, \quad \gamma = 2, \quad$ threshold = 0.8
FASA	DAA	$\text{lr} = 1e-3, \quad d_{\text{Latent}} \in \{1024, 2048\}, \quad \text{dropout} = 0.2, \quad \text{threshold} = 0.0652, \quad b_0 = -3.61$

CS = Chemical Structure; DAA = Drug–ADR Association; DPI = Drug–Protein Interaction

molecular views. The method integrates MPNN + Set2Set embeddings together with fingerprint-based features, producing a rich, feature-driven latent space before modelling ADR associations.

(iii) Galeano [24] projects the drug–ADR matrix into a low-dimensional latent space via linear matrix factorisation. The method captures co-occurrence structure but does not model non-linear dependencies and is constrained to bilinear relationships by construction.

(iv) LogitMF [7] applies Kernel PCA to reduce high-dimensional molecular fingerprints (ECFP; 2048 dimensions) to a 100-dimensional representation. These compressed structural features are then fed into a logistic matrix factorisation model. The overall performance depends heavily on the quality of the external chemical descriptors.

(v) MSC-MKL [8] enhances matrix factorisation by performing soft pseudo-labelling using Weighted k-Nearest Neighbours and combining multiple drug similarity matrices through multi-kernel learning. As with the previous baselines, its predictive capacity arises from auxiliary drug–drug similarity information rather than the incidence matrix itself.

(vi) WPLMF [9] formulates ADR prediction as a matrix completion task by factorising the drug–ADR matrix into low-dimensional latent representations. The method introduces a weighted pseudo-labelling scheme that iteratively updates the training matrix, allowing the model to mitigate sparsity during optimisation. Although WPLMF operates on the binary incidence matrix, its training procedure incorporates several external components, including node2vec embeddings derived from drug–protein interaction networks, frequency-based weighting strategies for ADRs, and an ensemble mechanism designed to stabilise the pseudo-label updates. These elements provide structured biological information that complements the incidence matrix and contribute substantially to its predictive performance. Despite these additions, WPLMF remains the closest baseline conceptually, as it directly addresses the sparsity of the drug–ADR matrix through a pseudo-labelling framework.

All baseline methods therefore rely on external molecular or similarity features and constrain learning to low-dimensional embeddings typically below 200 latent units. By contrast, our approach operates directly on the binary incidence matrix alone, without incorporating any chemical, biological, or phenotypic metadata. Furthermore, the autoencoder learns high-dimensional non-linear representations of 1024–2048 units per layer, enabling the model to capture complex, high-capacity structure in the incidence patterns that is inaccessible to shallow or feature-dependent factorization methods. This comparison highlights a key conceptual distinction: whereas existing ADR predictors rely on curated drug features and dimensionality reduction, the proposed model reconstructs ADR profiles solely from observed incidence patterns, demonstrating that rich pharmacovigilance structure can be recovered without auxiliary information. The quantitative results for the WPLMF and SIDER datasets are presented in Table 3 and Table 4.

Table 3: Performance comparison of the proposed autoencoder model against baseline methods on the WPLMF dataset. Best values are shaded blue and worst values red.

Method	**AUPR**	**F1**	**P@15**	**R@15**	**Precision**	**Recall**
MCS-MKL	0.6428 (1.3e-3)	0.6082 (1.2e-3)	0.6567 (7.6e-3)	0.7813 (1.9e-2)	0.6145 (1.2e-2)	0.6082 (1.1e-2)
FGRMF	0.5117 (4.5e-3)	0.4993 (2.6e-3)	0.4932 (5.9e-3)	0.7842 (1.3e-2)	0.4953 (1.4e-2)	0.5043 (1.5e-2)
idse-HE	0.5888 (1.1e-2)	0.5564 (1.1e-2)	0.5676 (8.5e-3)	**0.9136** (1.3e-2)	0.4738 (3.5e-2)	**0.6804** (3.9e-2)
Galeano's	0.6294 (2.8e-3)	0.5924 (1.9e-3)	0.6081 (1.3e-2)	0.7748 (1.1e-2)	0.6049 (1.9e-2)	0.5804 (1.2e-2)
LogitMF	0.6124 (3.3e-3)	0.5766 (3.5e-3)	0.6370 (6.6e-3)	0.7263 (5.0e-3)	0.5819 (1.3e-2)	0.5714 (1.6e-3)
WPLMF	0.6553 (3.0e-3)	0.6095 (1.8e-3)	0.6440 (11.8e-2)	0.77981 (5.5e-3)	0.6172 (1.4e-2)	0.6024 (1.1e-2)
Proposed Model	**0.7150** (4e-4)	**0.6621** (6e-4)	**0.8481** (35e-4)	0.2118 (11e-4)	**0.6934** (18e-4)	0.6335 (16e-4)

WPLMF dataset. Across all metrics evaluated on the WPLMF dataset, the proposed autoencoder achieves the highest scores in AUPR, F1, P@15 and Precision. The gain in AUPR is substantial, rising from 0.6553 for WPLMF to 0.7150 for the proposed model. This improvement is consistent with the objective of the method, since AUPR is driven by the quality of calibrated probability estimates under severe class imbalance. The cardinality-preserving regularisation term encourages the decoder to reproduce

Table 4: Performance comparison of the proposed autoencoder model against baseline methods on the SIDER dataset. Best values are shaded blue and worst values red.

Method	AUPR	F1	P@15	R@15	Precision	Recall
MCS-MKL	0.5747 (3.6e-3)	0.5624 (2.1e-3)	0.6061 (1.1e-2)	0.7741 (2.0e-2)	0.5693 (1.2e-2)	0.5590 (1.7e-2)
FGRMF	0.5434 (4.4e-3)	0.5347 (3.4e-3)	0.7304 (1.0e-2)	0.5416 (6.7e-3)	0.5347 (9.2e-3)	0.5351 (1.1e-2)
IDSE-HE	0.5303 (7.7e-3)	0.5069 (2.6e-2)	0.5228 (1.9e-2)	**0.9026** (1.3e-2)	0.4267 (5.8e-2)	**0.6402** (4.5e-2)
Galeano's	0.5698 (3.6e-3)	0.5457 (7.1e-3)	0.7500 (1.3e-2)	0.5567 (2.3e-2)	0.5617 (9.4e-3)	0.5306 (1.3e-2)
LogitMF	0.5479 (4.9e-3)	0.5402 (4.1e-3)	0.7324 (1.3e-2)	0.6135 (6.7e-3)	0.5379 (1.8e-2)	0.5425 (2.1e-2)
WPLMF	0.6031 (3.8e-3)	0.5744 (2.9e-3)	0.6100 (7.5e-3)	0.7816 (8.2e-3)	0.5858 (9.4e-3)	0.5635 (4.3e-3)
Proposed Model	**0.6456** (8e-4)	**0.6076** (1e-3)	**0.8145** (21e-4)	0.2004 (8e-4)	**0.6445** (45e-4)	0.5747 (35e-4)

realistic ADR label counts, which in turn stabilises probability mass allocation across positive and negative classes and contributes to higher precision at common evaluation thresholds.

The model also achieves the highest P@15 (0.8481), indicating that its top-ranked predictions are considerably more reliable than those of the baselines. This behaviour aligns with the ensemble architecture, where multiple autoencoders provide complementary reconstruction patterns that the meta-learner integrates into a sharper consensus. The resulting probability matrix concentrates its highest confidence on ADRs with consistent support across models, which increases both precision and P@k performance.

In contrast, the method exhibits the lowest R@15 (0.2118). This pattern is expected: a model that prioritises sparse, high-confidence predictions will retrieve fewer ADRs when only the top fifteen positions are considered. Several baselines achieve higher recall at this cutoff by producing broader, noisier score distributions. However, this comes at the cost of lower precision and reduced AUPR. In this regime, the proposed model behaves conservatively, favouring correctness over coverage in the early ranks while still maintaining competitive full-curve Recall and the highest overall F1 score.

Taken together, the WPLMF results show that the model concentrates probability mass more effectively and suppresses false positives more aggressively than feature-driven baselines, leading to a superior balance between precision and recall over the entire precision–recall curve.

SIDER dataset. A similar pattern is observed in the SIDER benchmark. The proposed model again obtains the highest AUPR (0.6456), F1, P@15 and Precision, while yielding the lowest R@15. The relative improvements over the baselines are smaller than in the WPLMF dataset, which is consistent

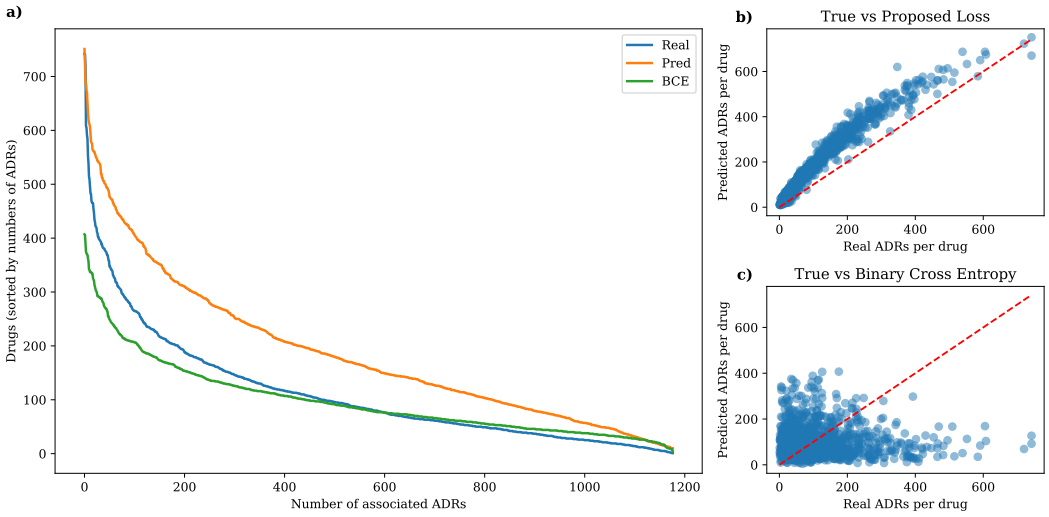


Figure 5: **Prediction structure analysis.** (a) Number of associated ADRs per drug for the ground truth, the proposed cardinality-aware model, and a BCE baseline. The BCE model underestimates high-cardinality drugs, whereas the cardinality-aware loss recovers the correct scale. (b) Predicted vs. true ADR counts using the cardinality-aware loss, showing close alignment with the $x = y$ diagonal. (c) Same scatter plot for the BCE baseline, illustrating its systematic under-prediction and loss of global cardinality structure.

with the increased noisiness and reduced curation of the SIDER matrix. Despite this, the model retains a clear advantage in AUPR, which reflects the stability of its probability estimates under different class distributions and sparsity levels.

The strong Precision and P@15 values (0.6445 and 0.8145) indicate that the high-confidence region of the score space remains well calibrated even when the underlying incidence matrix contains more heterogeneous noise patterns. The meta-learning stage is particularly influential in this setting, since the base models produce partially divergent reconstructions in the low-signal regions of SIDER, and the ridge regression layer effectively suppresses inconsistent contributions.

The low R@15 is again a consequence of conservative ranking behaviour. Because the model avoids distributing probability mass widely across ADRs with weak evidence, it retrieves fewer positive labels at very small cutoffs. However, its full-curve Recall remains competitive, demonstrating that the model is not simply underpredicting but is instead reallocating probability more selectively, which benefits AUPR and F1.

3.3 Reconstruction Behaviour and Cardinality Structure

Beyond aggregate metrics, it is essential to evaluate whether the autoencoder correctly captures the global structure of ADR incidence across drugs. Figure 5 provides a detailed characterisation of this behaviour and highlights clear differences between the proposed approach and a model trained solely with binary

cross-entropy (BCE).

Panel **(a)** shows the number of associated ADRs per drug, sorted from most to least annotated. The ground-truth curve follows a heavy-tailed distribution, with a small set of drugs exhibiting extremely high ADR cardinality. The BCE model fails to reconstruct this head region and saturates at 407 ADRs, substantially below the true maximum 742. This confirms that BCE optimisation alone cannot recover global cardinality structure in highly sparse binary matrices, leading to systematic underestimation of ADR-rich drugs. In contrast, our cardinality-aware loss successfully matches the true scale of the head distribution, reaching more than 740 predicted ADRs for the drugs with highest incidence. This demonstrates that the additional cardinality term enables the model to learn dataset-level structural properties that are invisible to pointwise cross-entropy.

Panels **(b)** and **(c)** further quantify these effects by comparing predicted versus true ADR counts per drug. Under the cardinality-aware loss (panel b), predictions align closely with the $x = y$ diagonal, indicating that the model reconstructs the overall number of ADRs per drug with high fidelity. Points lie slightly above the diagonal, reflecting the mild over-prediction intentionally encouraged for downstream case-study analysis, where high-recall candidates are prioritised before manual review. When trained with BCE alone (panel c), the model exhibits chaotic cardinality behaviour. The scatter collapses into a dense mass near the lower-left corner, revealing severe underestimation across nearly all cardinality ranges. Drugs with more than 400 real ADRs receive fewer than 100 predictions, while others with fewer than 100 true ADRs are assigned more than 400. This inconsistency is consistent with the curve observed in panel (a), and further confirms that BCE is insufficient for reconstructing the long-tailed incidence structure of pharmacovigilance data.

To formally quantify the distributional similarity between the predicted and ground-truth ADR cardinalities, we computed the Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences. The proposed model achieved a KL divergence of 0.0247 and a JS divergence of 0.0062, indicating a near-perfect alignment with the empirical distribution. In contrast, the BCE-only baseline yielded significantly higher divergence values (KL: 0.6215; JS: 0.1517). These metrics numerically confirm that the cardinality-aware loss effectively forces the predicted density to match the underlying data manifold, whereas standard BCE allows the reconstruction distribution to drift substantially from the true prior.

Overall, these structural analyses indicate that the proposed cardinality-aware training procedure is critical for enabling the autoencoder to model not only local ADR-level reconstruction accuracy but also global dataset properties such as drug-level incidence magnitude. This behaviour is not captured by standard BCE training, and it is a key contributor to the superior precision–recall performance observed in Tables 3 and 4.

3.4 Ablation Studies

An ablation study was conducted to quantify the contribution of each architectural and training component of the proposed FASA model, with the goal of isolating the effect of activation functions, regularisation

Table 5: Ablation study of the proposed FASA model. Best values are highlighted in bold.

Ablation Variant	**AUPR**	**F1**	**ΔAUPR**
No negative output bias	0.1709	0.2346	-0.5425
ReLU instead of LeakyReLU(0.1)	0.3555	0.4114	-0.3579
No dropout ($d = 0$)	0.6828	0.6325	-0.0306
No pseudo-labels	0.6867	0.6373	-0.0267
Single autoencoder (no ensemble)	0.6979	0.6480	-0.0155
No Second LeakyReLU(0.1) on first encoder layer	0.6985	0.6528	-0.0149
No cardinality loss, used BCE instead	0.7069	0.6534	-0.0065
Latent dims reduced: $d_{\text{Latent}} = \{512, 384\}$	0.7091	0.6552	-0.0043
Full Model (FASA)	**0.7134**	**0.6617**	–

strategies, weak supervision, latent dimensionality, ensemble averaging and the negative output bias term. The results, summarised in Table 5, report the change in AUPR relative to the full model.

The most influential component is the negative output bias: removing it causes the model to collapse to its lowest AUPR (0.1709) and F1 score (0.2346), which confirms that the bias is essential for calibrating output logits and preventing the uncontrolled overprediction observed during preliminary experiments. Activation design also plays a major role. Replacing LeakyReLU(0.1) with standard ReLU substantially reduces performance, indicating that a small negative slope is important for maintaining gradient flow throughout the extremely sparse ADR incidence space. Regularisation and weak supervision have a more modest but still meaningful impact. Eliminating dropout or pseudo-labels produces moderate decreases in AUPR, showing that both mechanisms help reduce overfitting and stabilise reconstruction, although neither is individually indispensable for competitive performance. Removing the ensemble and training a single autoencoder results in a mild performance decline, suggesting that while variance reduction is beneficial, the latent representation learned by FASA is stable enough for a single model to remain strong. Likewise, omitting the additional LeakyReLU in the first encoder layer produces only a small reduction, which implies that most representational capacity is concentrated in deeper layers.

Two further variants examine representational flexibility. Substituting the cardinality-aware loss with pure BCE leads to only a slight decrease in AUPR, which may suggest at first glance that the alternative loss has limited effect. However, the substitution severely disrupts the predicted distribution of ADR counts, as demonstrated in Subsection 3.3 , where BCE-trained models consistently fail to reproduce the

long-tail cardinality dynamics of the dataset. This shows that the benefit of the cardinality-aware loss arises not from per-entry accuracy alone but from correctly modelling global structural constraints. Similarly, reducing the latent dimensionality to $\{512, 384\}$ produces another small but consistent decline in AUPR, which aligns with earlier latent-search experiments indicating that dimensions in the 1024–2048 range are more suitable for capturing the intrinsic complexity of ADR incidence patterns.

Taken together, these results show that the full FASA model achieves the highest performance because it integrates complementary mechanisms whose effects reinforce each other. The negative output bias, activation design, regularisation, weak supervision, latent capacity and cardinality-aware loss all contribute to the model's overall accuracy and stability. Although the architecture remains robust under moderate simplifications, its best performance emerges when all components operate jointly and are optimised as a unified system.

4 Discussion

The results of this study demonstrate that a feature-free, high-capacity autoencoder can reconstruct adverse drug reaction (ADR) incidence patterns with substantially higher fidelity than established matrix factorisation and feature-driven baselines. This finding challenges the common assumption that accurate ADR prediction requires chemical, biological, or pharmacokinetic descriptors, and instead shows that the binary drug–ADR matrix alone contains recoverable and pharmacologically meaningful structure. The model achieves strong performance across all metrics despite operating exclusively on sparse incidence data, suggesting that high-dimensional non-linear representations may provide a viable alternative to feature-dependent architectures.

A central insight emerging from our experiments is that model depth is not a substitute for representational width in this domain. Increasing the number of layers consistently degraded reconstruction metrics, likely due to oversmoothing and vanishing gradient effects on extremely sparse targets. In contrast, widening layers up to several thousand units enabled the model to capture higher-order co-occurrence patterns that shallow matrix factorisation approaches cannot represent. The systematic sweep over latent dimensionalities revealed a broad performance plateau at $d_{\text{Latent}} \in \{1024, 2048\}$, indicating that the intrinsic dimensionality of ADR incidence patterns is moderate but not low-rank in the classical matrix completion sense.

Early design choices also provide additional context for the final architecture. We initially explored a denoising autoencoder under the intuition that artificially removing entries might regularise the reconstruction process and reduce over-prediction. However, removing even 20% of observed positives produced a severe loss of signal, given that positives account for less than 3% of the matrix. This experiment underscores the extreme sparsity of ADR datasets and the fragility of approaches that rely on masking or dropout of observed entries. Similarly, focal loss was initially tested to mitigate imbalance, but provided no improvements over binary cross-entropy once the negative output bias was introduced. The negative initialisation of the final bias term (set to $b_0 = -3.61$) proved to be a simple yet crucial mechanism to

counteract the overwhelming prevalence of zeros without requiring database-specific weighting schemes.

The performance of the proposed model must also be contextualised within the broader ADR prediction literature, where methodological inconsistencies make it difficult to establish genuine state-of-the-art progress. Recent scoping reviews [25, 26] indicate that there is little consensus on what constitutes the input space for ADR models. Kim et al. [25] analyzed 72 studies and reported a strong tendency for machine learning approaches to rely on heterogeneous combinations of datasets, with DrugBank being the primary source for ML-based tasks. Similarly, Toni et al. [26] classified features into four distinct categories: chemical (e.g., SMILES, fingerprints), biological (e.g., targets, transporters, enzymes), phenotypical, and general. They found that the majority of models depend on integrating these diverse views to achieve predictive stability. This heavy reliance on external metadata results in selective benchmarking and makes it difficult to compare feature-rich models against pure matrix-completion approaches.

In addition to feature-related inconsistencies, evaluation practices across ADR-prediction studies further limit comparability. Several papers report unusually high AUPR values due to the use of synthetic resampling methods such as SMOTE, random undersampling of negatives, or the restriction of experiments to "easy" subsets (e.g., only cardiovascular ADRs) [27, 28, 29]. These techniques inflate apparent performance by altering the underlying class distribution rather than improving the true predictive capacity of the model. For instance, Random Forest models have been reported to achieve AUPR scores as high as 0.977 [26]. This stands in sharp contrast to findings from a recent scoping review [26] which showed that, under fair and fully imbalanced conditions, the median AUPR across 22 methods was approximately 0.41, with only a small minority of models, such as Zhang et al. [18], exceeding 0.64. This discrepancy exposes how permissive evaluation pipelines can produce overly optimistic results, obscuring the genuine difficulty of ADR prediction and hindering progress toward robust benchmarking standards.

Given these challenges, WPLMF was selected as the primary baseline because it provides a reproducible framework built on publicly available datasets, a fully documented training pipeline, and an evaluation protocol that aligns with standard matrix-completion practice. In addition, WPLMF remains one of the few methods designed specifically to complete binary ADR incidence matrices, albeit by incorporating auxiliary chemical or biological features. Its use of pseudo-labels, frequency-weighted losses, and node2vec embeddings reflects a sophisticated integration of external information, making it a strong and relevant comparator for assessing the contribution of the autoencoder's high-capacity latent modelling. The fact that the proposed model outperforms WPLMF despite relying solely on the incidence matrix emphasises the potential of non-linear reconstruction methods for ADR prediction.

Furthermore, our results directly address two specific limitations acknowledged by the authors of WPLMF. In their discussion, Chen et al. [9] noted that their pseudo-labelling approach yielded poor improvements for 'extremely low-frequency drug-ADR pairs' and hypothesised that incorporating auxiliary molecular structures would be necessary to solve this. Our findings challenge this assumption: FASA demonstrates that the 'tail' of the distribution can be effectively reconstructed without additional chemical

features, provided the model possesses sufficient non-linear capacity and is guided by structural constraints like the cardinality-aware loss. Additionally, WPLMF admits that its reliance on GMF (Generalized Matrix Factorization) for pseudo-labelling 'may not be optimal'. By replacing their shallow bilinear factorization with a deep stacked autoencoder, we confirm that increasing architectural depth provides the computational expressivity needed to resolve complex incidence patterns that linear baselines fail to capture.

Despite these promising results, the proposed framework has specific limitations inherent to its feature-agnostic design. First, because FASA relies exclusively on historical incidence patterns (collaborative filtering), it faces the 'cold-start' problem [30]: it cannot predict ADRs for a completely novel compound that has no recorded associations, whereas feature-based models could theoretically infer risks from chemical structure alone. Our approach is therefore best suited for post-marketing surveillance of drugs with at least partial history, rather than pre-clinical screening of new chemical entities. Second, the scope of this evaluation is strictly computational. While the model identifies high-confidence candidates for unobserved associations, we did not conduct qualitative case studies to manually verify these predictions. Interpreting the biological plausibility of specific drug–ADR pairs requires specialized pharmacological or toxicological expertise, which was beyond the scope of this work. Consequently, the clinical validity of the model's novel predictions should be assessed in future collaborations with domain experts.

5 Conclusion

This study introduces FASA, a feature-agnostic stacked autoencoder framework that challenges the prevailing assumption that accurate ADR prediction requires comprehensive biological or chemical descriptors. By operating exclusively on the binary drug–ADR incidence matrix, our model consistently outperforms established baselines, including those enriched with molecular graphs and protein interaction networks, across AUPR, F1, and precision metrics. These results demonstrate that the incidence matrix itself contains a rich, recoverable latent structure that has been largely underutilized by shallow factorization methods.

A key methodological contribution of this work is the introduction of a structure-aware cardinality loss. We showed that standard binary cross-entropy is insufficient for reconstructing the long-tailed distribution of adverse effects, leading to the systematic underestimation of polypharmaceutic risks. By explicitly constraining the model to match the global label-count distribution, FASA stabilizes reconstruction in the 'tail' of the distribution, recovering rare associations that standard methods miss. Furthermore, the use of a deep, noise-suppressing architecture allows the model to capture non-linear co-occurrence patterns that remain inaccessible to linear matrix factorization. The practical implications of these findings are significant. Because FASA does not rely on external databases such as SIDER, DrugBank, or protein interaction networks, it is immune to the 'missing metadata' problem that plagues feature-dependent models. This makes it particularly valuable for early-stage pharmacovigilance or for analyzing legacy datasets where chemical fingerprints may be incomplete or standardized terminologies unavailable. [31]

Future work will focus on two directions. First, while we have proven that auxiliary features are not

strictly necessary for state-of-the-art performance, they remain complementary. A hybrid architecture that injects molecular embeddings into the bottleneck of our pre-trained autoencoder could potentially refine predictions for the rarest drugs. Second, the cardinality-aware loss function proposed here is generalizable; we intend to apply this objective to other high-sparsity biomedical tasks, such as gene–disease association prediction and drug–drug interaction screening, where class imbalance and long-tailed distributions present similar challenges.

Declarations

Abbreviations

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used in this study were obtained from the WPLMF repository available at https://github.com/WhoIsBalance/WPLMF. This dataset was originally constructed using the SIDER database (http://sideeffects.embl.de) and DrugBank (https://go.drugbank.com) as described in Chen et al. [9].

The source code and implementation of the proposed method (FASA), including scripts to reproduce the main experiments, are available at https://github.com/Owzok/FASA-ADR.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Author contributions

Martin G. Perez-Bonany Torrealva: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft; Rensso V. H. Mora-Colque: Supervision, Project administration; Edward Jorge Y. Cayllahua Cahuina: Supervision, Writing—review & editing. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

[1] Greta Masiliūnienė, Edgaras Stankevičius, and Edmundas Kaduševičius. Trends in adverse drug reaction reporting in eight selected countries after the implementation of new pharmacovigilance regulation in 2012: a joinpoint regression analysis. *European Journal of Clinical Pharmacology*, pages 1–13, 2025.

[2] Marie-Laure Laroche, Noémie Tarbouriech, Taha Jai, Marie-Blanche Valnet-Rabier, and Virginie Nerich. Economic burden of hospital admissions for adverse drug reactions in france: The iatrostat-eco study. *British Journal of Clinical Pharmacology*, 91(2):439–450, 2025.

[3] David W. Nierenberg, Kenneth L. Melmon, H. Morrelli, and Brian B. Hoffman. *Clinical Pharmacology: Basic Principles in Therapeutics*. McGraw-Hill Professional, New York, 1992.

[4] Diego Galeano, Shantao Li, Mark Gerstein, and Alberto Paccanaro. Predicting the frequencies of drug side effects. *Nature communications*, 11(1):4575, 2020.

[5] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[6] Wen Zhang, Xinrui Liu, Yanlin Chen, Wenjian Wu, Wei Wang, and Xiaohong Li. Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing*, 287:154–162, 2018.

[7] K. Fukuto, T. Takagi, and YS. Tian. Predicting the side effects of drugs using matrix factorization on spontaneous reporting database. *Sci Rep*, 2021.

[8] Yijie Ding, Jijun Tang, and Fei Guo. Identification of drug-side effect association via semisupervised model and multiple kernel learning. *IEEE journal of biomedical and health informatics*, 23(6):2619–2632, 2018.

[9] Chen. Junheng, Han. Fangfang, and He. et al. Mingxiu. A novel weighted pseudo-labeling framework based on matrix factorization for adverse drug reaction prediction. *BMC Bioinformatics*, 2025.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

[11] Onur Can Uner, Halil Ibrahim Kuru, R Gokberk Cinbis, Oznur Tastan, and A Ercument Cicek. Deepside: a deep learning approach for drug side effect prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1):330–339, 2022.

[12] Liyi Yu, Meiling Cheng, Wangren Qiu, Xuan Xiao, and Weizhong Lin. idse-he: Hybrid embedding graph neural network for drug side effects prediction. *Journal of Biomedical Informatics*, 131:104098, 2022.

[13] Xianyu Xu, Ling Yue, Bingchun Li, Ying Liu, Yuan Wang, Wenjuan Zhang, and Lin Wang. Dsgat: predicting frequencies of drug side effects by graph attention networks. *Briefings in Bioinformatics*, 23(2), 2022.

[14] Wuyong Liu, Jingyu Zhang, Guanyu Qiao, Jilong Bian, Benzhi Dong, and Yang Li. Hmmf: a hybrid multi-modal fusion framework for predicting drug side effect frequencies. *BMC bioinformatics*, 25(1):196, 2024.

[15] Haochen Zhao, Kai Zheng, Yaohang Li, and Jianxin Wang. A novel graph attention model for predicting frequencies of drug–side effects from multi-view data. *Briefings in bioinformatics*, 22(6):bbab239, 2021.

[16] Lin Wang, Chenhao Sun, Xianyu Xu, Jia Li, and Wenjuan Zhang. A neighborhood-regularization method leveraging multiview data for predicting the frequency of drug–side effects. *Bioinformatics*, 39(9):btad532, 2023.

[17] Liangwei Nathan Zheng, Chang George Dong, Wei Emma Zhang, Xin Chen, Lin Yue, and Weitong Chen. Devil in the tail: A multi-modal framework for drug-drug interaction prediction in long tail distinction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3395–3404, 2024.

[18] Wen Zhang, Hua Zou, Longqiang Luo, Qianchao Liu, Weijian Wu, and Wenyi Xiao. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing*, 173:979–987, 2016.

[19] Fangyu Zhou and Shahadat Uddin. Interpretable drug-to-drug network features for predicting adverse drug reactions. In *Healthcare*, volume 11, page 610. MDPI, 2023.

[20] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 11 2023.

[21] M. Kuhn, I. Letunic, L. Jensen, and P. Bork. The sider database of drugs and side effects. *Nucleic Acids Res.*, 2016.

[22] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[23] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[24] Diego Galeano and Alberto Paccanaro. A recommender system approach for predicting drug side effects. In *2018 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[25] Hae Reong Kim, MinDong Sung, Ji Ae Park, Kyeongseob Jeong, Ho Heon Kim, Suehyun Lee, and Yu Rang Park. Analyzing adverse drug reaction using statistical and machine learning methods: A systematic review. *Medicine*, 101(25):e29387, 2022.

[26] Esmaeel Toni, Haleh Ayatollahi, Reza Abbaszadeh, and Alireza Fotuhi Siahpirani. Machine learning techniques for predicting drug-related side effects: a scoping review. *Pharmaceuticals*, 17(6):795, 2024.

[27] Vladislav S Sukhachev, Sergey M Ivanov, and Alexander V Dmitriev. Prediction of adverse effects of drug–drug interactions on cardiovascular system based on the analysis of structure–activity relationships. *Biochemistry (Moscow)*, 88(5):630–639, 2023.

[28] John-Paul Cooper, Pierre Chue, Arno G Siraki, and Lusine Tonoyan. Predicting clozapine-induced adverse drug reaction biomarkers using machine learning. *Scientific Reports*, 15(1):25572, 2025.

[29] Jianxiang Wei, Guanzhong Feng, Zhiqiang Lu, Pu Han, Yunxia Zhu, and Weidong Huang. Evaluating drug risk using gan and smote based on cfda's spontaneous reporting data. *Journal of healthcare engineering*, 2021(1):6033860, 2021.

[30] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, 2002.

[31] Annette Haerdtlein, Elisabeth Debold, Marietta Rottenkolber, Anna Maria Boehmer, Yvonne Marina Pudritz, Faiza Shahid, Jochen Gensichen, and Tobias Dreischulte. Which adverse events and which drugs are implicated in drug-related hospital admissions? a systematic review and meta-analysis. *Journal of Clinical Medicine*, 12(4):1320, 2023.