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.

Informative alignment patterns between chemical structure and clinical signals support context-aware toxicity assessment

Hiroto Asano
The University of Osaka
Yu-Shi Tian
yushi-tian@phs.osaka-u.ac.jp
The University of Osaka https://orcid.org/0000-0002-8988-9453
Asuka Hatabu
The University of Osaka https://orcid.org/0000-0001-9886-1447
Kaori Fukuzawa
The University of Osaka https://orcid.org/0000-0001-5357-8250
Daisuke Takaya
The University of Osaka
Kenji Ikeda
The University of Osaka

Article

Keywords: Computational toxicology, drug development, drug repurposing, pharmacovigilance, structure–activity relationship

Posted Date: March 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8923363/v1

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

Additional Declarations: There is **NO** Competing Interest.

Informative alignment patterns between chemical structure and clinical signals support context-aware toxicity assessment

Hiroto Asano[1], Yu-Shi Tian[1]*, Asuka Hatabu[1], Kaori Fukuzawa[1], Daisuke Takaya[1], Kenji Ikeda[1]

[1]Graduate School of Pharmaceutical Sciences, The University of Osaka, 1-6 Yamadaoka, Suita, Osaka 565-0871, Japan

ORCIDs:

Hiroto Asano: 0000-0002-0973-6773

Yu-Shi Tian: 0000-0002-8988-9453

Asuka Hatabu: 0000-0001-9886-1447

Kaori Fukuzawa: 0000-0001-5357-8250

Daisuke Takaya: 0000-0001-6020-8151

Kenji Ikeda: 0000-0001-8299-4464

***Corresponding author:**

Yu-Shi Tian

Graduate School of Pharmaceutical Sciences, The University of Osaka

1-6 Yamadaoka, Suita City, Osaka 565-0871, Japan

Phone: +81-6-6879-8242

Fax.: +81-6-6879-8242

Email: yushi-tian@phs.osaka-u.ac.jp

Key Words:

Computational toxicology, drug development, drug repurposing, pharmacovigilance, structure–activity relationship

Abstract

Drug safety remains a major challenge in drug development. In clinical practice, safety outcomes are influenced not only by chemical structure but also by patient characteristics and treatment conditions, limiting the ability of structure-based predictions to fully capture real-world adverse events. Here, we integrated post-marketing safety data from the FDA Adverse Event Reporting System (FAERS; 2004Q1–2025Q2) with chemical structure and drug label information. Using 875 drugs, we quantified concordance between structure-based similarity and FAERS safety signals, as well as agreement between FAERS signals and label-reported risks. To support large-scale, reproducible signal detection, we developed a dedicated analytical application. Based on these measures, drugs were classified into four safety patterns. Pattern 1 (56.4%) showed consistent alignment among chemical structure, real-world safety signals, and labels, indicating informative structure–safety relationships. Pattern 2 (38.4%) diverged from structural similarity yet agreed with label information, suggesting safety profiles influenced by individual biology or clinical context. Pattern 3 (2.7%) reflected treatment-dependent attenuation of expected risks, whereas Pattern 4 (2.5%) revealed unanticipated risks not predicted by structure or labels. Collectively, this framework provides a pattern-based perspective on drug safety and suggests that such stratification may facilitate drug repurposing and support structure–activity relationship analyses.

Introduction

Early identification of potential safety risks is a fundamental objective in modern drug discovery.[1] Evaluating these risks through structure–activity relationship (SAR) analyses at the earliest stages of development enables the rational design of safer drug candidates and the implementation of more targeted evaluation strategies. This proactive approach allows research teams to prioritize compounds with favorable safety profiles and to better anticipate biological responses before clinical testing.[2]

Despite extensive preclinical evaluation using *in vitro* and *in vivo* models, unanticipated clinical toxicity remains a leading cause of drug attrition and post-marketing withdrawal.[3,4] Many adverse events are not readily detected in standard experimental models, particularly when the effects are rare, delayed, or associated with long-term use. Therefore, bridging the gap between preclinical safety predictions and real-world clinical outcomes remain a critical challenge in pharmaceutical development.

A major contributor to this gap is that clinical safety outcomes are influenced by factors beyond the intrinsic chemical properties of drug compounds. Drug responses in real-world settings are shaped by patient-specific variables such as age, genetic background, organ function, comorbidities, and disease severity.[5] These multifaceted clinical factors introduce a level of complexity that cannot be fully captured by chemical structure-based models alone.

Safety evaluation is therefore an iterative process that spans the entire lifecycle of a drug.[6] Early discovery relies on structure-based predictions to link chemical similarity with biological effects.[7] This is followed by randomized clinical trials, which assess safety in human populations within controlled environments, following strict protocols to minimize bias and characterize risks under predefined conditions. Finally, post-marketing pharmacovigilance systems capture spontaneous adverse event reports, reflecting a drug's performance across diverse patient populations and real-world settings.[6] Each stage provides a complementary perspective on the safety profile of drugs. SAR-based methods are essential for early screening and molecular design, as they identify risks associated with specific chemical motifs. Clinical trials generate high-quality safety data in human subjects under defined conditions. In contrast, pharmacovigilance data can uncover safety signals that emerge only in complex clinical environments, such as those involving comorbidities, polypharmacy, or uncommon patient profiles.

Therefore, integrating structure-based predictions with clinical trial outcomes and post-marketing pharmacovigilance data is necessary for a comprehensive understanding of drug safety. Consistency across these sources strengthens confidence in a drug's safety profile, whereas differences highlight the complementary roles of clinical trials and real-world surveillance. Importantly, these differences often reveal risks that emerge only in real-world

settings—driven by patient variability, comorbidities, or drug combinations—that are not detected from structure-based predictions alone.

In this study, we integrated 21 years of data (2004–2025) from the FDA Adverse Event Reporting System (FAERS), with chemical structure information and drug labeling data from OnSIDES. This integrated dataset enabled systematic characterization of the safety profiles of 875 drugs by directly comparing chemical similarity with real-world adverse event patterns. We conducted a multidimensional comparative analysis incorporating chemical descriptors, label-reported risks, and real-world adverse event profiles. By synthesizing these distinct data modalities, we classified drugs into four safety patterns representing different degrees of alignment or divergence between structure-based expectations and clinical outcomes. This framework highlights where chemical similarity correlates with safety outcomes and, more importantly, where clinical factors drive divergence from structural predictions.

The objective of this study was to provide a practical, pattern-based framework that clarifies the relationship between chemical profiles and real-world safety. By delineating the complementary roles of SAR, clinical trials, and post-marketing surveillance, this approach supports more informed safety assessment, targeted SAR studies, and improved decision-making throughout drug development, establishing a foundation for more proactive and efficient safety monitoring.

Results

Large-scale signal detection across 21 years of post-marketing surveillance data

We constructed a curated FAERS dataset covering Q1 2004 to Q2 2025 (Figure 1).[8] Preprocessing included harmonizing adverse event terminology according to the Medical Dictionary for Regulatory Activities (MedDRA) version 28.1 and standardizing units and date formats across records. Duplicate reports and records with temporal inconsistencies were also excluded. To address heterogeneity in drug naming within the DRUG table, we standardized drug strings using the DiAna Dictionary[9] and an expanded normalization procedure designed to capture ingredient-level names from noisy entries better, increasing drug name normalization coverage within the curated FAERS database from 97.7% to 99.03% for all mentions and from 39.8% (351,563/883,058) to 66.8% (590,277/883,058) for unique entries.

We implemented high-throughput disproportionality analyses using the Adverse Event Signal-detection Tool (AEST), an originally developed web-based signal detection platform optimized for large-scale FAERS processing. AEST uses label encoding and an HDF5-based storage design with a Virtual Dataset (VDS)[10,11] layout to support efficient longitudinal querying without physically merging quarterly files. When applied to the curated FAERS records stored in the AEST database, this optimized workflow reduced the storage footprint by 77% (from 5.47 GiB to 1.25 GiB) relative to a baseline representation without label encoding or HDF5. To

accelerate the large-scale aggregation, computations were parallelized across quarterly segments using Ray[12] in a 20 vCPUs/30 GiB environment, achieving an average processing time of approximately 6 s per compound.

Using established thresholds for the Reporting Odds Ratio (ROR) and Proportional Reporting Ratio (PRR),[13–15] we defined an adverse event (AE) signal as meeting the criteria for either method and generated PT-level binary safety profiles for downstream analyses. Across 3,110 drugs with Anatomical Therapeutic Chemical (ATC) codes, this workflow yielded analyzable post-marketing AE signal profiles.

Chemical–AE signal similarity alignment stratifies drugs into high-, low-, and intermediate-alignment groups

Among the 3,110 drugs with valid ATC codes, 2,159 compounds with available chemical information were included in the chemical–AE signal similarity alignment analysis. For these compounds, we analyzed a 2,159 × 696 2D chemical descriptor matrix and a corresponding 2,159 × 27,163 PT-level binary AE signal matrix derived from significant ROR or PRR signals. Aligned-UMAP[16] embeds the chemical descriptor and AE signal matrices into a shared low-dimensional space, achieving moderate preservation of the local neighborhood structure (Figure 2). Quantitative evaluation showed trustworthiness/continuity[17] values of 0.70/0.66 for chemical

descriptors and 0.71/0.73 AE respectively. Robustness analyses across 81 runs with independent ±10% hyperparameter perturbations showed minimal variation in embedding quality (chemical: trustworthiness 0.701±0.027, continuity 0.671±0.034; AE signal: trustworthiness 0.715±0.003, continuity 0.738±0.012), indicating that the aligned representations are stable to parameter changes.

The chemical–AE signal similarity alignment was quantified using two complementary measures capturing global and local alignments across the two embedding spaces (Figure 3a, 3b). We defined thresholds based on the elbow points of the cumulative distribution functions for these metrics: a Euclidean distance of 0.835 and a Jaccard index of 0.037. Based on these cutoffs, the drugs were classified into three groups: high alignment (low embedding distance and high neighbor overlap), low alignment (high embedding distance and low neighbor overlap), and intermediate (all remaining drugs). This procedure yielded 1,229 high-alignment (56.92%), 804 low-alignment (37.24%), and 126 intermediate-alignment (5.84%) drugs.

We evaluated whether these alignment-defined groups reflected the separable structures in the original high-dimensional data using PERMANOVA and PERMDISP.[18,19] In the AE signal space, the high- and low-alignment groups were significantly separated by PERMANOVA (pseudo-F=1.37, Bonferroni-adjusted p=0.012; raw p=0.001), whereas no significant separation was observed in the chemical descriptor space (pseudo-F=0.84, p=0.53). PERMDISP suggested

a dispersion difference in the AE signal space before multiple-testing correction (pseudo-F =7.89, raw p=0.006); however, this effect was not significant after Bonferroni adjustment (p_adj=0.072).

These results suggested that high- and low-alignment drugs formed statistically distinguishable groups in the adverse-event signal space. When interpreted alongside the PERMDISP results, the group separation detected by PERMANOVA was more likely attributable to differences in the group centroids than to differences in the within-group dispersion. In contrast, no significant differences were observed between the two groups in terms of the chemical descriptor space. These findings indicate that the separation observed in the UMAP projection reflects distinctions intrinsic to adverse event signal profiles, rather than differences in chemical properties. Thus, although high- and low-alignment drugs may share similar chemical characteristics, they appear to differ in their adverse event profiles. This may partly explain the limitations of toxicity prediction approaches that rely primarily on chemical features, including quantitative structure–activity relationship (QSAR) models.

Signal–label alignment stratifies drugs into high- and low-alignment groups

Among the 3,110 drugs with valid ATC codes, 1,450 overlapped with drugs reported in OnSIDES,[20] a database of AEs extracted from drug labels, and were included in the signal–label

alignment analysis. For each drug, we encoded label-listed events across 27,163 PTs and quantified signal–label alignment by computing Yule's Q between the FAERS-derived binary AE-signal vector and the corresponding OnSIDES label vector, with Yule's Q ranging from −1 to +1 as a bounded association measure appropriate for sparse, highly imbalanced adverse-event profiles.

To stratify drugs by signal–label alignment, we fitted Gaussian mixture models (GMMs)[21] (k=1–5) to the distribution of Yule's Q and selected the optimal model using the Bayesian Information Criterion (BIC).[22] This procedure favored a two-component model, yielding high-alignment (95.2%, n=1,380) and low-alignment (4.8%, n=70) groups (Figure 4).

Four-patterns risk stratification framework construction

We constructed a four-pattern risk stratification framework for 875 drugs possessing both chemical descriptors and OnSIDES labeling information by integrating structure–signal alignment (similarity between chemical descriptors and signals) and signal–label alignment. For structure–signal alignment, the analysis was restricted to high- versus low-alignment classes, as supported by a pairwise PERMANOVA utilizing a Euclidean distance of 0.835 as the separation threshold. Signal–label alignment was dichotomized using Yule's Q GMMs.

In the integrated framework, the x-axis represents Yule's Q (higher values indicate stronger correlation with labeling), and the y-axis represents the Euclidean distance (lower values indicate higher similarity alignment between chemical descriptors and signals). Based on these thresholds, the drugs were classified into four distinct safety information alignment patterns (Figure 5, Supplementary Tables 1–4).

Pattern 1 (consistent safety profile; high-x/low-y) was characterized by high alignment in both dimensions (56.4%, n=493).

Pattern 2 (individual biology-driven risk; high-x/high-y) was characterized by high label alignment but low structural alignment (38.4%, n=336).

Pattern 3 (treatment environment: low-x/low-y) was characterized by low label alignment despite high structural alignment (2.7%, n=24).

Pattern 4 (hidden risk; low-x/high-y) was characterized by low alignment in both dimensions (2.5%, n=22).

Therapeutic category composition differed significantly across patterns at ATC level 1 (Chi-square test, $p = 5.05\times10^{-6}$). Standardized residual analysis identified specific overrepresentations: ATC A (alimentary tract and metabolism; residual = 2.02) and ATC R (respiratory system; 2.20) in pattern 3; ATC D (dermatologicals; 3.56) and ATC S (sensory

organs; 3.40) in pattern 4; and ATC H (systemic hormonal preparations, excluding sex hormones and insulins; 2.07) in pattern 2.

At ATC level 3, no subgroup remained significant after the Bonferroni correction. This indicated that the enrichment signal observed at ATC level 1 did not robustly localize to specific ATC level 3 subcategories under conservative multiple-testing control.

To explore the mechanistic and contextual characteristics of our framework, we evaluated Pattern 2 for biological underpinnings and Pattern 3 for potential polypharmacy effects.

For Pattern 2, FDA-listed pharmacogenetic associations[23] mentioning active metabolites were significantly more frequent than in Pattern 1 (odds ratio 9.8, $p = 0.0078$; Figure 6), with Patterns 3 and 4 excluded as they contained no drugs on the list. This observation suggests that risk expression in this category may be linked to metabolic factors beyond parent chemical profiles alone. For Pattern 3, our analysis of prescription claims data[24] suggests that drugs in this category are often used in multi-drug regimens, with the relative frequency reaching an approximately 3.7-fold peak in triple-therapy combinations compared with monotherapy (Figure 7). Despite the limited sample size, the observed trend suggests that Pattern 3 captures risk signals arising in complex clinical settings, rather than from isolated pharmacological effects.

Discussion

We integrated 21 years of post-marketing surveillance signals, chemical informatics, and label-derived adverse event profiles to stratify drugs based on the alignment of safety evidence across three orthogonal data sources. By quantifying the alignment between intrinsic chemical descriptors, real-world reporting patterns, and drug labels, we established a four-pattern risk-stratification framework. Since spontaneous reporting captures drug safety within actual clinical contexts, the resulting patterns reflect a synthesis of pharmacological properties and real-world factors, enabling system-level standardization of drug safety profiles beyond isolated signal detection.

Pattern 1 (Information Convergence) included the majority of analyzed drugs (56.4%, n=493), which structural predictions, signals, and approved labeling showed strong alignment. This served as a validation baseline for the framework, confirming that the chemical descriptors, drug labels, and signals were generally consistent across most drugs.

Notably, ATC enrichment analysis showed no significant over-representation of specific drug classes, unlike the other patterns. The absence of bias indicates that information convergence represents a general baseline state that is not limited to specific therapeutic categories.

Pattern 2 (individual biology-driven risk) comprised a substantial proportion of the analyzed drugs (38.4%, n = 336). Drugs in this quadrant display adverse event signals that diverge from predictions based on chemical structure yet remain consistent with established safety labeling (OnSIDES). This structure–signal inconsistency indicates that the safety profiles of these agents are not primarily encoded in their chemical features alone but are conditionally expressed through individual- and context-dependent factors, including metabolic variability, genetic susceptibility, comorbidities, and treatment conditions.

Canonical examples illustrate this concept. The risk of hepatotoxicity associated with isoniazid is influenced by NAT2 acetylation phenotypes,[25] while severe cutaneous adverse reactions induced by carbamazepine are closely linked to specific HLA alleles.[26] Similarly, the clinical safety profiles of codeine and tramadol depend on CYP2D6-mediated bioactivation, resulting in substantial inter-individual variability driven by genetic polymorphisms and drug–drug interactions.[27,28] Consistent with this context-dependency, residual analysis revealed a significant enrichment of systemic hormonal agents (ATC code H, excluding sex hormones and insulin) within this pattern. Synthetic glucocorticoids such as dexamethasone exemplify this class, acting across interconnected endocrine and metabolic pathways in which both efficacy and safety are profoundly shaped by patient-specific physiology, disease state, and hormone sensitivity.[29]

Importantly, this context-dependent manifestation of risk suggests that drugs with pattern 2 can exhibit distinct clinical characteristics across different therapeutic settings. Because the observed adverse events often reflect secondary or downstream effects in humans under appropriate conditions, these same agents may confer meaningful therapeutic benefits. In line with this notion, several drugs with this pattern have recently received regulatory approval or are currently under active investigation for novel indications.

For example, colchicine expanded beyond its long-standing use in gout to receive FDA approval for cardiovascular risk reduction in 2023,[30] targeting inflammatory risk through pleiotropic anti-inflammatory mechanisms that extend beyond its classical role as a microtubule inhibitor.[31] Similarly, amisulpride has long been used for the treatment of schizophrenia, with dose-dependent pharmacological effects. Although lower oral doses preferentially act on presynaptic dopamine D2/D3 receptors, higher doses primarily antagonize postsynaptic receptors, resulting in distinct clinical profiles.[32] In 2020, modification of both the dose and route of administration, specifically, low-dose intravenous delivery (5 mg), allowed its repurposing for the prevention and rescue treatment of postoperative nausea and vomiting,[33] reflecting a therapeutic application that differs from its original psychiatric indication.

Additionally, ganaxolone, a synthetic analog of the endogenous neurosteroid allopregnanolone, features a 3β-methyl modification that enhances metabolic stability and reduces hormonal

effects.[34] This modification enables oral administration and confers a distinct pharmacological profile by selectively modulating the extrasynaptic $GABA_A$ receptors.[35] Ganaxolone, which was approved in 2022 for CDKL5 deficiency disorder, is also under investigation for refractory status epilepticus and tuberous sclerosis complex,[36] highlighting the potential clinical impact of its subtle structural differences and unique pharmacological properties.

Taken together, these observations suggest that drugs with inconsistent chemical profiles and real-world safety signals constitute a distinct and biologically meaningful class. These drugs highlight the importance of individual and clinical contexts in shaping drug responses and may constitute promising candidates for drug repositioning.

Pattern 3 (treatment environment), accounting for 2.7% (n = 24) of the dataset, represents a subset of drugs that exhibit high structural concordance and persistent adverse event signals in FAERS, yet show limited alignment with the safety descriptions provided in their drug labels. This discordance does not necessarily indicate failure of the prediction framework or a lack of safety data. Rather, it likely reflects a gap between controlled, label-based evaluations of individual drugs and complex multidrug treatment environments captured in real-world pharmacovigilance data.

A characteristic of this group is that although the drugs share chemical features suggestive of specific toxicities or adverse events and show corresponding signals in pharmacovigilance analyses, these structural risks are not always clearly reflected as clinically apparent adverse outcomes when the agents are evaluated individually.

Paromomycin is an example. In the United States, paromomycin is administered orally and is primarily used to treat acute and chronic intestinal amoebiasis. Although it is classified as an aminoglycoside, the drug class commonly associated with nephrotoxicity and ototoxicity is not specifically listed in the DailyMed information. This finding may be related to limited gastrointestinal absorption and the resulting low level of systemic exposure.[37] In developed countries, intestinal amoebiasis has been increasingly reported in HIV-infected individuals living with HIV.[38] This population may present with several clinical characteristics, including HIV-associated nephropathy, otologic symptoms, and the use of concomitant medications with nephrotoxic potential, such as tenofovir and atazanavir.[39–42] According to outpatient prescription claims databases,[24] paromomycin can be co-prescribed in situations where drugs such as tenofovir or atazanavir are used. These factors may be informative when considering the relationship between structural features and pharmacovigilance data.

Residual analysis further revealed that drugs classified under the alimentary and metabolic systems (ATC code A) and the respiratory system (ATC code R) were significantly enriched in

this group. Many of these medications are used concomitantly with other therapies and are often administered over the long term. As FAERS records drugs reported at the time of adverse event occurrence, co-prescribed agents or those producing similar clinical manifestations can contribute to the apparent signal overlap. As a result, structural risk predictions and real-world signals may coincide, even when the intrinsic toxic potential of the compound is low.

Taken together, the signals observed for Pattern 3 drugs are not necessarily false positives, but may instead reflect adverse outcomes arising in patients with severe underlying conditions or those receiving intensive multi-drug therapy. At the same time, these findings underscore the shared limitations of both SAR models and spontaneous reporting data, highlighting the importance of considering biological modifiers of exposure and clinical context when interpreting individual drug risks.

Pattern 4 (hidden risk), accounting for 2.5% (n = 22) of the dataset, represents a distinct subset of drugs for which structural features, approved label information, and pharmacovigilance signal detection results do not consistently align. This pattern suggests the presence of risks that are challenging to predict from chemical structures or during clinical trials and may become increasingly apparent in post-marketing real-world data, often after longer periods of follow-up. Residual analysis revealed that this group was significantly enriched for dermatological agents

(ATC code D) and sensory drugs (ATC code S). These drugs are primarily intended for topical or local administration. They are often assumed to have minimal systemic effects compared with oral or intravenously administered medications. However, this assumption can be misleading. Under certain clinical conditions, topical agents may be systemically absorbed, leading to serious adverse effects.[43]

A representative example is mafenide acetate, a topical antimicrobial that is widely used to prevent infections in patients with burns. Mafenide acetate is known to inhibit carbonic anhydrases through its metabolite p-sulfamylbenzoic acid, leading to metabolic acidosis. This risk is particularly pronounced in patients with extensive burns, especially those with inhalation injuries and concomitant respiratory acidosis.[44] Indeed, a recent clinical report described a case in which metabolic acidosis worsened rapidly after mafenide application, even under contemporary burn management.[45]

Importantly, the occurrence of systemic toxicity depends on the clinical context, including usage patterns (dose and frequency), application sites, and individual factors such as age.[46] These factors may not be fully apparent in the early stages of drug development, including in clinical trials, and often become better understood over time through routine post-marketing surveillance, which monitors drugs across broader populations and longer periods of use.[47]

The inclusion of ranitidine in this quadrant further validates our framework. Ranitidine is a

systemic drug, and safety concerns arise primarily from external factors, such as contamination with N-nitrosodimethylamine (NDMA), rather than from its intrinsic chemical descriptors or pharmacological mechanism.[48,49] The classification of ranitidine into a pattern characterized by low structural risk demonstrates that our approach can capture clinically relevant risks driven by manufacturing and external factors that traditional structure-based analyses fail to detect. These results highlight the need for continuous real-world pharmacovigilance to capture safety signals that are not evident solely through chemical descriptors or clinical trials.

This study stratified drugs into four distinct categories based on the alignment of chemical descriptors, drug label information, and safety signals, offering a unified framework for interpreting heterogeneity in adverse drug reactions. By jointly considering structure-based predictions and real-world evidence, this alignment approach provides practical insights into safety assessments in clinical trials, pharmacovigilance, and hypothesis generation for drug repositioning.

Our exploratory analyses further suggest that this stratification can inform the construction of more effective QSAR models for adverse event prediction (Supplementary Note 2; Supplementary Table 7). In a benchmark dataset,[50] random forest QSAR models trained within pattern-defined subsets showed higher AUCs for three of the four adverse events examined—

acute kidney injury and acute myocardial infarction in Pattern 1 (AUC 0.92 and 0.77; ΔAUC 0.17 and 0.25, respectively) and upper gastrointestinal bleeding in Pattern 2 (AUC 0.91; ΔAUC 0.22)—compared with size- and ratio-matched random baselines, with Benjamini–Hochberg adjusted q-values of 0.09 meeting the pre-specified exploratory FDR threshold ($q < 0.10$). The improvement in AUC observed for upper gastrointestinal bleeding in Pattern 2 is notable, as it may reflect the influence of biological factors. Pattern 2 includes several non-steroidal anti-inflammatory drugs, such as piroxicam, for which polymorphisms in CYP2C9 have been associated with an increased risk of upper gastrointestinal bleeding.[51,52] In addition, Pattern 2 contains antithrombotic agents such as clopidogrel, whose CYP2C19-mediated bioactivation contributes to variability in bleeding risk.[53] Although based on limited examples, pattern-based stratification may reduce biological and mechanistic heterogeneity within datasets, thereby contributing to improved QSAR performance.

Interestingly, the same pattern-based stratification approach appears to be applicable to drug repositioning. In a separate exploratory analysis (Supplementary Note 3; Supplementary Table 8), we applied a similar four-pattern framework to evaluate the repositioning potential of approved drugs. Pattern 2, which covers a wide range of therapeutic areas while maintaining a hit rate comparable to or higher than the overall hit rate obtained without any pattern-based

grouping, serves as a practical starting point for prioritization in DR studies. These findings suggest that applying this framework could enhance the efficiency of drug repositioning.

This study has several limitations that should be considered when interpreting the results. First, FAERS data are subject to well-recognized biases, including under-reporting,[54] Weber effects,[55] and lack of denominator information. While disproportional metrics such as ROR and PRR partially mitigate these issues, they cannot fully eliminate them. Second, the use of 2D chemical descriptors constrains the representation of the stereochemical, conformational, and dynamic properties that may contribute to toxicity. Third, although OnSIDES provides a standardized and curated source of label-derived adverse events, it is a secondary resource derived from regulatory texts and may not fully capture the complexity underlying regulatory decision-making. These considerations regarding data provenance and chemical representation have been expanded further (Supplementary Note 4).

Further investigations are required to address these limitations. Nevertheless, this framework offers a scalable and interpretable approach to integrate chemical profiles, label information, and real-world safety data. By explicitly modeling the alignment and discordance across the three dimensions, the proposed strategy supports a more informed safety anticipation from early drug discovery through post-marketing surveillance.

Conclusion

In this study, we integrated chemical structure data, drug label information, and post-marketing safety signals derived from real-world clinical practice, and quantitatively examined their agreement and differences to propose a pattern-based perspective on drug safety. By analyzing 21 years of FAERS data, together with chemical structure descriptors and label-derived adverse events, we showed that drugs can be grouped into those for which structure-based expectations are generally consistent with observed clinical safety and those for which safety signals are not fully accounted for by structure alone and may be influenced by factors such as interindividual variability, treatment context, or usage conditions.

Although the analyses were exploratory and did not address causal mechanisms, the results suggested that organizing drugs according to these patterns may be useful for interpreting safety signals and informing subsequent analyses, including exploratory drug repurposing considerations and QSAR-based adverse event modeling. The four safety patterns introduced here are intended as an interpretive framework for relating structural and clinical safety information rather than as definitive safety classifications.

As a secondary contribution, we developed AEST, a web-based platform that enables large-scale reproducible analysis of FAERS data and supports transparent pharmacovigilance analyses.

Overall, explicitly considering the agreement and differences among chemical structure descriptors, post-marketing safety signals, and label information offers a preliminary framework for organizing safety-related evidence across the drug lifecycle and understanding how structure-based prediction and clinical context contribute to drug safety assessment.

Methods

FAERS data acquisition, processing, and quality control

FAERS data spanning Q1 2004 through Q2 2025 were obtained from FDA public repositories.[8] FAERS is a voluntary reporting system with substantial free-text entry; therefore, data are subject to quality issues, such as input errors and inconsistent terminology, which can affect the accuracy of signal detection.[56]

We standardized patient characteristics (e.g., age, weight, sex, date, and reporting country). Body weight was recorded in multiple units (e.g., g, lb) and converted to kilograms; non-convertible values were set to missing values. Age was converted to years, records with non-convertible ages were excluded, and values outside 0–125 years were treated as missing, considering the plausible physiological limits (maximum documented lifespan: 122 years).[57] Event onset dates were converted to the YYYY-MM-DD format and unparseable values were

missing. Reporting countries were harmonized with ISO 3166 codes,[58] and obvious typographical errors were manually corrected. Sex was classified as male (M), female (F), or unknown (all other entries).

Adverse events were coded using MedDRA®, an international standardized medical terminology developed under the auspices of ICH.[59] As FAERS contains records based on different MedDRA versions and includes non-standard term variants, we harmonized all adverse event terms with PTs version 28.1. Indications were not consistently MedDRA compliant; therefore, we mapped indication terms to the PT where possible and excluded records that could not be standardized after manual review.

We merged DRUG, THER, and INDI using primary_id and drug_seq and merged all other tables using primary_id. We defined duplicates as records identical across drug name, therapy start and end dates, indication, adverse event, sex, event onset date, weight, age, and reporting country. Following FDA guidance, we retained the most recent record and removed the remainder. Finally, we excluded logically inconsistent reports, i.e., those in which the recorded therapy start date occurred after the date of adverse event onset.

AEST development and signal detection methodology

We developed AEST, a web-based application designed to enable comprehensive, high-

throughput detection of pharmacovigilance signals (Supplementary Note 1; Supplementary Figures 1–6 and Supplementary Tables 5, 6). AEST integrates (i) database optimization for efficient storage and querying, (ii) name normalization to mitigate name heterogeneity, and (iii) distributed parallel computing for accelerating large-scale aggregation.

For data management, string fields in the FAERS (e.g., drug names and adverse event terms) were label-encoded and stored in a hierarchical HDF5.[10] Quarterly FAERS releases are presented as a single logical dataset via a VDS, eliminating the need for physical file merging.[11] This architecture reduces the memory footprint and improves I/O performance, enabling efficient longitudinal querying and aggregation.

To address heterogeneous drug names (e.g., generic versus brand names), we adopted the DiAna Dictionary (v.2025.06.25),[9] which provides a high coverage of FAERS. We further extended the dictionary through text normalization, which focuses on ingredient extraction. Specifically, ingredient names were parsed from entries such as “80% (propranolol)” using non-alphanumeric delimiters and then matched against known substances. This improved the consolidation of reports mapped to the same active ingredient, thereby increasing the accuracy of downstream signal counts.

For scalable computations, AEST leverages rays for distributed parallel processing.[12] FAERS data were partitioned into manageable segments, aggregations required for signal detection were

computed in parallel within each segment, and the intermediate results were merged to produce the final counts. In this study, rays were deployed in 20 vCPUs and 30 GiB environments and used to analyze 5,721 compounds.

AEST signals were evaluated using ROR and PRR with established thresholds. For ROR, a signal required a minimum report count of ≥2 and a lower bound of the 95% confidence interval >1.0. For PRR, a signal required a minimum report count of ≥3, PRR ≥2.0, and χ^2 ≥4.0. A drug–event pair was considered a signal if it met the criteria for either method.[13–15]

Chemical descriptor computation and chemical–AE signal similarity alignment analysis

We acquired chemical structure data via the PubChem[60] and Chemical Identifier Resolver (CIR)[61] using standardized drug names, yielding SMILES notations for 2,159 compounds. Two-dimensional molecular descriptors (696 features) were computed using Mordred (Mordred-community),[62,63] resulting in a 2,159 × 696 chemical descriptor matrix. Correspondingly, we constructed a 2,159 × 27,163 binary AE signal matrix (PT level), where each element denoted the presence (1) or absence (0) of a signal for the 27,163 PTs.

We employed the Aligned-UMAP[16] to embed these two high-dimensional matrices in a shared low-dimensional representation. The algorithm was configured with two embedding

dimensions: a Euclidean distance metric, alignment regularization strength of 20, 30 nearest neighbors, and a minimum distance of 0.1.

Embedding quality was evaluated using trustworthiness and continuity, which measure local neighborhood preservation within each modality.[17] These metrics were calculated using the default neighborhood size (k = 5), employing the Euclidean distance for the chemical descriptor matrix and the cosine distance for the binary AE signal matrix.

Parameter robustness was examined by systematically perturbing each model parameter by ±10% around its baseline value, including the alignment strength (alignment_regularisation), embedding neighborhood size (n_neighbors), a minimum distance (min_dist), and the evaluation neighborhood size (baseline values: 20, 30, 0.1, and 100, respectively). The stability of trustworthiness and continuity was assessed across all 81 parameter settings.

The chemical–AE signal alignment was quantified using two complementary measures: (i) the Euclidean distance between each compound's coordinates in the chemical descriptor embedding and the AE signal embedding (global), and (ii) the Jaccard overlap of the k=100 nearest-neighbor sets across the two embeddings (local). We then defined two thresholds using the elbow method applied to the cumulative distribution functions and classified the drugs into three groups: high-alignment (low embedding distance and high neighbor overlap), low-alignment (high embedding distance and low neighbor overlap), and intermediate (all remaining drugs).

To validate the separation of these groups in the original high-dimensional data, we performed PERMANOVA[18] and PERMDISP[19] using a chemical descriptor matrix and a binary AE signal matrix. We used the Euclidean distance for the continuous chemical descriptor matrix, which was z-score-normalized, and the Jaccard distance for the sparse binary AE signal matrix, with the significance assessed via 999 permutations. Multiple testing was controlled using the Bonferroni correction applied to a single family comprising all PERMANOVA and PERMDISP p-values (including all pairwise group comparisons).

OnSIDES integration and signal–label alignment quantification

Drug labels were obtained from the OnSIDES database (v3.1.0; May 2025 updated).[9] OnSIDES is an adverse drug event resource that is generated by extracting AE terms from drug labels using a fine-tuned PubMedBERT language model.[20] In this study, we identified 1,450 overlapping compounds between OnSIDES and our AEST compound list and included them in downstream analyses.

For each compound, we encoded the presence (1) or absence (0) of label-listed events across 27,163 PTs, resulting in a 1,450 × 27,163 binary label matrices. To quantify the signal–label alignment for each compound, we computed Yule's Q (Yule's coefficient of association) between the FAERS-derived signal vector and the corresponding OnSIDES label vector. Yule's

Q is derived from the odds ratio and is defined as:

$$Q = \frac{ad - bc}{ad + bc}$$

where a, b, c, and d denote the true positives (signal present/label present), false positives, false negatives, and true negatives, respectively. This measure ranges from −1 (complete negative association) to +1 (complete positive association) and provides a bounded association metric for binary contingency tables suitable for sparse, highly imbalanced adverse-event data. If the coefficient was undefined owing to a zero denominator, it was set to zero to represent a lack of association.

To stratify drugs by signal–label alignment, we modeled the univariate distribution of Yule's Q using GMMs.[21] We fitted the GMMs with k=1 to k=5 components and selected the optimal model based on the BIC.[22] To determine the optimal number of components, we used the elbow criterion on the BIC curve by identifying the point with the maximum perpendicular distance from the line connecting the BIC values at k=1 and k=5.

Four patterns risk stratification framework construction

Based on the relationship between chemical descriptors and signals, we stratified compounds into groups showing high versus low chemical–AE signal similarity alignment in UMAP embedding and validated group separation in the original high-dimensional space using

PERMANOVA and PERMDISP. Independently, we quantified the signal–label alignment between FAERS-derived signals and label-derived events using Yule's Q and classified drugs into high- versus low-alignment groups via GMMs of Yule's Q distribution. We then integrated these two binary classification axes and constructed a four-pattern placement for the subset of compounds with available chemical descriptors and OnSIDES labeling information (n = 875). To evaluate whether the therapeutic categories were nonrandomly enriched in specific quadrants, we performed statistical analyses using the ATC classification system. At ATC level 1, we tested the independence between quadrant membership and ATC categories using the chi-square test and assessed deviations from a random distribution. To identify categories that contributed most strongly to the observed association, we computed standardized residuals for each cell of the contingency table and interpreted ATC categories with absolute residual values greater than 2.0, as significantly over-represented (positive) or under-represented (negative) relative to expectations.

For ATC level 1 categories showing evidence of nonrandom quadrant enrichment, we repeated the chi-square analysis after stratifying to ATC level 3 to localize the contributing subgroups. To control for multiple comparisons across repeated subgroup tests, we adjusted the significance thresholds using Bonferroni correction and defined ATC level 3 subgroups as key drivers of

quadrant enrichment when Bonferroni-adjusted $p < 0.05$ and the absolute standardized residual exceeded 2.0.

Analysis of Mechanistic and Contextual Consistency Across Patterns

To further characterize the biological and clinical relevance of the proposed framework, we focused our analysis on Pattern 2 and Pattern 3.

Pattern 2 (high label alignment, low structural alignment) comprises drugs with high concordance with safety label information but low structural alignment, suggesting that risk expression may be influenced by biological factors not fully captured by chemical descriptors. As a representative mechanism, the involvement of active metabolites was examined. Pharmacogenetic association data were obtained from the table of pharmacogenetic associations created by the FDA,[23] which compiles gene variants or genotype-inferred phenotypes that affect drug metabolism or response based on scientific evidence. Analyses were restricted to Section 1 (pharmacogenetic associations for which the data support therapeutic management recommendations) and Section 2 (pharmacogenetic associations for which the data indicate a potential impact on safety or response). Each entry was screened for the explicit presence of the term “active metabolite” in the description, and counts were aggregated by pattern. As no drugs classified as Pattern 3 or Pattern 4 were represented in these sections, statistical comparison was

limited to Pattern 1 and Pattern 2. Differences in the proportion of reports mentioning active metabolites were assessed using a Fisher's exact test.

Pattern 3 (low label alignment, high structural alignment) comprises drugs with high structural alignment but low concordance with safety labels, potentially reflecting divergence between controlled-label evaluations and real-world multidrug treatment environments. Using outpatient prescription claims data,[24] the mean number of exact counts per drug was calculated for each pattern across strata defined by the number of concomitant medications. Counts reported as "<100" for privacy protection were excluded from the analysis. Values were normalized to the mean prescription frequency under monotherapy to assess relative changes in prescription frequency with increasing numbers of concomitant drugs. Given the limited sample sizes in some patterns, statistical testing was not performed, and trends were evaluated descriptively.

Statistical information

All statistical tests were two-sided unless otherwise specified. The unit of analysis was defined as an individual drug compound, and all n values represent the number of unique drugs included in each analysis.

For signal detection, a disproportionality analysis was conducted for 5,721 compounds, of which 3,110 were drugs with ATC codes and were used for downstream profiling. The

chemical–AE signal similarity alignment analysis included 2,159 compounds with both chemical descriptors and FAERS-derived binary AE profiles. Signal–label alignment was quantified for 1,450 compounds overlapping with the OnSIDES database. The final four-pattern risk stratification and ATC enrichment analyses were performed on a subset of 875 compounds with complete data across chemical, label, and AE signal information.

To validate group separation in the original high-dimensional spaces, PERMANOVA and PERMDISP were performed using 999 permutations. Multiple testing across all pairwise group comparisons for PERMANOVA and PERMDISP was controlled using Bonferroni correction, and adjusted P values were reported. Therapeutic category enrichment across the four patterns was assessed using chi-square tests of independence. To identify specific over- or under-representation, standardized residuals were calculated for each cell of the contingency table, and categories with absolute residuals greater than 2.0 were considered significantly enriched or depleted. For ATC level 3 subcategory analyses, P values were adjusted using Bonferroni correction to account for multiple comparisons. All statistical values, including pseudo-F statistics, P values, and Bonferroni-adjusted P values, are reported in the corresponding sections of the Results.

To further contextualize the proposed framework, additional statistical analyses were conducted with a focus on Pattern 2. Differences in the proportion of reports reporting active metabolites between Pattern 1 and Pattern 2 were assessed using a two-sided Fisher's exact test.

All statistical analyses in this study were performed using Python, including the SciPy, statsmodels, and Scikit-bio[64] libraries.

Data availability

All data used in this study are publicly available. FAERS data were obtained from the FDA website (https://www.fda.gov/drugs/fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files). The adverse event signal detection results generated in this study, as well as the AEST analysis framework, are accessible through the AEST web application (https://alain003.phs.osaka-u.ac.jp/AEST/). Drug safety label information from the OnSIDES database was obtained from its public repository (https://github.com/tatonetti-lab/onsides).

Code availability

The programming code is available from the authors upon reasonable request.

References

1. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *Br J Pharmacol* **162**, 1239–1249 (2011) https://doi.org/10.1111/j.1476-5381.2010.01127.x.

2. Tong, W., Welsh, W. J., Shi, L., Fang, H. & Perkins, R. Structure-activity relationship approaches and applications. *Environmental Toxicology and Chemistry* **22**, 1680–1695 (2003) https://doi.org/10.1897/01-198.

3. Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery* **18**, 495–496 (2019) https://doi.org/10.1038/d41573-019-00074-z.

4. Onakpoya, I. J., Heneghan, C. J. & Aronson, J. K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med* **14**, 10 (2016) https://doi.org/10.1186/s12916-016-0553-2.

5. Amorim, A. M. B. *et al.* Advancing Drug Safety in Drug Development: Bridging Computational Predictions for Enhanced Toxicity Prediction. *Chem. Res. Toxicol.* **37**, 827–849 (2024) https://doi.org/10.1021/acs.chemrestox.3c00352.

6. Ball, G. *et al.* A Framework for Safety Evaluation Throughout the Product Development Life-Cycle. *Ther Innov Regul Sci* **54**, 821–830 (2020) https://doi.org/10.1007/s43441-019-00021-5.

7. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **57**, 3186–3204 (2014) https://doi.org/10.1021/jm401411z.

8. FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files. *FDA* (2019) https://www.fda.gov/drugs/fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files.

9. Fusaroli, M. *et al.* Enhancing Transparency in Defining Studied Drugs: The Open-Source Living DiAna Dictionary for Standardizing Drug Names in the FAERS. *Drug Saf* **47**, 271–284 (2024) https://doi.org/10.1007/s40264-023-01391-4.

10. The HDF Group. The HDF5 Field Guide. https://support.hdfgroup.org/documentation/hdf5/latest/index.html.

11. Introduction to the Virtual Dataset - VDS. *The HDF Group Support Site* https://support.hdfgroup.org/documentation/hdf5-docs/advanced_topics/intro_VDS.html.

12. Moritz, P. *et al.* Ray: a distributed framework for emerging AI applications. in *Proceedings of the 13th USENIX conference on Operating Systems Design and Implementation* 561–577 (USENIX Association, USA, 2018).

13. Rothman, K. J., Lanes, S. & Sacks, S. T. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf* **13**, 519–523 (2004) https://doi.org/10.1002/pds.1001.

14. Fukazawa, C., Hinomura, Y., Kaneko, M. & Narukawa, M. Significance of data mining in routine signal detection: Analysis based on the safety signals identified by the FDA.

Pharmacoepidemiology and Drug Safety **27**, 1402–1408 (2018) https://doi.org/10.1002/pds.4672.

15. Evans, S. J., Waller, P. C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* **10**, 483–486 (2001) https://doi.org/10.1002/pds.677.
16. Dadu, A. *et al.* Application of Aligned-UMAP to longitudinal biomedical studies. *Patterns* **4**, 100741 (2023) https://doi.org/10.1016/j.patter.2023.100741.
17. Lee, J. A. & Verleysen, M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* **72**, 1431–1443 (2009) https://doi.org/10.1016/j.neucom.2008.12.017.
18. Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA). in *Wiley StatsRef: Statistics Reference Online* 1–15 (John Wiley & Sons, Ltd, 2017). doi:10.1002/9781118445112.stat07841.
19. Bakker, J. D. PERMDISP. (2024) https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/permdisp/.
20. Tanaka, Y. *et al.* OnSIDES database: Extracting adverse drug events from drug labels using natural language processing models. *Med* **6**, 100642 (2025) https://doi.org/10.1016/j.medj.2025.100642.
21. Reynolds, D. Gaussian Mixture Models. in *Encyclopedia of Biometrics* 659–663 (Springer,

Boston, MA, 2009). doi:10.1007/978-0-387-73003-5_196.

22. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2012) https://doi.org/10.1002/wics.199.

23. U.S. Food and Drug Administration. Table of Pharmacogenetic Associations. (2022) https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations.

24. Quinn, K. J. & Shah, N. H. A dataset quantifying polypharmacy in the United States. *Sci Data* **4**, 170167 (2017) https://doi.org/10.1038/sdata.2017.167.

25. Surarak, T., Chumnumwat, S., Nosoongnoen, W. & Tragulpiankit, P. Efficacy, safety, and pharmacokinetics of isoniazid affected by NAT2 polymorphisms in patients with tuberculosis: A systematic review. *Clin Transl Sci* **17**, e13795 (2024) https://doi.org/10.1111/cts.13795.

26. Dean, L. Carbamazepine Therapy and HLA Genotype. in *Medical Genetics Summaries* (eds Pratt, V. M. et al.) (National Center for Biotechnology Information (US), Bethesda (MD), 2012).

27. Dean, L. & Kane, M. Codeine Therapy and CYP2D6 Genotype. in *Medical Genetics Summaries [Internet]* (National Center for Biotechnology Information (US), 2025).

28. Dean, L. & Kane, M. Tramadol Therapy and CYP2D6 Genotype. in *Medical Genetics Summaries [Internet]* (National Center for Biotechnology Information (US), 2025).

29. Scherholz, M. L., Rao, R. T. & Androulakis, I. P. Modeling inter-sex and inter-individual variability in response to chronopharmacological administration of synthetic glucocorticoids.

Chronobiol Int **37**, 281–296 (2020) https://doi.org/10.1080/07420528.2019.1660357.

30. Mohammadnia, N. *et al.* Cardiovascular Benefit of Colchicine in Relation to Baseline Risk: A Secondary Analysis of the LoDoCo2 Trial. *J Am Heart Assoc* **14**, e038687 (2025) https://doi.org/10.1161/JAHA.124.038687.

31. Buckley, L. F. & Libby, P. Colchicine's Role in Cardiovascular Disease Management. *Arterioscler Thromb Vasc Biol* **44**, 1031–1041 (2024) https://doi.org/10.1161/ATVBAHA.124.319851.

32. Elias, M. *et al.* Perioperative utility of amisulpride and dopamine receptor antagonist antiemetics-a narrative review. *Front Pharmacol* **14**, 1274214 (2023) https://doi.org/10.3389/fphar.2023.1274214.

33. Zhang, H. *et al.* Generic Intravenous Amisulpride (QLG2069) for the Prevention of Postoperative Nausea and Vomiting in Adults: A Phase III, Multicenter, Randomized, Placebo-Controlled Study. *Drug Des Devel Ther* **19**, 7707–7718 (2025) https://doi.org/10.2147/DDDT.S529526.

34. Ganaxolone - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/neuroscience/ganaxolone.

35. Yasmen, N., Sluter, M. N., Yu, Y. & Jiang, J. Ganaxolone for management of seizures associated with CDKL5 deficiency disorder. *Trends Pharmacol Sci* **44**, 128–129 (2023)

https://doi.org/10.1016/j.tips.2022.11.007.

36. Fitch, W. L. *et al.* Complex Metabolism of the Novel Neurosteroid, Ganaxolone, in Humans: A Unique Challenge for Metabolites in Safety Testing Assessment. *Drug Metabolism and Disposition* **51**, 753–763 (2023) https://doi.org/10.1124/dmd.122.001218.

37. DailyMed - PAROMOMYCIN SULFATE capsule. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=333d55f0-ab49-459e-8d32-1590b2623ef3.

38. Hung, C.-C., Chang, S.-Y. & Ji, D.-D. Entamoeba histolytica infection in men who have sex with men. *Lancet Infect Dis* **12**, 729–736 (2012) https://doi.org/10.1016/S1473-3099(12)70147-0.

39. Melendez Rivera, J. G. & Hashmi, M. F. HIV Nephropathy. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

40. Rzewnicki, I., Olszewska, E. & Rogowska-Szadkowska, D. HIV infections in otolaryngology. *Med Sci Monit* **18**, RA17–RA21 (2012) https://doi.org/10.12659/MSM.882505.

41. Asirvatham, E. S. *et al.* A review of Tenofovir Disoproxil Fumarate associated nephrotoxicity among People Living with HIV: Burden, risk factors and solutions. *Clinical Epidemiology and Global Health* **25**, 101462 (2024) https://doi.org/10.1016/j.cegh.2023.101462.

42. Hara, M. *et al.* Atazanavir nephrotoxicity. *Clin Kidney J* **8**, 137–142 (2015)

https://doi.org/10.1093/ckj/sfv015.

43. Systemic poisoning in adults following skin absorption of pharmaceutical agents. *Journal of the American Academy of Dermatology* **76**, AB246 (2017) https://doi.org/10.1016/j.jaad.2017.04.957.

44. Church, D., Elsayed, S., Reid, O., Winston, B. & Lindsay, R. Burn Wound Infections. *Clinical Microbiology Reviews* **19**, 403–434 (2006) https://doi.org/10.1128/cmr.19.2.403-434.2006.

45. Abdominal Compartment Syndrome in Burn Patients: Not Always a Consequence of Excess Fluid Resuscitation. *ACS* https://www.facs.org/for-medical-professionals/news-publications/journals/case-reviews/issues/v3n8/vigneswaran-abdominal-compartment-syndrome/.

46. Law, R. M., Ngo, M. A. & Maibach, H. I. Twenty Clinically Pertinent Factors/Observations for Percutaneous Absorption in Humans. *Am J Clin Dermatol* **21**, 85–95 (2020) https://doi.org/10.1007/s40257-019-00480-4.

47. Kircik, L., Sung, J. C., Stein-Gold, L. & Goldenberg, G. United States Food and Drug Administration Product Label Changes. *J Clin Aesthet Dermatol* **10**, 20–29 (2017) https://www.ncbi.nlm.nih.gov/pubmed/28367259.

48. Commissioner, O. of the. FDA Requests Removal of All Ranitidine Products (Zantac) from the Market. *FDA* (2024) https://www.fda.gov/news-events/press-announcements/fda-requests-

removal-all-ranitidine-products-zantac-market.

49. You, S. C. *et al.* Ranitidine Use and Incident Cancer in a Multinational Cohort. *JAMA Netw Open* **6**, e2333495 (2023) https://doi.org/10.1001/jamanetworkopen.2023.33495.

50. Ryan, P. B. *et al.* Defining a reference set to support methodological research in drug safety. *Drug Saf* **36 Suppl 1**, S33-47 (2013) https://doi.org/10.1007/s40264-013-0097-8.

51. Macías, Y., Gómez Tabales, J., García-Martín, E. & Agúndez, J. A. G. An update on the pharmacogenomics of NSAID metabolism and the risk of gastrointestinal bleeding. *Expert Opin Drug Metab Toxicol* **16**, 319–332 (2020) https://doi.org/10.1080/17425255.2020.1744563.

52. Dean, L. Piroxicam Therapy and CYP2C9 Genotype. in *Medical Genetics Summaries* (eds Pratt, V. M. et al.) (National Center for Biotechnology Information (US), Bethesda (MD), 2012).

53. Niu, X. *et al.* CYP2C19 polymorphism and clinical outcomes among patients of different races treated with clopidogrel: A systematic review and meta-analysis. *J Huazhong Univ Sci Technolog Med Sci* **35**, 147–156 (2015) https://doi.org/10.1007/s11596-015-1404-7.

54. Alatawi, Y. M. & Hansen, R. A. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opin Drug Saf* **16**, 761–767 (2017) https://doi.org/10.1080/14740338.2017.1323867.

55. Hoffman, K. B., Dimbil, M., Erdman, C. B., Tatonetti, N. P. & Overstreet, B. M. The Weber Effect and the United States Food and Drug Administration's Adverse Event Reporting System

(FAERS): Analysis of Sixty-Two Drugs Approved from 2006 to 2010. *Drug Saf* **37**, 283–294 (2014) https://doi.org/10.1007/s40264-014-0150-2.

56. Giunchi, V., Fusaroli, M., Hauben, M., Raschi, E. & Poluzzi, E. Challenges and Opportunities in Accessing and Analysing FAERS Data: A Call Towards a Collaborative Approach. *Drug Saf* **46**, 921–926 (2023) https://doi.org/10.1007/s40264-023-01345-w.
57. Robin-Champigneul, F. Jeanne Calment's Unique 122-Year Life Span: Facts and Factors; Longevity History in Her Genealogical Tree. *Rejuvenation Res* **23**, 19–47 (2020) https://doi.org/10.1089/rej.2019.2298.
58. ISO - ISO 3166 — Country Codes. *ISO* https://www.iso.org/iso-3166-country-codes.html.
59. ICH Official web site : ICH. https://www.ich.org/page/meddra.
60. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Res* **53**, D1516–D1525 (2025) https://doi.org/10.1093/nar/gkae1059.
61. NCI/CADD Group Chemoinformatics Tools and User Services. https://cactus.nci.nih.gov/index.html.
62. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J Cheminform* **10**, 4 (2018) https://doi.org/10.1186/s13321-018-0258-y.
63. Burns, J. JacksonBurns/mordred-community. (2025).
64. Aton, M. *et al.* Scikit-bio: a fundamental Python library for biological omic data analysis. *Nat*

Methods 1–3 (2025) https://doi.org/10.1038/s41592-025-02981-z doi:10.1038/s41592-025-02981-z.

Acknowledgements

MedDRA® trademark is registered by ICH.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

H.A. and Y.S.T. contributed to the study conception and design; H.A. developed the AEST application and performed the analyses; H.A. and Y.S.T. wrote the first draft of the manuscript; Y.S.T., D.T., and K.F. provided methodological and technical support; A.H. and K.I. confirmed the results and provided suggestions to improve the analysis; All the authors have reviewed and accepted the final submitted manuscript.

Competing interests

All authors declare no competing interests.

Figures

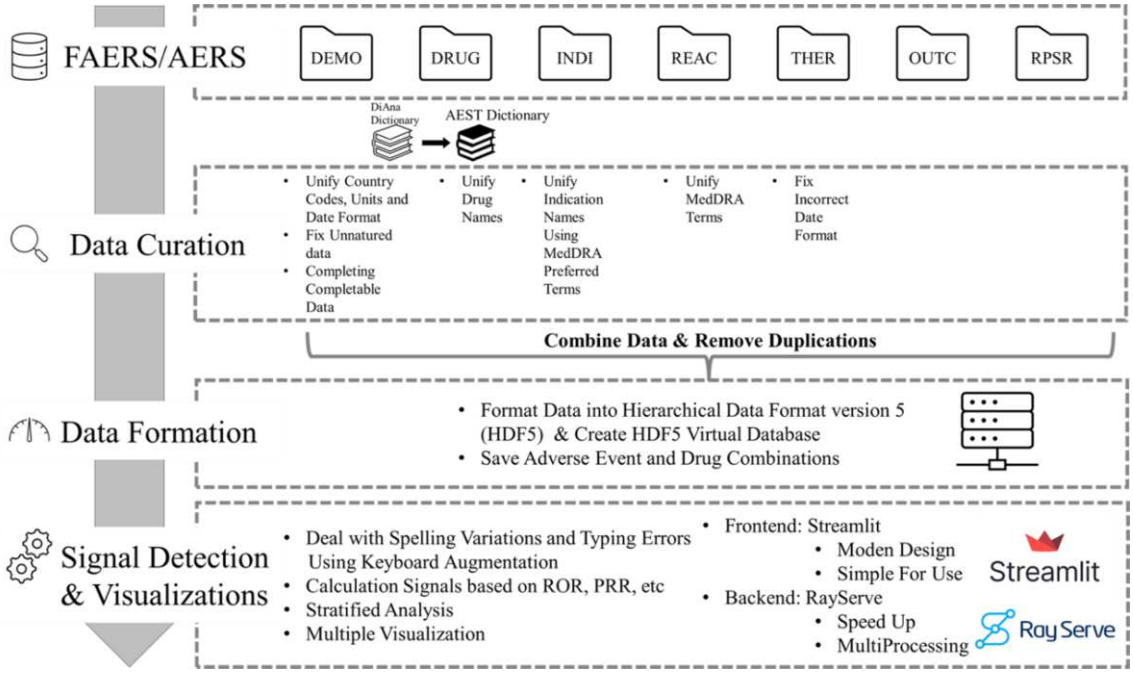


Figure 1. Data-processing pipeline and architecture of the Adverse Events Signal-detection Tool (AEST). The FAERS quarterly reports from Q1 2004 to Q2 2025 were assembled across relational tables (DEMO, DRUG, INDI, REAC, and THER) to construct a unified post-marketing safety resource. During data curation, records were cleaned through country code harmonization, standardization of measurement units (e.g., age and body weight), and unification of date formats; adverse event terms were mapped to MedDRA version 28.1 Preferred Terms (PTs). Drug names were standardized using the DiAna Dictionary and a custom ingredient-level normalization procedure to ensure robust integration of heterogeneous drug entries. After removing duplicate and temporally inconsistent reports, the curated data were organized in the data formation stage using a hierarchical HDF5-based architecture with VDS,

enabling efficient longitudinal access across quarterly partitions. In the signal detection and visualization stage, disproportionality analyses based on the ROR and PRR were conducted in parallel, and the resulting post-marketing safety signal profiles for drugs with valid ATC codes were made accessible through an interactive web-based visualization interface.

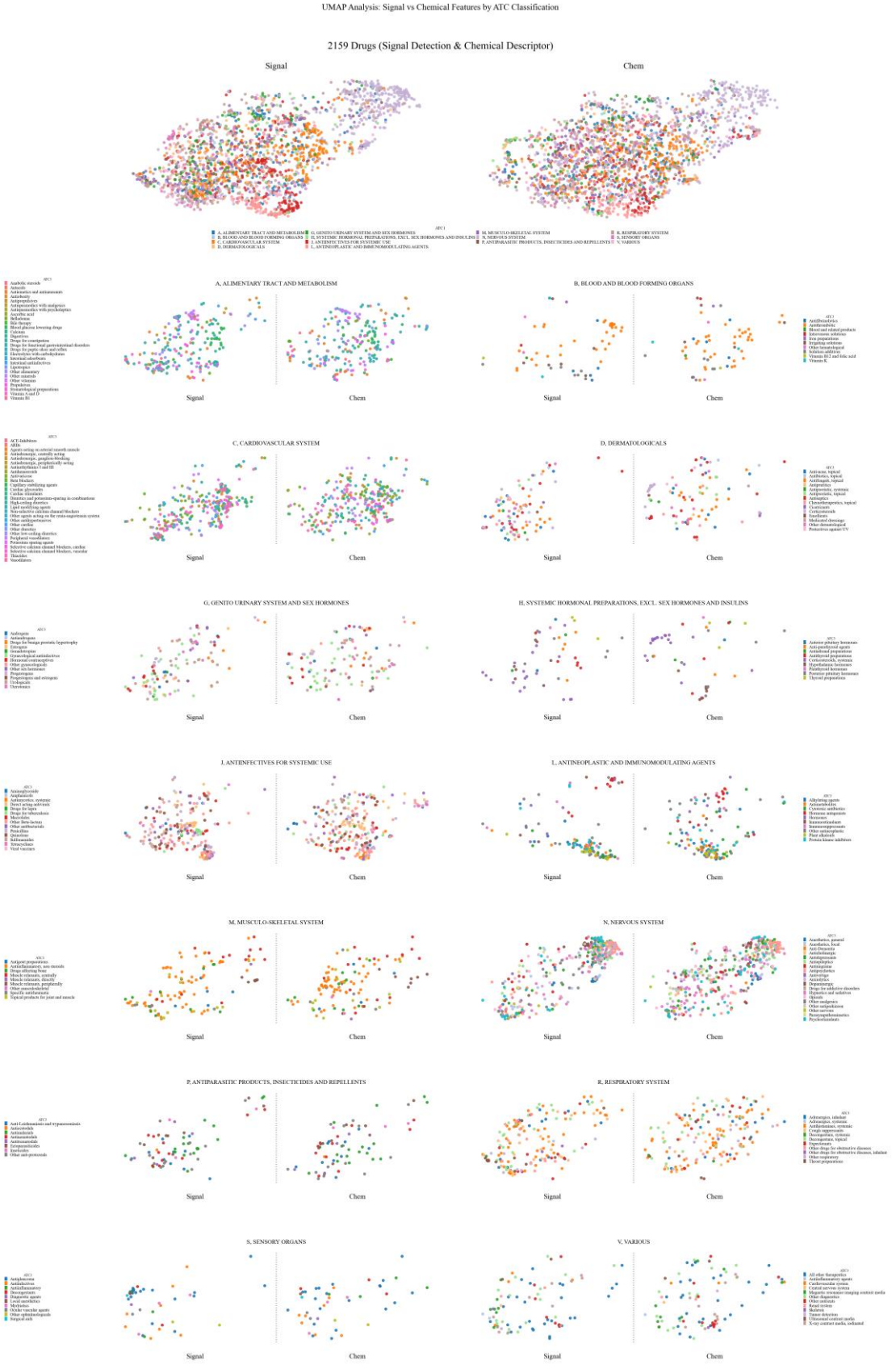
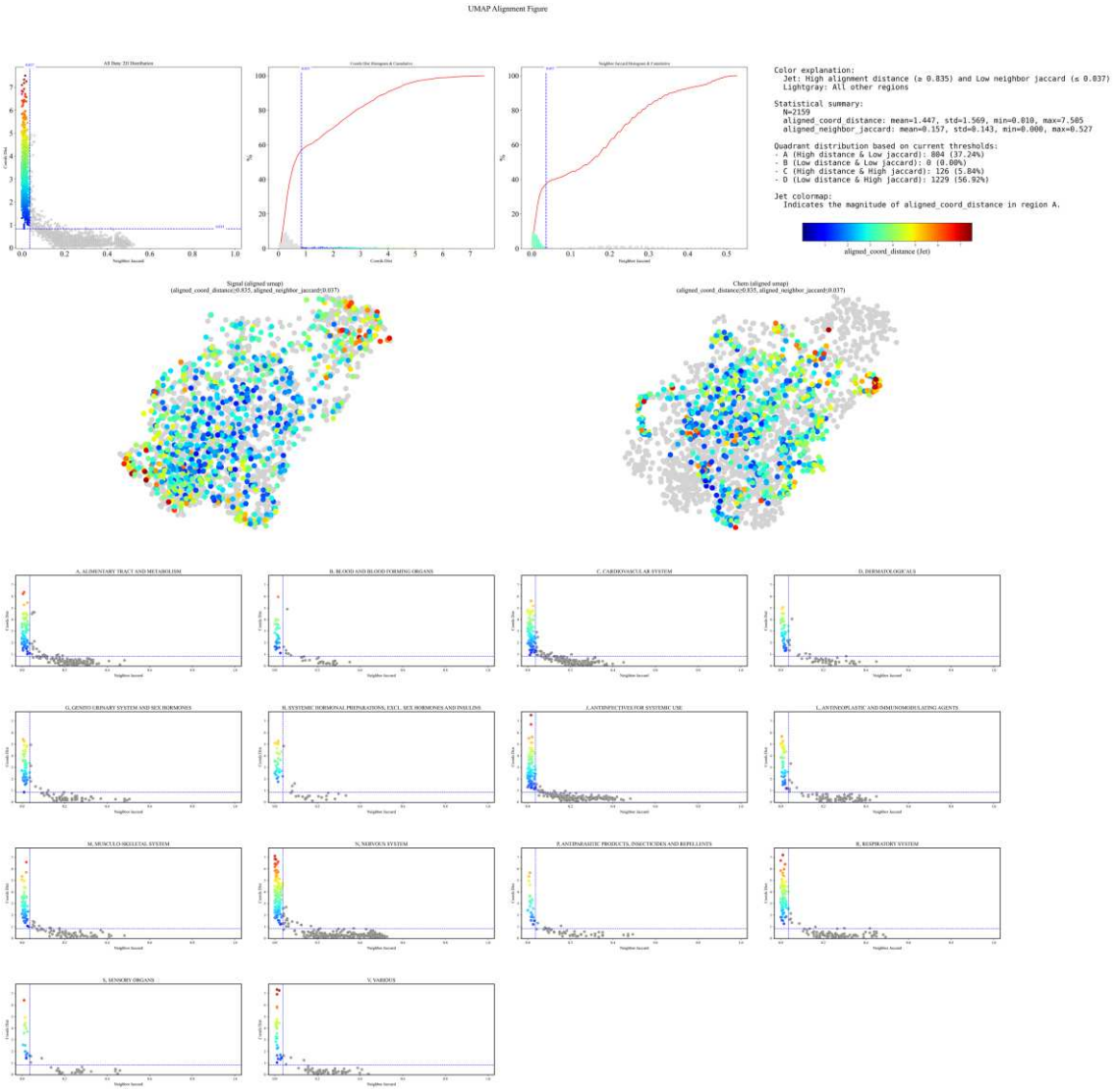

UMAP Analysis: Signal vs Chemical Features by ATC Classification
2159 Drugs (Signal Detection & Chemical Descriptor)
Signal
Chem
ATC1
A, ALIMENTARY TRACT AND METABOLISM
B, BLOOD AND BLOOD FORMING ORGANS
C, CARDIOVASCULAR SYSTEM
D, DERMATOLOGICALS
G, GENITO URINARY SYSTEM AND SEX HORMONES
H, SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS
J, ANTIINFECTIVES FOR SYSTEMIC USE
L, ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS
M, MUSCULO-SKELETAL SYSTEM
N, NERVOUS SYSTEM
P, ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS
R, RESPIRATORY SYSTEM
S, SENSORY ORGANS
V, VARIOUS

Figure 2. Visualization of the alignment between chemical similarity and adverse event (AE) signal similarity using Aligned-UMAP. Aligned-UMAP was used to embed PT-level binary adverse event (AE) signal profiles (2,159 × 27,163) and corresponding chemical descriptor profiles (2,159 × 696) of 2,159 drugs into a shared low-dimensional space. Each point represents a drug and is colored according to its ATC level 1 classification. To enable direct visual comparison between modalities, embeddings derived from post-marketing AE signals ("Signal") and chemical descriptors ("Chem") are displayed side-by-side. Preservation of local neighborhood structure in the aligned embeddings was confirmed by trustworthiness and continuity values of 0.70/0.66 for chemical descriptors and 0.71/0.73 for AE signals, respectively. The top panel shows the globally-aligned embedding across all therapeutic classes, whereas the subsequent panels present subsets stratified by ATC level 1 categories, along with corresponding ATC level 3 subgroups. Drugs occupying similar positions in both Signal and Chem embeddings indicate concordance between intrinsic chemical features and real-world post-marketing safety signal profiles.


UMAP Alignment Figure
Color explanation:
Jet: High alignment distance (≥ 0.835) and Low neighbor jaccard (≤ 0.037)
Lightgray: All other regions
Statistical summary:
N=2159
aligned_coord_distance: mean=1.447, std=1.569, min=0.010, max=7.505
aligned_neighbor_jaccard: mean=0.157, std=0.143, min=0.000, max=0.527
Quadrant distribution based on current thresholds:
- A (High distance & Low jaccard): 804 (37.24%)
- B (Low distance & Low jaccard): 0 (0.00%)
- C (High distance & High jaccard): 126 (5.84%)
- D (Low distance & High jaccard): 1229 (56.92%)
Jet colormap:
Indicates the magnitude of aligned_coord_distance in region A.
aligned_coord_distance (Jet)
Signal (aligned umap)
(aligned_coord_distance≥0.835, aligned_neighbor_jaccard≤0.037)
Chem (aligned umap)
(aligned_coord_distance≥0.835, aligned_neighbor_jaccard≤0.037)

3a

UMAP Analysis: Signal vs Chemical Features by ATC Classification



3b

Figure 3. Quantification of chemical–AE signal alignment and stratification of drugs. (**a**) Quantitative assessment of the alignment between chemical descriptor–based embeddings and adverse event (AE) signal–based embeddings using complementary local and global similarity metrics, together with their representation in aligned low-dimensional space. Local alignment was quantified using the distribution of Jaccard indices (k = 100), which measure the overlap of k-nearest neighbors between the chemical and AE signal embeddings, with higher values indicating a stronger correspondence of the local neighborhood structure across modalities. Global alignment was assessed by the distribution of Euclidean distances between the chemical and AE signal embedding coordinates for individual drugs, with smaller distances indicating a stronger overall correspondence between chemical and AE signal similarities. The cumulative distribution functions are shown for both metrics, and the vertical dashed lines indicate the thresholds determined by the elbow points (Jaccard index = 0.037, Euclidean distance = 0.835). Based on these cut-offs, drugs (n = 2,159) were classified into three groups: high-alignment (56.92%, n = 1,229; low Euclidean distance and high Jaccard overlap), low-alignment (37.24%, n = 804; high Euclidean distance and low Jaccard overlap), and intermediate alignment (5.84%, n = 126; remaining drugs). Alignment scores were visualized on aligned UMAP embeddings, where drugs were colored according to Euclidean distance using a continuous color scale,

directly linking quantitative global alignment to spatial organization in the embedding space. (**b**) Visualization of chemical AE signal alignment stratified by therapeutic class. Drugs with small Euclidean distances are shown in gray to provide a spatial context, whereas drugs with larger distances are highlighted and colored according to their ATC classification, which is consistent with the visualization scheme shown in Figure 2. The top panel displays the global-aligned embedding across all included drugs, whereas the subsequent panels show subsets stratified by ATC Level 1 categories. These panels reveal heterogeneous alignment patterns across therapeutic areas, including classes with strong correspondence between chemical and AE signal similarities, as well as classes with weak alignment. Drugs occupying similar positions in both embeddings indicated a concordance between the intrinsic chemical features and post-marketing safety signal profiles.

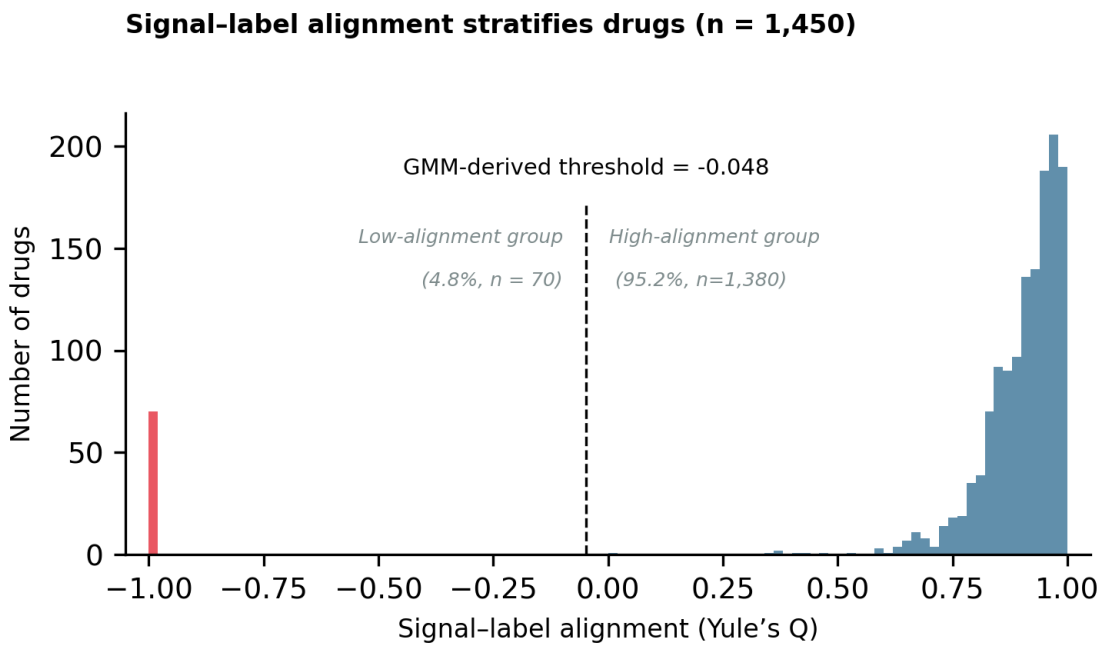


Figure 4. Stratification of drugs based on signal–label alignment using GMMs. The distribution of Yule's Q coefficients (red/blue bars) quantifies the association between FAERS-derived adverse event signals and OnSIDES-extracted label information across 1,450 drugs. Yule's Q values range from −1 to +1, where higher values indicate stronger alignment between real-world reporting and regulatory labeling. To objectively define a data-driven cutoff for stratification, GMMs with k = 1–5 components were fitted to the observed distribution. Bayesian Information Criterion (BIC) analysis favored a two-component solution. Based on posterior probabilities from the selected GMM, a threshold of −0.048 (vertical dashed line) stratifies the drugs into a high-alignment group (blue bars; 95.2%, n = 1,380) and a low-alignment group (red bars; 4.8%, n = 70).

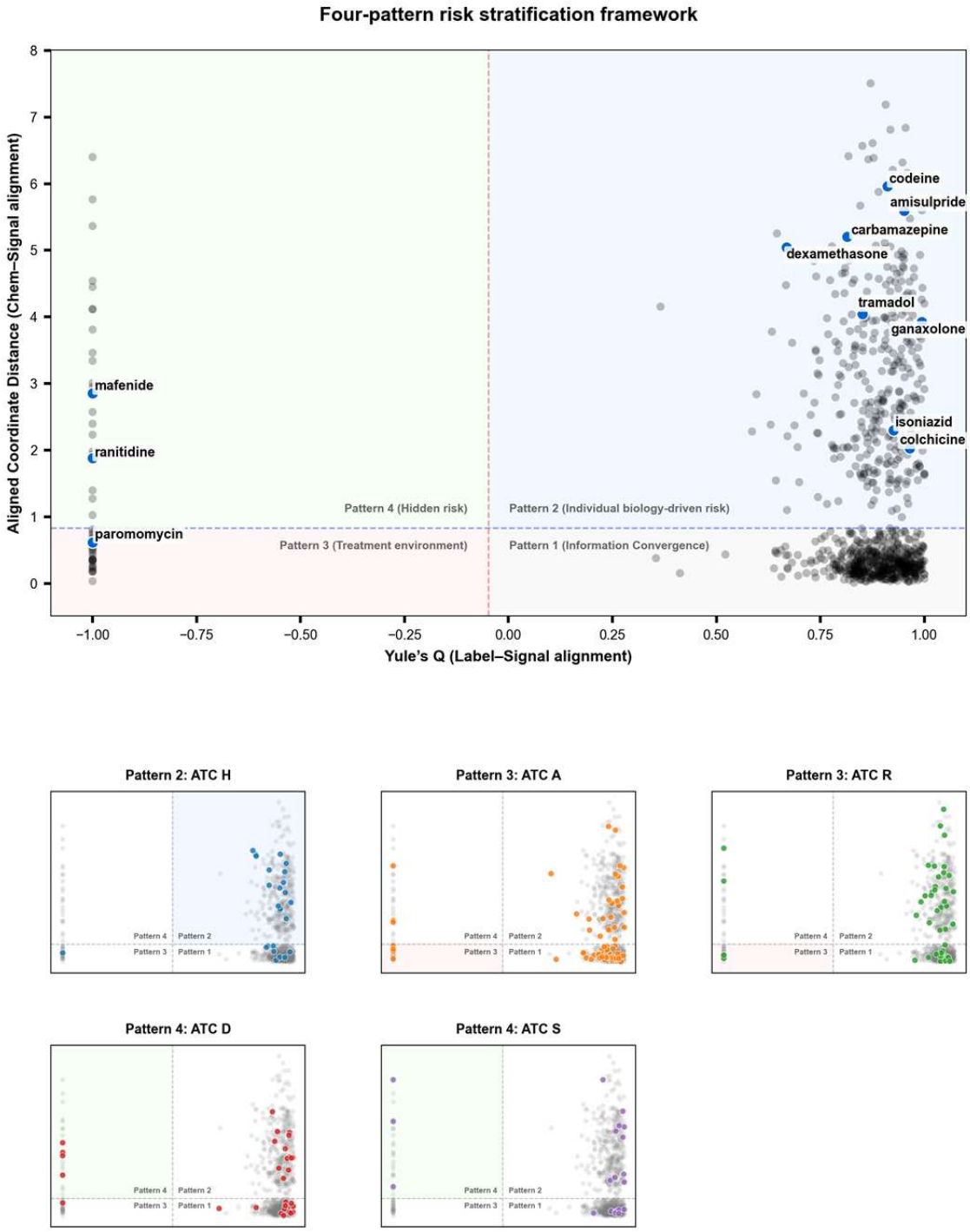


Figure 5. Integrated four-pattern risk stratification framework for drug safety information alignment. Building on the alignment metrics derived in Figures 3 and 4, this framework integrates structure-signal alignment and signal-label alignment for 875 drugs. The

x-axis represents the signal-label alignment (Yule's Q), where higher values indicate stronger correlation with established labeling, and the y-axis represents the structure-signal alignment (Euclidean distance in the aligned embedding space), where lower values indicate higher similarity between chemical descriptors and real-world safety signals. Horizontal (blue dashed line, Euclidean distance = 0.835) and vertical (red dashed line, Yule's Q = -0.048) thresholds divide the space into four distinct patterns: Pattern 1 (bottom-right: high-alignment in both dimensions, 56.4%, n=493), Pattern 2 (top-right: individual biology-driven risk, 38.4%, n=336), Pattern 3 (bottom-left: treatment environment, 2.7%, n=24), and Pattern 4 (top-left: hidden risk, 2.5%, n=22). The representative drugs for each pattern were labeled to illustrate their clinical relevance. Statistically significant enrichment of therapeutic categories was observed across patterns (Chi-square test, $p = 5.05 \times 10^{-6}$), as visualized in the bottom panels where points are colored by ATC level 1 class and quadrants with significant over-representation are highlighted for ATC H in Pattern 2, ATC A and R in Pattern 3, and ATC D and S in Pattern 4 (standardized residuals > 2.0).

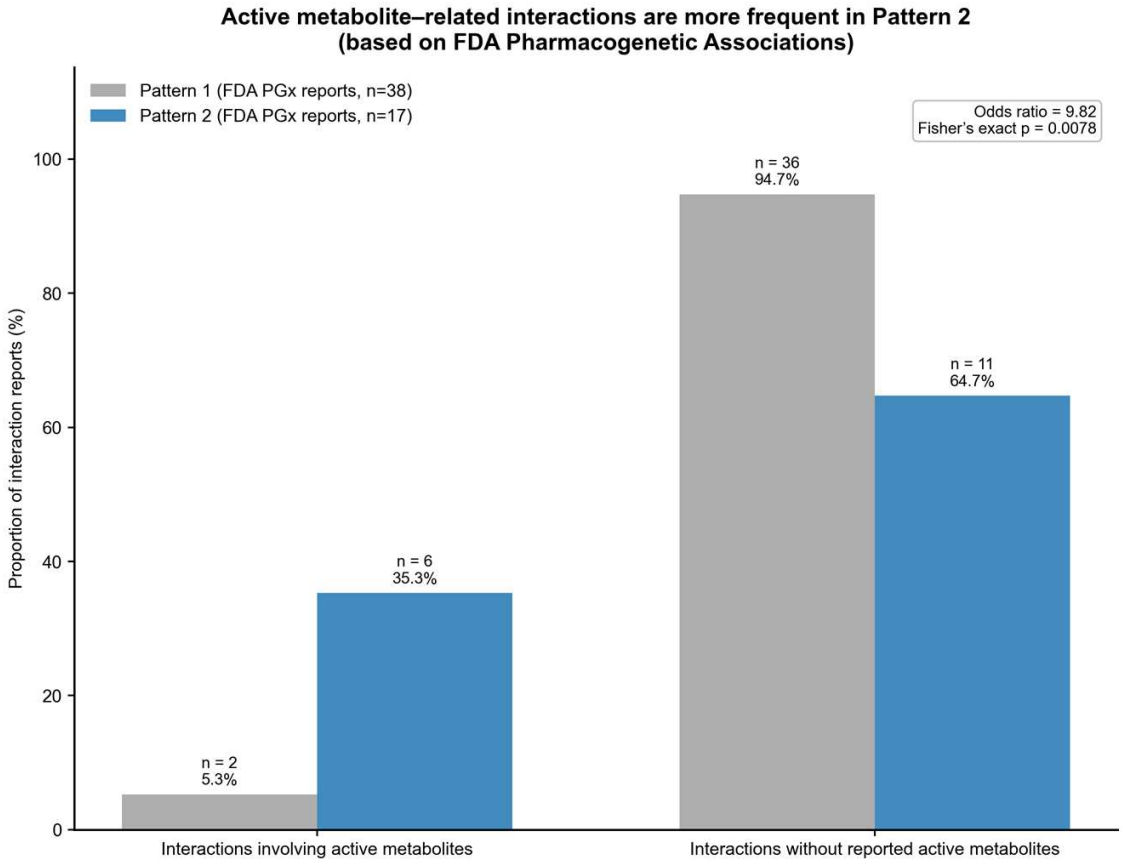


Figure 6. Enrichment of active metabolite–related pharmacogenetic associations in Pattern 2

Proportions of FDA pharmacogenetic association reports (Sections 1 and 2) that explicitly mention active metabolites are shown for Pattern 1 and Pattern 2. Reports involving active metabolites were more frequent in Pattern 2 (35.3%, 6/17) than in Pattern 1 (5.3%, 2/38), corresponding to an odds ratio of 9.82. Statistical significance was assessed using a Fisher's exact test ($p = 0.0078$).

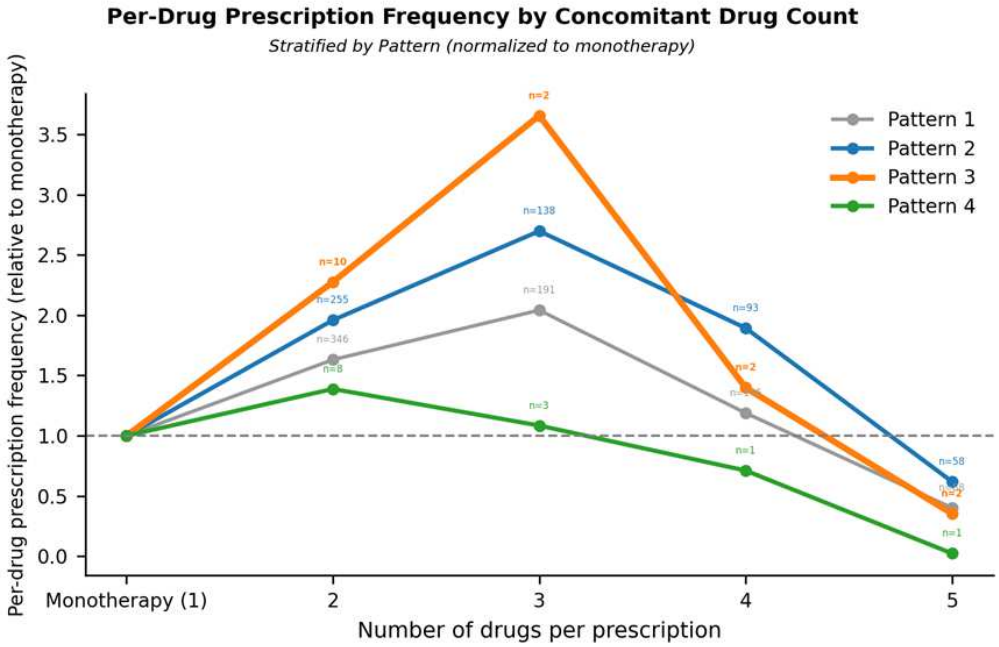


Figure 7. Mean prescription frequency across levels of concomitant medication, stratified by risk pattern.

The x-axis indicates the number of concomitant drugs per prescription. The y-axis represents the mean prescription frequency, calculated by averaging the number of prescriptions across drugs within each risk pattern for a given concomitant-drug stratum, and normalized to the corresponding monotherapy value. Drugs classified as Pattern 3 showed an increase in normalized mean prescription frequency with increasing numbers of concomitant medications, peaking at triple-drug regimens. The number of drugs contributing to each pattern is shown adjacent to the data points. Trends were evaluated descriptively, as sample sizes were smaller for some patterns.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables14.xlsx
- SupplementaryFig.4.png
- SupplementaryFig.2.png
- SupplementaryFig.6.png
- SupplementaryFig.5.png
- SupplementaryFig.3.png
- MSSupplementaryInformation.docx
- SupplementaryFig.1.png