

Contrastive learning of adverse events to provide effective and interpretable vector representations for machine-assisted pharmacovigilance

Olivér M. Balogh^{1,2,3,4}, Mátyás Pétervári^{1,2,3,5}, Áron M. Csernák^{1,2}, Eszter Puhl^{1,2}, András Horváth⁶, Péter Ferdinandy^{1,2,3,4}, Bence Ágg^{1,2,3,4,*}

¹Department of Pharmacology and Pharmacotherapy, Semmelweis University, Üllői út 26., Üllői út 26., H-1085 Budapest, Hungary

²Center for Pharmacology and Drug Research and Development, Semmelweis University, Üllői út 26., Üllői út 26., H-1085 Budapest, Hungary

³HUN-REN-SU System Pharmacology Research Group, Department of Pharmacology and Pharmacotherapy, Semmelweis University, Üllői út 26., H-1089 Budapest, Hungary

⁴Pharmahungary Group, Szeged, Hajnóczy utca 6., H-6722 Szeged, Hungary

⁵Sanovigado Kft, Somfa köz 11., Somfa köz 11., H-1107 Budapest, Hungary

⁶Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter utca 50/A., H-1083 Budapest, Hungary

*Corresponding author. Department of Pharmacology and Pharmacotherapy, Semmelweis University, H-1085 Budapest, Hungary. E-mail: agg.bence@semmelweis.hu

Abstract

Post-marketing surveillance is crucial for drug safety, yet the tools of pharmacovigilance rely solely on text-based data that may limit contemporary machine learning methodologies in the support of decision-making. With the recent surge of employing large language models (LLMs) for text-based tasks, there also arises an unmet need for a different approach which is not grounded in the linguistic patterns of unfiltered natural text, like LLMs, but rather based on real-world drug safety data. Here, we adapt contrastive learning algorithms to generate adverse event vector representations from spontaneous adverse event reports to serve as machine-readable (i.e. numerical) resources for downstream pharmacovigilance applications, such as drug–event association prediction for signal detection or causality assessment. We present comprehensive interpretability analyses of the resulting representations through density-based clustering, semantic evaluation, and comparison of multivariate dispersions, revealing patterns that reflect both functional and causal relations of the adverse events while also capturing drug-safety-related information better than existing medical terminologies and encoder-only LLMs. Furthermore, we demonstrate the applicability of our representations as input features in our downstream classifier model, outperforming the reporting odds ratio method, commonly used by regulatory agencies, and also LLM-generated representations (area under the receiver operating characteristic curve: 0.88 versus 0.76–0.83) on drug–event association prediction benchmarks. Therefore, we propose an interpretable adverse event vector representation, serving as a general resource that could enable the development of a wide array of machine learning applications to support decision-making in pharmacovigilance and facilitate patient safety.

Keywords drug safety, adverse drug reaction, embedding, causality assessment, interpretable ai

Introduction

Adverse events are any harmful medical occurrences experienced during the administration of a medicinal product. When suspicion is conveyed about a possible causal relationship between an adverse event and a product, the terms adverse effect, adverse reaction, or adverse drug reaction (ADR) are used instead [1].

The role of pharmacovigilance is to minimize the risk-to-benefit ratio of drugs via monitoring their adverse reactions throughout both the development and post-marketing phases. Several adverse reactions appear only during the latter, recorded in spontaneous

adverse event reports (AERs). AERs are collected for analysis by various pharmacovigilance entities, such as the EudraVigilance system of the European Medicines Agency (EMA) [2], the Adverse Event Reporting System of the United States Food and Drug Administration (FAERS) [3], and the VigiBase database of the World Health Organization [4]. AERs suggest only the potential for causality between drugs and adverse events (or equivalently in this case, adverse reactions), and thus they require careful and time-consuming causality assessment by experts [5] to distinguish coincidental observations from genuine ADRs before the appropriate regulatory action can be issued. Given the

Received: March 3, 2026. **Revised:** May 15, 2026. **Accepted:** July 28, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

sheer volume of AERs (e.g. more than 2 million received by FAERS in 2024 [6]), there is an increasing need for efficient automation methods to either improve the identification of new safety concerns from the reports (a process called signal detection) or to accelerate the causality assessment of already suspected drug–event pairs. Within our framework, we define this supporting task as “drug–event association prediction,” where a positive association corresponds to a relationship substantiated by clinical or regulatory evidence rather than strictly proven causal effects. Drug safety professionals have expressed their keenness to engage with artificial intelligence systems in order to shift their volume-based tasks to a value-based work [7–11], and so, our aim here is to provide an interpretable resource that could enable the application of a wide array of machine learning models for drug–event association prediction, facilitating machine-assisted pharmacovigilance.

A core limitation of handling AERs is their inherent text-based structure, lacking any informative, non-linguistic descriptor of the suspect drugs and the accompanying adverse events beyond their names, rendering contemporary machine learning models unreliable in providing evidence and supporting the decision-making of regulatory experts. For drug substances, several chemical and biological descriptors are available from external databases, such as DrugBank [12] or PubChem [13], that can provide numeric features for machine learning. There is, however, no similarly rich, machine-readable resource for adverse events. Representation learning could serve as a substitute by producing information-preserving embedding vectors; yet, to the best of our knowledge, no such data specifically made for general pharmacovigilance applications are readily available.

Representation learning of medical concepts is, however, not a new field of study. In terms of embedding style, we can identify: (i) methods with traditional word2vec-like approaches [14–22], (ii) methods where knowledge graphs or term hierarchies are also constructed, resembling graph embeddings [22–26], and (iii) methods that are based on pretrained encoder-only large language models (LLMs) [26–33]. Regarding the application of these models in pharmacovigilance, one of the most frequent obstacles is that they embed on the level of various Unified Medical Language System subsets [16, 17, 19, 20, 22, 26] with only a partial mapping to the preferred terms (PTs) of the Medical Dictionary for Regulatory Activities (MedDRA®), which is the most commonly used terminology standard in pharmacovigilance [34]. For static embeddings (i.e. word2vec-like and graph-based ones), this obstacle introduces either a complicated task of needing to translate between terminologies, or significantly limits the number of embedding vectors readily available. For contextual embeddings produced by encoder-only LLMs, this problem is bypassed through subword- or character-level tokenization, and the use of attention. However, LLMs are rarely utilized to provide term embeddings in a static fashion, as the methods to do so (e.g. context aggregation, decontextualization) have their own challenges and limitations, especially in terms of interpretability [26, 35, 36]. The second recurring aspect hindering the reuse of medical concept embeddings is that they are often learned through a very specific, non-generalizable task and are part of a larger end-to-end prediction model, with no description of their applicability in other downstream models. For example, encoder-only LLMs fine-tuned for language tasks, such as document classification, natural language inference, and named entity recognition, all belong to this category [27–33], as well as some of the word2vec-like models [18] and graph-based embeddings [23, 25]. The third notable limitation of reusing medical concept embeddings in

pharmacovigilance is that most of them utilize natural text as their input, ranging from PubMed abstracts and full text articles [16, 17, 20, 22, 28–32] to clinical notes and electronic health records [16, 17, 27, 30, 31] to even insurance claims and social media posts [16, 17, 19, 29], making AERs an underutilized resource for representation learning, despite their relevance in pharmacovigilance. Lastly, availability is also an issue as some embeddings have simply not been made public [15, 21].

Motivated by these limitations of previous studies, we created our own representation learning framework that is based on a data processing module (Fig. 1a), handling raw FAERS reports, and an embedding module (Fig. 1b) responsible for learning the vector representation, i.e. embedding, of adverse events (from here on referred to as reactions) on the level of MedDRA® PTs. We developed two models: one based on normalized temperature-scaled cross entropy [37] (here referred to as NTX), and one based on the negative sampling skip-gram version of word2vec [38] (here referred to as NSG). In both cases, the embeddings are learned via noise-contrastive estimation [39] for which the reactions are first paired up with their context terms as observed in the AERs. We defined three different context-sampling approaches based on what is considered as context for a given “target” reaction: only drugs it was reported with (reaction2drug), only the other reactions it was reported with (reaction2reaction), or both (reaction2all), resulting in six different embedding sets overall. Comparing this framework to learning embeddings from natural text, the reports are treated similarly to the sentences of a corpus, and our context-sampling approaches function as skip-gram sampling methods with “dynamic” context window, operating within only the current sentence. As such, in our framework the distribution of context terms is influenced by the drug-safety-related behavior, i.e. reporting patterns, of the given reaction, rather than its occurrences in written language. Consequently, the resulting embeddings encode the pharmacovigilance profile of reactions, instead of their linguistic characteristics, which makes these embeddings not only effective descriptors for downstream applications, such as for machine learning models performing drug–event association prediction, but also semantically interpretable through various analysis methods.

Materials and methods

For more details on the following subjects, see the extended methods description in Section 1.1–1.12 of the Supplementary.

Data sources

As outlined in Fig. 1a, raw AERs were accessed in XML format from the publicly available quarterly data extracts of the FAERS database, from the interval 2012/Q4–2023/Q1, totaling >15 M AERs [6]. Given that drug names are not standardized in FAERS, multiple dictionaries (DrugBank [12], RxNorm [40], and the Anatomical, Therapeutic and Chemical classification (ATC) [41]) were employed to identify the canonic DrugBank substance name of each drug, when possible, enabling us to later trivially link them to their molecular attributes in DrugBank [12] and PubChem [13] for the downstream drug–event association prediction model. As for reaction names, we used MedDRA®, which is an extensive, international medical terminology for the support of regulatory communication and data analysis in pharmacovigilance [34]. Validated (positive only) drug–ADR associations for dispersion analysis and training data were collected from three

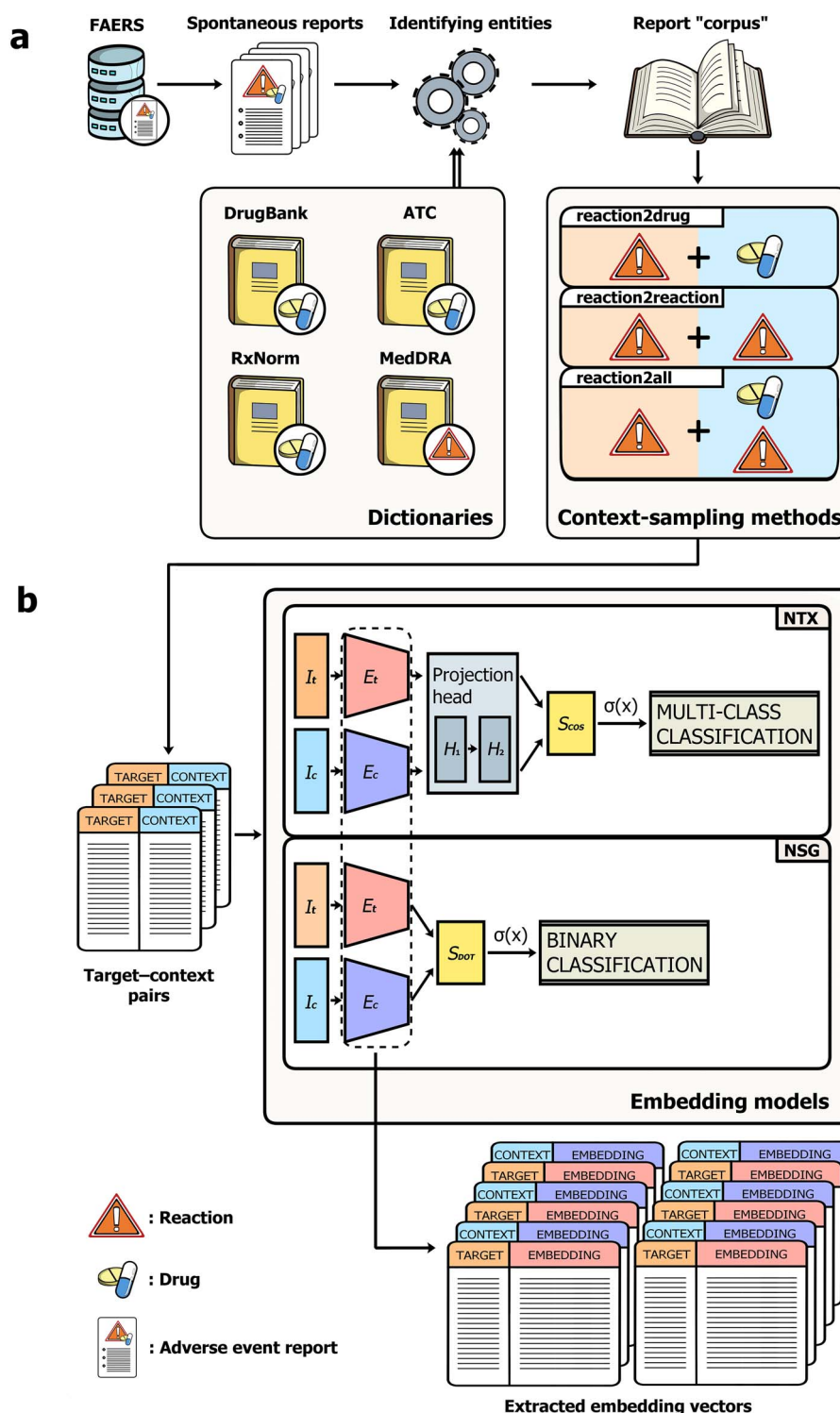


Figure 1 Schematic visualization of the main framework; (a) data processing: we accessed spontaneous AERs from the FAERS database, and used a combination of various naming dictionaries to identify the suspect drugs and reactions, mapping them into canonic DrugBank substance names and MedDRA® PTs, respectively; this step also acts as a quality filter, excluding poorly made reports; we performed three different context-sampling methods on the resulting report "corpus," based on what is considered as context for a given reaction: the drugs it was reported with (reaction2drug), the other reactions it was reported with (reaction2reaction), or both (reaction2all); (b) embedding: the three types of target–context pairs were utilized as input data for our embedding models, which learn the vector representation, i.e. embedding, of the reactions through two different contrastive learning algorithms; the NTX model is based on the normalized temperature-scaled cross entropy loss, and performs multiclass classification on batches of target–context pairs, where each target is expected to be classified as belonging only to its corresponding context and vice-versa; the NSG model uses negative sampling, and performs binary classification on the target–context pairs (positive class) that are complemented with randomly sampled negative pairs (negative class); in both cases, the embedding models are discarded after training, and only their embedding matrices are kept for further tasks and analysis.

sources: The Side Effect Resource (SIDER) [42], the “PROTECT” project of the Innovative Medicines Initiative (IMI-PROTECT) [43], and the British National Formulary (BNF) [44]. Benchmark (positive and negative) drug–ADR associations for performance testing of the association prediction model were collected from another three sources: Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge (EU-ADR) [45], Observational Medical Outcomes Partnership (OMOP) [46], and FDA revisions from the MedWatch website (FDA-rev) [47].

Data preprocessing

We selected AERs only with the report type of “spontaneous,” which corresponds to those made by healthcare professionals, patients, and pharmaceutical companies, but excludes studies. The deduplication and the exclusion of AERs with no identifiable drugs/reactions resulted in about 10 M reports remaining in the report “corpus.” For the embedding models, AERs were reduced to simple lists of target–context pairs, where the target term is always a reaction and the associated context term is determined by the type of context-sampling (Fig. 1a): only drugs in the same AER (reaction2drug), only other reactions in the same AER (reaction2reaction), or both (reaction2all). The three validated drug–ADR datasets and the three benchmark sets were subjected to the same drug and reaction identification processes as the AERs, then merged separately, eliminating duplicate entries, to serve as the basis for a training and a test dataset for association prediction, respectively.

Embedding

Contrastive learning can be considered as a self-supervised approach, where the labels are not provided by an external source of positive–negative ground truth samples but rather determined in the training data internally, by a chosen association, like the spatial proximity or semantic similarity of the unlabeled samples [48]. Both our models (Fig. 1b) are a form of noise-contrastive estimation, where the associated word pairs (positive label) are contrasted with noisy random pairs (negative label). In particular, one of the models is inspired by the negative sampling skip-gram version of word2vec from natural language processing [38], here referred to as NSG, which treats co-occurring words as the positive samples and randomly selected non-co-occurring ones as negative samples, leading to a binary classification problem in practice. The other model is based on the normalized temperature-scaled cross entropy loss from computer vision [37], here referred to as NTX, which handles co-occurrence as a multi-class classification problem, expecting each word to be classified as belonging to only its current context out of all the other words present in the same batch. The NSG model training was done for 30 epochs (embeddings extracted after every fifth), using Root Mean Square Propagation (RMSProp) with learning rate $\eta = 0.01$, smoothing parameter $\alpha = 0.75$, subsampling parameter $t = 0.00001$, number of sampled negatives $k = 7$, and eight positive target–context pairs per batch, thus a total batch size of 64. The NTX model training was done for 1000 epochs (embeddings extracted after every 100th), using RMSProp with learning rate $\eta = 0.01$, temperature parameter $\tau = 0.3$, and a batch size of 16.

For performance comparison, reaction embeddings were also created via various pretrained BERT-based models [30–33] by providing them as input the same list of reaction terms as used for the reaction2all embeddings (largest coverage). Each term was embedded separately, like a sentence, and mean pooling was applied to

generate single static embeddings, which is a method known as de-contextualization [35]. It should be noted that the other option for creating static embeddings with encoder-only LLMs, namely context aggregation, could possibly lead to a stronger baseline by exploiting richer contexts. However, this method was shown to underperform as scaling the amount of context used can quickly saturate the embeddings and eliminate all preserved information [36]. Consequently, a rigorous fine-tuning would be required to determine how and which reports should be utilized for providing the best possible context for each LLM with each reaction term, which we leave for a possible future study.

Interpretability analyses

To characterize the interpretability of the reaction embeddings, we conducted three different analyses, consisting of visualizations (Fig. 2a), semantic analyses (Fig. 2b), and statistical evaluations (Fig. 2c).

Visualizing the embeddings enables us to gain a general intuition about the underlying data structures and so, with the inherent limitations kept in mind, we applied *t*-distributed stochastic neighbor embedding (*t*-SNE) [49] to reduce the original representations into 2D. Supporting our visualization are the clusters deemed most prominent in this 2D space by the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [50], which we chose for its robust performance on variable densities and ability to prune off noisy elements. By performing clustering after dimension reduction, small patterns might emerge that are mere artifacts of the *t*-SNE optimization process; however, larger, well-separated clusters are still provably and consistently retrieved [51]. In our case, the resulting clusters were further examined according to their ratios of highest level MedDRA® categories, called System Organ Classes (SOCs), which are specific terms assigned to the reaction PTs (one primary SOC term to each).

Semantic analysis was conducted by manually assessing the pathophysiological connections and clinical relevance of the closest neighboring reactions of a selected MedDRA® PT, “Myocardial Ischemia,” based on several international medical guidelines, categorizing them into one of four distinct semantic groups with MedDRA® SOC categorization as reference. Group 1 consists of synonyms, such as “Ischemia,” and the various signs/symptoms considered during the diagnosis of “Myocardial Ischemia” itself, like “Angina Pectoris” (chest pain) and “Electrocardiogram Abnormal.” Group 2 contains the most neighbors, corresponding to various pathophysiological risk factors, like “Arteriosclerosis Coronary Artery” (partial or complete obstruction in the coronary arteries) or “Sinus Tachycardia/Bradycardia” (heart beats faster/slower than normal), and possible complications of “Myocardial Ischemia,” such as “Arrhythmia” (rhythm abnormalities) or “Cardiac Failure” (impaired pump function). For Group 3, we selected reactions that share their pathophysiological cause with “Myocardial Ischemia,” and have identifiable common risk factors other than old age. For example, both “Peripheral Artery Disease” and “Myocardial Ischemia” can be caused by arteriosclerosis, with diabetes mellitus, hypertension, and dyslipidemia (abnormal levels of lipids in the blood) being well-known risk factors. Finally, Group 4 consists of reactions with semantic connections that are too indirect to be relevant. See [Supplementary Table S8](#) for further details on the sources used for these categorizations.

At the core of the statistical evaluation was a permutation test for homogeneity of multivariate dispersions that enables the use of arbitrary distance measures, and calculates group dispersion by

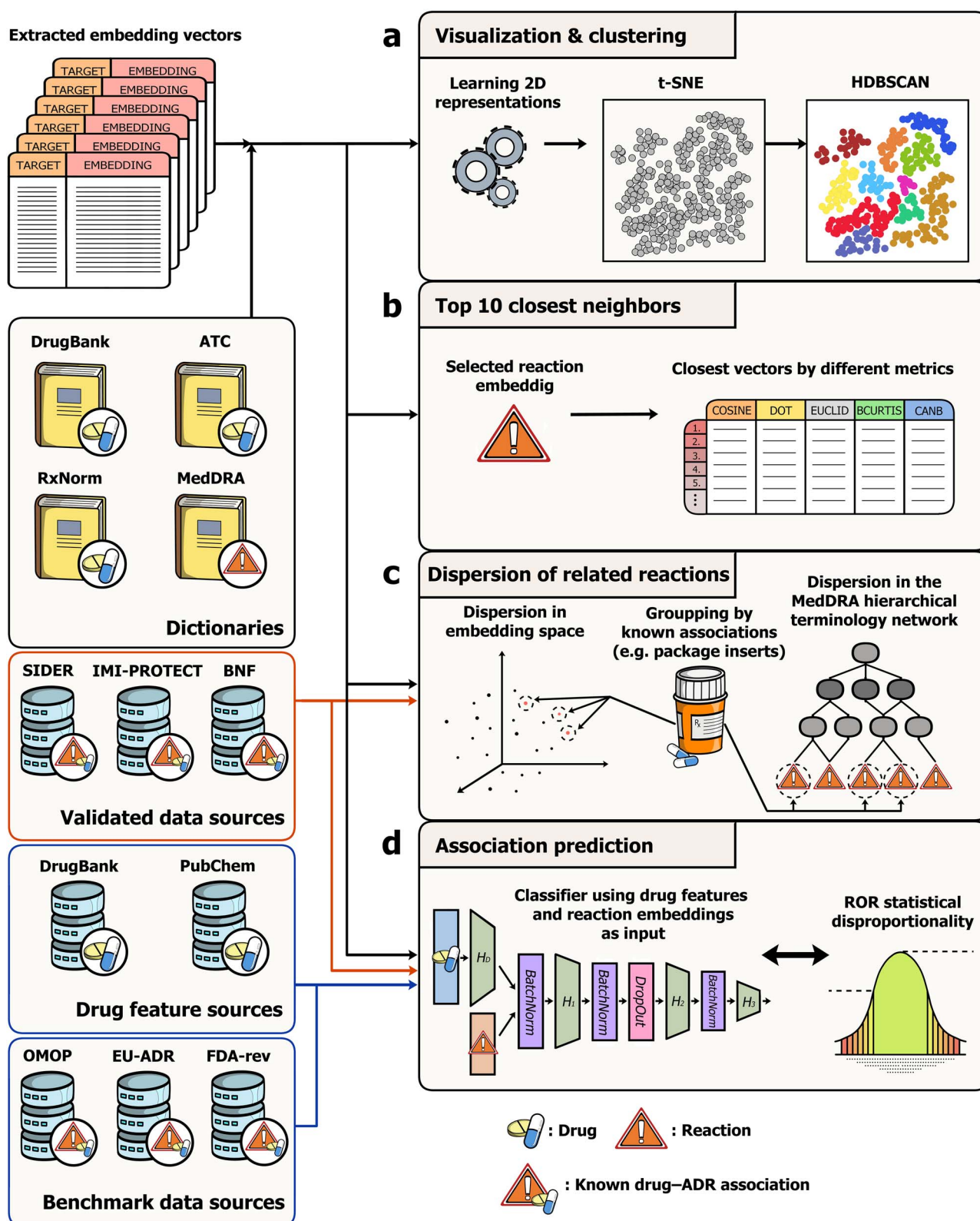


Figure 2 Summary of our evaluation methods and the data sources used to characterize the interpretability and applicability of the reaction embeddings; (a) we applied t-SNE for dimensionality reduction and HDBSCAN clustering to support the visualization by discovering potentially meaningful clusters of reactions; (b) by calculating various distance and similarity measures in the embedding space, we evaluated the semantic similarities of top 10 closest neighbors of a selected reaction across the representation spaces; (c) using validated drug-ADR data, we analyzed the multivariate dispersion of embedding vectors for reactions that belong to the same drug; dispersions of validated groups were statistically tested against randomly selected groups, and also compared to the dispersions measured in the MedDRA[®] hierarchical terminology network; (d) we constructed a deep neural network classifier to utilize the embedding vectors as input features, along with drug features from DrugBank and PubChem, for drug-event association prediction; the model was trained on the validated drug-ADR sources with various negative sampling approaches applied, and its performance was evaluated on established benchmark sets, comparing it with using embeddings generated by LLMs and the ROR statistical disproportionality analysis method.

considering the distances of samples from the estimated group center [52]. As described by Gijbels and Omelka [53], the test can be modified to use pairwise interpoint distances instead. This version of the test was not only used directly, but it also served as an inspiration for employing another dispersion difference measure [54], here simply termed “gain norm.” Calculation of the gain norm involves the difference of [0, 1]-normalized average pairwise interpoint distances, similar to the modified homogeneity test. With this measure, we evaluated how well reactions that are caused by the same drug separate from the background noise in the embedding space, as well as in the MedDRA hierarchical terminology network, enumerating the preserved drug-safety-related information.

Association prediction model

To demonstrate the real-life applicability of the reaction embeddings, we introduced a drug–event association prediction model, with a relatively simple deep neural classifier architecture (Fig. 2d). The input of this model consists of the embedding vector of the given reaction as well as the so-called chemical and biological feature vectors of the given drug [55]. The output, that is produced by sigmoid activation and optimized against the expected label through binary cross entropy loss, can be considered as the probability of the given drug and reaction being associated, either causally, functionally, or on a regulatory basis. Consequently, we emphasize that in the context of pharmacovigilance data, this association may also reflect strong co-reporting patterns and correlating regulatory warnings rather than strictly proven mechanistic causality, which is often subject to confounding and reporting biases. For the training data, we employed the validated drug–ADR associations, treating them as positive samples, and created negative pairs via various negative sampling (i.e. random recombination) techniques. Namely, we defined six methods according to which constituent of the positive drug–ADR pairs to keep (drug, reaction, none) and whether to sample the other constituent uniformly from the available terms or follow their frequency distribution as observed in the positive pairs. Hyperparameter search was performed in a 0.7–0.3 train-test split fashion. After retraining the final model on all data, performance testing was conducted using the withheld benchmark drug–ADR associations, averaged across 10 trainings. Performance was measured via area under the receiver operating characteristic curve (AUROC), as well as accuracy, and area under the precision–recall curve (AUPRC). Model training was done for 50 epochs, using stochastic gradient descent with learning rate $\eta = 0.001$ and a batch size of 512.

Software and hardware

Data preprocessing was implemented in C++ 17 standard, while model training and result analysis were done using Python (version 3.10). The contrastive embedding and classifier models were implemented using Tensorflow/Keras (version 2.11) [56]. We used the scikit-learn (version 1.3) [57] implementations of *t*-SNE and HDBSCAN for dimensionality reduction and clustering analysis, respectively, and plotly (version 5.19) [58] for visualization. BERT-based pretrained LLMs were accessed from Hugging Face through the transformers package (version 4.48) [59], and used to generate embeddings (inference only) with Python (version 3.12) and PyTorch (version 2.6) [60]. Model trainings were performed on an NVIDIA A100 Tensor Core GPU, taking 1–3 h for the NTX models and 46–221 h for the NSG models, depending on the context-sampling method used. This marks a notable advantage in our NTX model as compared to the NSG model,

in terms of efficiency. Classifier training took 1–2 h per run, depending on the negative sampling methods used.

Results

Visualization and clustering

For a deeper analysis, we selected the NSG-reaction2all embedding, which was the best performing representation in the downstream drug–event association prediction task (Fig. 7). See [Supplementary Figs S1–S5](#) for the others. The reaction2all-based embeddings have a perspicuous explanation, i.e. vectors will be close to each other for reactions that were reported with the same or similar drugs and reactions. For the NSG models specifically, a logarithmic frequency-based coloring (Fig. 3a) reveals a further, yet evident property: during training, the model encountered the often reported reactions much more frequently, even with subsampling applied, thus learning embedding vectors for them that are a lot more refined and pulled farther away from the randomly initialized values. We can observe the most frequently reported reactions congregating in the right side of the plot and on the right-facing edges of many of the visually well-separated clusters. These are general terms, such as “Drug Ineffective,” “Fatigue,” “Nausea,” “Pain,” “Headache,” and “Off Label Use.” At the left and left-center side of Fig. 3a, we can also observe a large, elongated group with low frequency, lacking a clear set of frequent reactions to gravitate toward. This group holds embedding vectors that had such a low occurrence during training that they barely moved from the noise initialization. Here, we have a diverse repertoire of niche reactions, such as “Vaccine Breakthrough Infection,” “Benign Rolandic Epilepsy,” “Nicotinamide Decreased,” “Mechanic’s Hand,” “Perthes Disease,” and “Coccidioides Encephalitis.”

Small-scale density differences are not preserved by *t*-SNE, but with the elimination of noise, HDBSCAN still proved successful in highlighting well-separated clusters in the dimension-reduced space for further analysis (Fig. 3b). For example, we can examine the composition of highest level MedDRA® categories, i.e. SOCs, in the resulting clusters to get an approximate idea of how closely the embedding space resembles them (Fig. 3c). Given that there are 27 distinct SOC terms, we selected only the top three with the highest ratios (with respect to the size of the cluster) in each cluster, and chose 15 clusters out of 161 in total, focusing on some of the larger and most representative ones. The first and second clusters are almost homogeneous, filled with reactions related to tumors (neoplasms) and blood tests, respectively, while the other clusters are more heterogeneous, with a pronounced plurality class. This is unsurprising, as reactions related to each other on the basis of prescribed therapies (and evidently, anatomical considerations) are often delegated into different SOCs, which is a phenomenon most prevalently caused by the umbrella SOCs, like those relating to investigations, injuries, infections, and procedures. For example, “Corneal Erosion” is categorized by MedDRA® SOCs as being primarily an eye disorder, while “Injury Corneal” is an injury and “Corneal Transplant” is a surgical procedure. Yet, the strong relation of these terms is reflected by the embedding space, with all three appearing close to each other in the sixth cluster, despite the embedding model lacking word and subword information (e.g. corneal) due to the term-level “tokenization.” This shows that the embedding space has a clear, interpretable structure that deviates from MedDRA® categories, but various anatomically and functionally consistent groups still emerge on the macro-scale. From the analysis of the NSG-reaction2drug embedding, which showed the highest gain

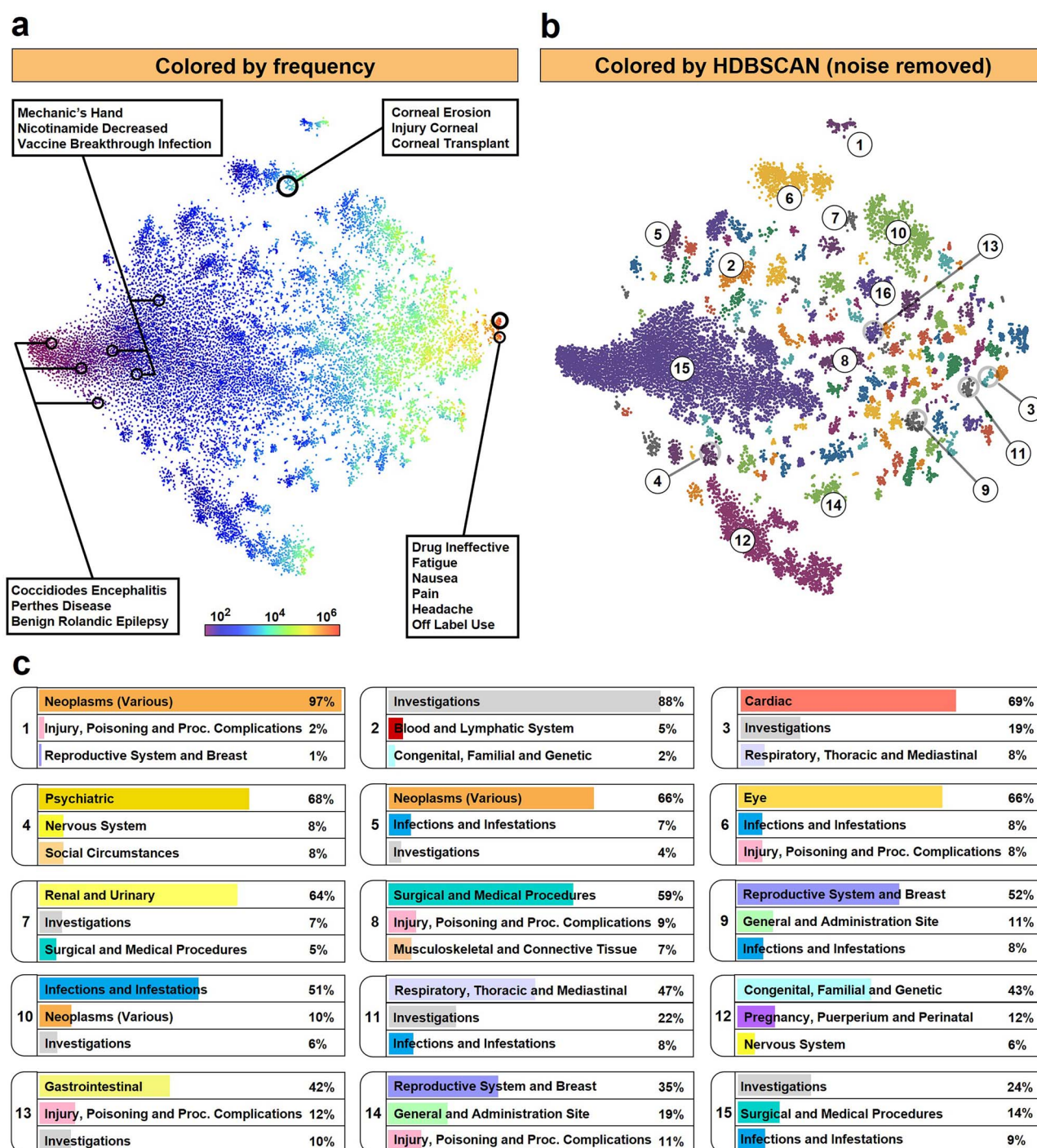


Figure 3 T-SNE visualization of the embedding space generated by the NSG model with reaction2all context-sampling; (a) frequency-based coloring of the *t*-SNE plot, using occurrence counts from the processed AERs with a logarithmic scale for smoother color transitions; example reactions discussed in section Visualization and clustering have their positions indicated; (b) HDBSCAN-based coloring of the *t*-SNE plot, with noise removed for a better visual separation; (c) ratio of the top three MedDRA® SOC in each selected cluster; selected clusters are numbered in descending order of their highest SOC ratio; SOC names are shortened to fit the figure by leaving out the words “disorders” and “conditions,” as well as the list of neoplasm types.

norm values during the statistical evaluation (Fig. 5), almost identical conclusions can be drawn (Supplementary Fig. S1).

Top 10 closest neighbors

Here we present a brief semantic analysis through the example of a MedDRA® PT, “Myocardial Ischemia,” and its 10 closest neighbors across the representations to provide a closer insight into the local

patterns of the embedding spaces without resorting to dimensionality reduction (Fig. 4 and Supplementary Tables S1–S7). We selected “Myocardial Ischemia” as the focus of this evaluation because of its relevance to global health [61, 62], rich literature [63], and connection to our previous studies [64, 65].

We can observe several different tendencies between the embeddings of the two models. NSG is consistent in emphasizing closely connected concepts across the three context-sampling methods, with

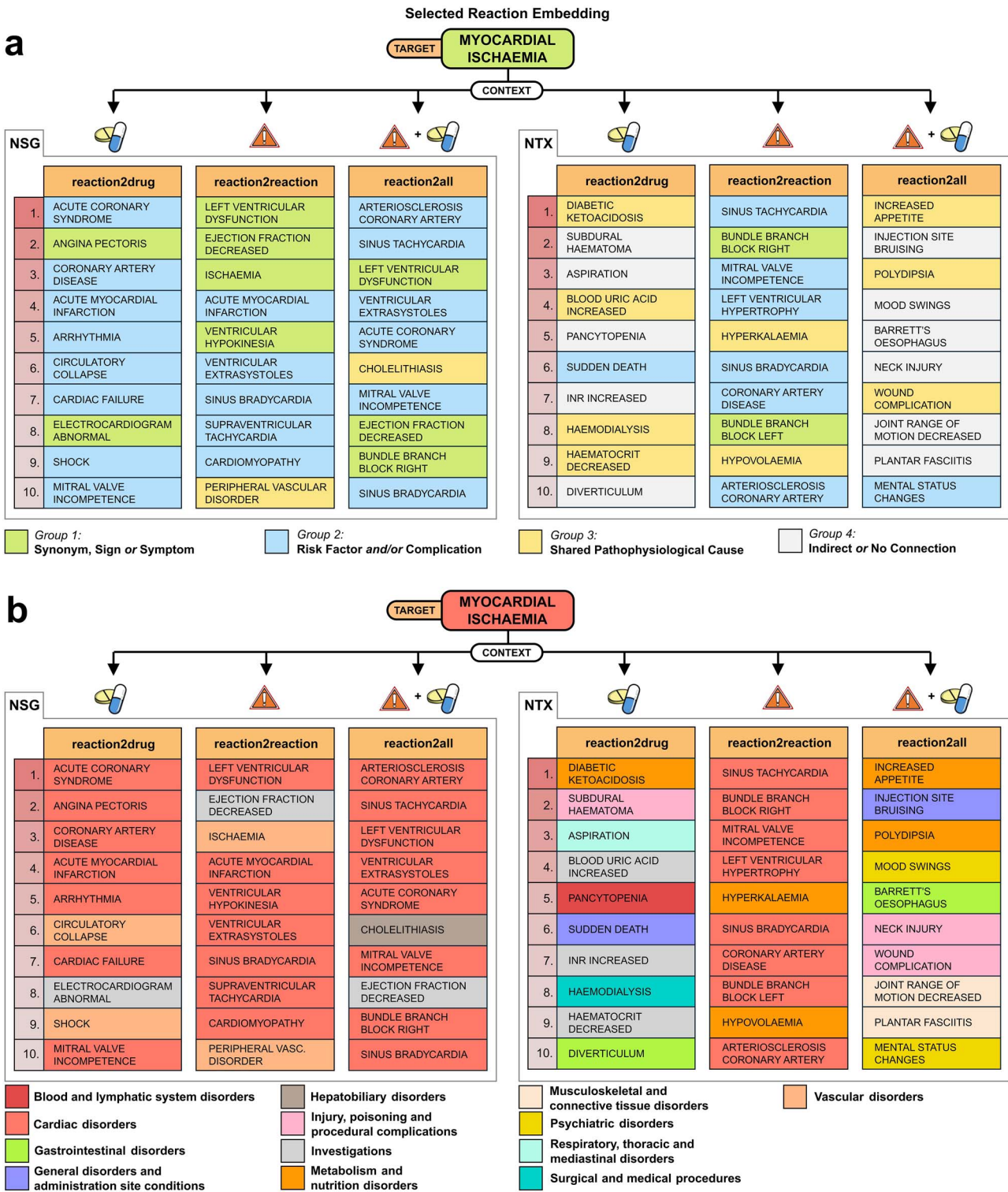


Figure 4 Lists of the top 10 closest neighbors of the MedDRA® PT myocardial ischemia across the six reaction embeddings, measured by cosine similarity in the original vector space, colored (a) by semantic evaluation based on expert knowledge and literature, and (b) by primary MedDRA® SOC.

distinguishable preferences among them: reaction2drug prioritizes the most well-known complications and symptoms, reaction2reaction leans toward terms describing impaired pump function and heart rhythm problems, and reaction2all results in a balanced set of the two. In contrast, NTX embeddings resemble such patterns only with reaction2reaction, indicative of its downstream performance (Section

Association prediction), while the other two sampling methods lead to several terms that share pathophysiological causes with “Myocardial Ischemia,” like “Polydipsia” (excessive thirst) or “Diabetic Ketoacidosis” (acidification of blood due to lack of insulin), which are typical symptoms of untreated diabetes mellitus. Furthermore, many indirect entries appear as well, suggesting patient-specific connections that are

less based on pathophysiology and more on circumstantial factors. For example, “Subdural Hematoma” (bleeding in the skull) might occur from falling or collapsing caused by hemodynamic instability, and “Aspiration” is common in patients who require intensive care due to an acute exacerbation of “Coronary Artery Disease.”

Overall, this analysis demonstrates how pathophysiological and clinical properties of various diseases (reactions) can be captured purely from pharmacovigilance data by our proposed representation learning framework, encoding them on the micro-scale, in local neighborhoods.

Dispersion of related reactions

So far, we observed that the embedding vectors tend to cluster differently than MedDRA[®] SOC (Fig. 3), yet their immediate neighbors display strong semantic relations (Fig. 4). For pharmacovigilance purposes, we would expect the representations to also preserve drug-safety-related information, which we measure by using curated datasets of drug-ADR associations [42–44], considering them as pairs in which a strong positive relationship is validated (from here on referred to as validated data). A fair assumption would be that reactions caused by the same drug (a “validated group” of reactions) share more similarities than unrelated reactions (a “random group”), thus should be more closely packed together (i.e. less dispersed) in the embedding space. The statistical test for the homogeneity of dispersions between these validated and random groups showed that above 95% of the validated groups are significantly less dispersed than their corresponding random groups, in almost all embeddings (Supplementary Table S9), which motivated us to switch to a finer measure of separation, here referred to as “gain norm.” We expect the gain norm to be large for embedding spaces that well preserve drug-safety-related information (i.e. a large separation of validated groups from background noise), therefore we also utilized it as a heuristic measure for tuning the hyperparameters of our embedding models, most importantly the number of epochs for selecting the final checkpoints (Fig. 5b and Supplementary Figs S6–S10). Through node-based, hierarchical semantic similarity measures [66], we can calculate the gain norm for the MedDRA[®] terminology network as well. Comparison to the embedding-based gain norms (Fig. 5c) reveals a stark difference in the distinguishability of these validated groups from the background noise. While the embedding-based values consistently reach 0.25–0.3 across all measures (except Euclidean), going as high as 0.46, the MedDRA-based ones are all in the range of 0.02–0.04. Distributions plotted in Fig. 5a reveal that the MedDRA-based measures are hindered by the larger validated groups, which blend in with the random groups. An exception is the Jiang–Conrath measure that shows a clear separation, albeit with a very small margin.

As such, this evaluation demonstrates that even though the embedding vectors are organized into a different structure as compared to the MedDRA[®] terminology network, drug-safety-related information is better reflected by their emerging patterns, correlating with real-life pharmacovigilance data, which could provide a more suitable resource in practice.

Association prediction

As the focal point of our work, here we present an application of the embedding vectors as input features for a deep neural network

model (Fig. 2d) performing drug–event association prediction. The previously used validated drug-ADR datasets [42–44] provide us only with positively labeled associations, which are insufficient for supervised training. Although there are datasets with negative pairs [45–47], they are substantially smaller in scale and utilized mostly as benchmarks during performance evaluations. We followed this trend, resorting to synthetic negative data for training, which can be difficult to create, as easy negatives will lead to an overestimated model performance during validation but poor generalization when tested on real samples. Consequently, the data generated by different negative sampling methods cannot be used reliably in a train-validation split to determine the best-performing methods themselves, without using external testing data. Due to the aforementioned benchmark datasets being already heavily limited, we chose not to allocate parts of them for this purpose, and instead we evaluated the final model with all six negative sampling methods (Fig. 6). The average training loss revealed that the most difficult training set was the one that also led to the best performance (highlighted with a blue frame in Fig. 6), demonstrating how robust generalization is best achieved with hard negatives. Additional trainings for the following comparisons were done only with this negative sampling method. See the Discussion and Section 8 of the Supplementary for further assessment on the performance-influencing effects of the various types of negative sampling approaches.

As a simple ablation study, we trained a version of the classifier that receives only the one-hot vector representation of reactions and learns the embeddings “in-place,” during training, so that the added benefit of using separately created representations can be observed. The results (Fig. 7) clearly indicate one of the primary detriments of this approach, i.e. that many of the reactions in the benchmark datasets may not appear in the training data, making it impossible for the model to generalize to unseen samples, leading to a random classifier (AUROC ~0.5) in practice. This necessitates the use of separately learned reaction embeddings.

To position our work in the larger scope of the modern representation learning landscape, we also trained versions of the same classifier using embedding vectors that are generated by various state-of-the-art BERT-based models, fine-tuned for clinical and biological language tasks (PubMedBERT [32], BioMed-RoBERTa [33], Clinical ModernBERT [31], and BioClinical ModernBERT [30]). Although encoder-only LLMs are rarely employed to create static embeddings (as discussed in the Introduction section), their usage here demonstrates how natural-language-informed reaction embeddings compare to those trained on structured pharmacovigilance data. As the overall “Merged” results show (Fig. 7), the NSG-variant contrastive embeddings all outperformed the LLM-based ones, while the NTX-variants outperformed one of them. However, dispersion analysis of these LLM-based representation spaces, as done in Section Dispersion of related reactions, reveals a very poor separation of validated groups from the background noise (Supplementary Table S9), suggesting that they are unable to capture interpretable drug-safety-related information and encode patterns that are harder to decipher.

As the final baseline, we calculated the performance of the reporting odds ratio (ROR), which is a statistical disproportionality measure and the cornerstone of the signal detection method prominently used and recommended by EMA [67]. With only two exceptions (AUPRC for NTX-reaction2drug and NTX-reaction2all), all six contrastive embeddings resulted in the classifier outperforming the ROR method when evaluated on the merged benchmark set (Fig. 7), which demonstrates

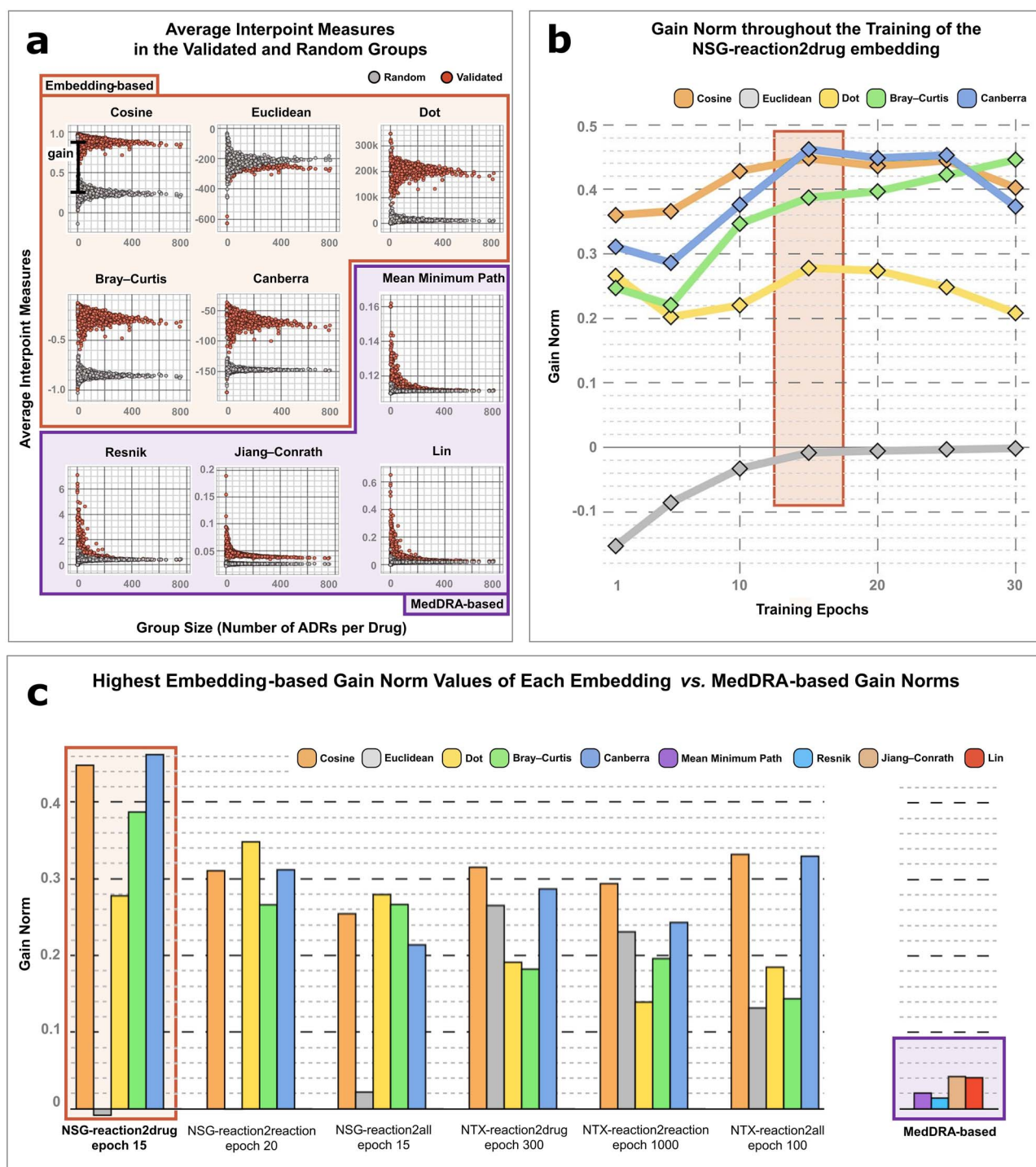


Figure 5 Representation of how the “gain norm” was calculated and used; (a) the concept of (unnormalized) gain: for each drug, we collected their validated reactions from known drug–ADR association data (validated group), and randomly sampled an equal number of other reactions (random group); the within-group average interpoint measures are plotted for both group types, as a function of group size, where the scale of the Euclidean, Bray–Curtis, and Canberra distances are negated so that a higher value on the y-axis means more closely related groups on all plots; gain is calculated as the average difference between the validated and the random groups; (b) changes in the gain norm (normalized gain) as seen throughout the training of the NSG-reaction2drug embedding; we utilized the gain norm as a heuristic method for selecting the best epoch for each embedding; highlighted here is the epoch selected for NSG-reaction2drug, from which data is visualized in panel (a) as well; (c) summary of the highest scoring embedding results and their embedding-based gain norm values; highlighted are the results belonging to the NSG-reaction2drug embedding, used as an example in panel (a) and (b); MedDRA-based gain norms were calculated by using hierarchical similarity measures on the MedDRA® terminology network, which is independent of the embeddings, and so the results are summarized separately.

Negative sampling		Prediction method	EU-ADR			OMOP			FDA-rev			Merged		
Keeping	Sampling		Acc	AUROC	AUPRC	Acc	AUROC	AUPRC	Acc	AUROC	AUPRC	Acc	AUROC	AUPRC
Drugs	Reactions with frequency	NSG-reaction2drug	0.8110	0.8915	0.8738	0.7512	0.8159	0.7616	0.6632	0.7303	0.6509	0.7438	0.8094	0.7501
		NSG-reaction2reaction	0.8211	0.8875	0.8498	0.7449	0.8242	0.7831	0.6440	0.7011	0.6003	0.7372	0.8081	0.7414
		NSG-reaction2all	0.8147	0.8973	0.8816	0.7557	0.8478	0.7995	0.6648	0.7381	0.6386	0.7472	0.8324	0.7717
		NTX-reaction2drug	0.7495	0.8207	0.8122	0.6288	0.6487	0.5685	0.6336	0.6962	0.5928	0.6525	0.6944	0.6180
		NTX-reaction2reaction	0.7807	0.8552	0.8615	0.6418	0.6729	0.6083	0.6560	0.7001	0.6082	0.6708	0.7134	0.6507
		NTX-reaction2all	0.7422	0.8182	0.8225	0.6244	0.6569	0.5361	0.6096	0.6654	0.5837	0.6431	0.6898	0.6030
Reactions	Drugs with frequency	NSG-reaction2drug	0.8339	0.9310	0.9187	0.8019	0.8717	0.8128	0.6784	0.7482	0.6852	0.7813	0.8551	0.7999
		NSG-reaction2reaction	0.8541	0.9187	0.9127	0.8388	0.9167	0.8838	0.6688	0.7329	0.6662	0.8056	0.8775	0.8322
		NSG-reaction2all	0.8358	0.9213	0.9133	0.8291	0.9098	0.8660	0.6808	0.7544	0.7000	0.7986	0.8801	0.8373
		NTX-reaction2drug	0.8046	0.8915	0.8684	0.7036	0.7587	0.6665	0.6736	0.7376	0.6691	0.7153	0.7791	0.7013
		NTX-reaction2reaction	0.8413	0.8972	0.9030	0.7222	0.7957	0.7303	0.6656	0.7175	0.6596	0.7312	0.7968	0.7424
		NTX-reaction2all	0.7670	0.8486	0.8182	0.7291	0.7990	0.7036	0.6608	0.7041	0.6411	0.7222	0.7863	0.7010
None	Both with frequency	NSG-reaction2drug	0.7972	0.8987	0.8959	0.7657	0.8319	0.7787	0.6632	0.7240	0.6398	0.7492	0.8182	0.7580
		NSG-reaction2reaction	0.8119	0.8971	0.8770	0.7898	0.8761	0.8433	0.6280	0.6915	0.6026	0.7591	0.8357	0.7787
		NSG-reaction2all	0.8183	0.9102	0.9043	0.7884	0.8860	0.8447	0.6640	0.7273	0.6443	0.7674	0.8531	0.8026
		NTX-reaction2drug	0.7651	0.8423	0.8339	0.6878	0.7453	0.6701	0.6344	0.6853	0.5800	0.6913	0.7505	0.6683
		NTX-reaction2reaction	0.7560	0.8548	0.8686	0.6895	0.7466	0.7018	0.6400	0.6860	0.5995	0.6905	0.7484	0.6924
		NTX-reaction2all	0.7239	0.8175	0.8094	0.6762	0.7584	0.6711	0.5888	0.6524	0.5720	0.6669	0.7421	0.6552
Drugs	Reactions uniformly	NSG-reaction2drug	0.5945	0.7501	0.7031	0.5452	0.6513	0.6020	0.6032	0.6560	0.5745	0.5637	0.6690	0.5973
		NSG-reaction2reaction	0.5835	0.7598	0.7072	0.5313	0.7054	0.6823	0.5792	0.6160	0.5300	0.5493	0.6947	0.6223
		NSG-reaction2all	0.5807	0.7583	0.7373	0.5427	0.6905	0.6438	0.5912	0.6439	0.5388	0.5576	0.6903	0.6129
		NTX-reaction2drug	0.5312	0.6748	0.6725	0.4983	0.5711	0.4856	0.5688	0.6069	0.5128	0.5173	0.5975	0.5193
		NTX-reaction2reaction	0.5587	0.7105	0.7072	0.5338	0.5898	0.5351	0.5544	0.6096	0.5359	0.5413	0.6154	0.5498
		NTX-reaction2all	0.5413	0.6641	0.5951	0.5075	0.5882	0.4966	0.5488	0.5794	0.5096	0.5209	0.5991	0.5122
Reactions	Drugs uniformly	NSG-reaction2drug	0.7578	0.8869	0.8703	0.8144	0.8945	0.8521	0.6488	0.6967	0.6255	0.7684	0.8515	0.7984
		NSG-reaction2reaction	0.7495	0.8652	0.8447	0.8460	0.9215	0.8902	0.6488	0.6924	0.6340	0.7857	0.8636	0.8140
		NSG-reaction2all	0.7422	0.8759	0.8717	0.8400	0.9142	0.8746	0.6616	0.7026	0.6559	0.7839	0.8645	0.8225
		NTX-reaction2drug	0.7394	0.8419	0.8213	0.7463	0.8179	0.7506	0.6320	0.6618	0.5908	0.7194	0.7873	0.7185
		NTX-reaction2reaction	0.7312	0.8333	0.8290	0.7615	0.8427	0.7815	0.6344	0.6745	0.6018	0.7282	0.7985	0.7276
		NTX-reaction2all	0.6670	0.7734	0.7529	0.7706	0.8495	0.7909	0.5960	0.6274	0.5721	0.7138	0.7857	0.7126
None	Both uniformly	NSG-reaction2drug	0.6037	0.7897	0.7673	0.6227	0.7889	0.7421	0.5912	0.6385	0.5449	0.6098	0.7528	0.6815
		NSG-reaction2reaction	0.6055	0.7800	0.7203	0.6377	0.8144	0.7895	0.5608	0.6049	0.5225	0.6127	0.7610	0.6860
		NSG-reaction2all	0.5972	0.7917	0.7673	0.6488	0.8122	0.7668	0.5936	0.6377	0.5386	0.6255	0.7682	0.6908
		NTX-reaction2drug	0.5817	0.7034	0.6830	0.5584	0.5953	0.5010	0.5823	0.7310	0.6761	0.5745	0.6923	0.6049
		NTX-reaction2reaction	0.6037	0.7338	0.7059	0.6285	0.7574	0.7090	0.5736	0.6241	0.5251	0.6100	0.7192	0.6310
		NTX-reaction2all	0.5523	0.6761	0.6074	0.6019	0.7506	0.7091	0.5336	0.5761	0.4946	0.5759	0.6952	0.5979

Figure 6 Performance of the drug–event association prediction model with the six different embeddings as reaction features, with the six different negative samplings, tested on benchmark datasets from three sources (here referred to as EU-ADR [45], OMOP [46], and FDA-rev [47]); highest performance metrics for each dataset are highlighted in bold, and the overall best performing negative sampling method in highlighted with a blue frame; Acc: accuracy; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision–recall curve.

the ability of the embeddings to serve as a general resource in the support of decision-making, improving the causality assessment process done by experts.

Comparison between our embeddings themselves shows the performance-influencing effect of the various context sampling methods and model choices. The reaction2reaction sampling performed better for the NTX model than both reaction2drug and reaction2all, matching their semantic evaluation trends (Fig. 4a). From this, we can conclude that reaction2reaction is the most suitable sampling method for the NTX model, possibly because it resembles the original use case of such architectures the most, where target–context pairs consist of augmented versions of the same images. On the other hand, the reaction2all sampling method was not only the best-performing NSG-variant but it also resulted in the overall best classification results

across all methods (Fig. 7). We can observe a clear increase in the performance metrics going from reaction2drug to reaction2reaction to reaction2all for the NSG model, which might be explained by the available information (number of context terms) also increasing in this order. In general, embeddings made by NSG performed better than their NTX counterparts, which suggests that the *N*-pairs-based noise contrastive technique employed by NTX is less suited for learning effective embeddings from AERs as compared to the negative sampling noise contrastive technique of NSG.

As a final analysis, we examined the temporal robustness of the best-performing embedding (NSG-reaction2all) by training it with incrementally more AERs, based on date, and testing the resulting new embeddings with our drug–event association prediction model the same way as with the other embeddings. Results showed

Methods		EU-ADR			OMOP			FDA-rev			Merged		
		Acc	AUROC	AUPRC	Acc	AUROC	AUPRC	Acc	AUROC	AUPRC	Acc	AUROC	AUPRC
No embedding (learnt in-place)		0.5404	0.5641	0.5602	0.7452	0.8071	0.7746	0.5296	0.5078	0.4599	0.6599	0.6990	0.6399
Contrastive embeddings	NSG-reaction2drug	0.8339	0.9310	0.9187	0.8019	0.8717	0.8128	0.6784	0.7482	0.6852	0.7813	0.8551	0.7999
	NSG-reaction2reaction	0.8541	0.9187	0.9127	0.8388	0.9167	0.8838	0.6688	0.7329	0.6662	0.8056	0.8775	0.8322
	NSG-reaction2all	0.8358	0.9213	0.9133	0.8291	0.9098	0.8660	0.6808	0.7544	0.7000	0.7986	0.8801	0.8373
	NTX-reaction2drug	0.8046	0.8915	0.8684	0.7036	0.7587	0.6665	0.6736	0.7376	0.6691	0.7153	0.7791	0.7013
	NTX-reaction2reaction	0.8413	0.8972	0.9030	0.7222	0.7957	0.7303	0.6656	0.7175	0.6596	0.7312	0.7968	0.7424
	NTX-reaction2all	0.7670	0.8486	0.8182	0.7291	0.7990	0.7036	0.6608	0.7041	0.6411	0.7222	0.7863	0.7010
LLM embeddings	PubMedBERT	0.8413	0.9210	0.9182	0.7529	0.8324	0.7859	0.6984	0.7452	0.7098	0.7560	0.8281	0.7882
	BioMed-RoBERTa	0.8239	0.9032	0.8992	0.7839	0.8543	0.8116	0.6640	0.7263	0.6810	0.7660	0.8336	0.7913
	Clinical ModernBERT	0.8000	0.8816	0.8843	0.6906	0.7466	0.6910	0.6464	0.6994	0.6605	0.6992	0.7576	0.7102
	BioClinical ModernBERT	0.8349	0.9177	0.9130	0.7607	0.8332	0.7805	0.6880	0.7547	0.7093	0.7572	0.8289	0.7832
Reporting odds ratio		0.8158	0.9094	0.9021	0.6867	0.7096	0.6710	0.6535	0.7592	0.7328	0.7071	0.7558	0.7363
End-to-end models (results from Joopudi et al. ²⁵)	aer2vec	-	0.9120	0.9140	-	0.8210	0.8020	-	-	-	-	-	-
	T-HARE	-	0.9230	0.9350*	-	0.8430	0.8140	-	-	-	-	-	-

Figure 7 Drug–event association prediction results; performance of the classifier was tested by training it with no embeddings (learning them in-place), with using the presented six contrastive embeddings, and with using the four LLM generated embeddings; as an additional baseline, the performance of the ROR signal detection method was also calculated, and for further comparison, the results of two more end-to-end models are included, as reported by Joopudi *et al.* [25] (see the Discussion section); highest values of the present study are highlighted in bold, with one result being outperformed by previous studies, denoted by an asterisk; three different sources of benchmark data were utilized (here referred to as EU-ADR [45], OMOP [46], and FDA-rev [47]), from which a merged set was created, as the individual datasets had duplicate entries and varying sizes; Acc: accuracy; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision–recall curve.

(Supplementary Fig. S11) an already high baseline with the smallest of splits and also a steady increase with more data. See Section 9 of the Supplementary for our further assessment.

Discussion

Here, we present a framework to generate vector representations of adverse events from spontaneous AERs to offer an interpretable and effective resource for general machine learning applications in pharmacovigilance. A comprehensive interpretability analysis of the results was provided, revealing numerous mathematical and semantic patterns, which imply high levels of preserved, drug-safety-related information, with anatomical and pathophysiological patterns emerging.

To showcase a possible application of the embedding vectors and demonstrate their effectiveness as input features, we presented a deep neural classifier performing drug–event association prediction. We trained the model on data that was independent of AERs used for the embeddings, as not to reuse them and to avoid some of the complications arising if we labeled parts of the reported pairs as positives and others as negatives via some mathematical or other unfounded consideration. An example with this issue would be the Transformer-based Hierarchy-aware Adverse Reaction Embeddings (T-HARE) model introduced by Joopudi *et al.* [25], which is a type of representation learning we already discussed in the Introduction section, one that learns and uses the embeddings for a specific end-to-end prediction task instead of general purpose. For its training, a bipartite graph from FAERS data was constructed, and an edge-sampling method was utilized to select positive and negative pairs, based on their drug-normalized reporting frequency. Similar pharmacovigilance networks can indeed be used to perform signal detection, as shown by Pétervári *et al.* [68] and Barbieri *et al.* [69]. However,

the edge weights that are based on report counts need to be properly normalized via network topological features even for outperforming the ROR method [68], suggesting that simpler normalization techniques might be inadequate for creating ground truth training data for machine learning. This limitation was indeed observed by Joopudi *et al.* during the training of T-HARE, which they remedied by scaling the loss on negative samples by their normalized pointwise mutual information. Another notable case is the aer2vec model introduced by Portanova *et al.* [18], which is a similar negative sampling skip-gram version of word2vec as our NSG model; however, it was utilized to perform association prediction directly, after training it on FAERS data. The first issue with this approach is that such shallow, noise contrastive embedding models are not meant to perfectly optimize their training objective, making them a suboptimal choice for this purpose. The other possible issue is the same as with T-HARE, that given its training regimen, the model might learn that all frequently reported pairs are positive, while only the rarely or never reported noisy pairs are negative, which might fail to be of any use in practice.

Instead of relying on AERs for the training data to our drug–event association prediction model, we utilized data where the drug–event pairs are validated to have strong positive associations [42–44]. Although these same three validated datasets were used for the dispersion analysis of related reactions, they presented a fundamentally different objective there, i.e. a heuristic measurement of the interpretable drug-safety-related information. This difference is best showcased by the LLM-derived embeddings achieving relatively comparable results to our contrastive embeddings on the association prediction task (Fig. 7) but orders of magnitude worse gain norm values during dispersion analysis (Supplementary Table S9). Further deviating from this task, we generated additional negative associations for the supervised training of our classifier by randomly recombining the established positive pairs. Here, we also encountered the problem of

easy negatives. We observed that the main contributor to the varying performance of the classifier was whether we sampled the other constituent of the negative pair uniformly or proportionally to its occurrence frequency (Fig. 6). Sampling uniformly made it too easy for the model to separate the uniform negatives from the heavily skewed positives, which resulted in high performance metrics during training but low values during testing. Matching the frequency of the constituents created a more balanced dataset and an overall harder training task (as shown by the loss values in Supplementary Table S10), leading to better generalization capabilities. The best performance coincided with the hardest negative sampling approach, where by keeping the reactions and sampling the drugs by their frequency, the distributions of reactions on the positive and negative sides remain the exact same, enforcing the model to pay more attention to how the reaction embedding vectors and the drug feature vectors interact. The results demonstrate the effect of strong, hard negative samples on model generalizability and thus how much they are required for real-world deployment. As outlined in the recently published 2025 Artificial Intelligence in Pharmacovigilance report, by the Council for International Organizations of Medical Sciences [70], real-world application is highly context-dependent and the impact of dataset limitations heavily relies on the specific objective. For instance, detecting low-prevalence adverse events requires fundamentally different benchmark statistics than training a mechanistic pharmacological model, where low-prevalence values might actually be excluded. Thus, future validation steps must extend beyond small-scale, general benchmarking, and focus on task-specific, expert-curated regulatory data distributions that can properly reflect the challenges of pharmacovigilance.

This leads us to the most important limitation of studies dealing with machine-assisted pharmacovigilance, one that plagues countless other biomedical research areas as well: the lack of proper gold standard benchmark data. Even with our diverse selection of sources (EU-ADR [45], OMOP [46], and FDA-rev [47]), we faced the issue of having a very small test dataset, which imposes serious limitations on estimating real-life performance for association prediction models. As there are no comprehensive, large-scale benchmark data with curated positive and negative drug–event associations, and many of the existing smaller sets are either unavailable [71] or heavily restricted in scope [72, 73], we resorted to some of the more commonly used datasets, enabling comparison with results from other studies (Fig. 7). However, a closer look into these datasets reveals that they also contain varying degrees of easy negatives, which is also reflected in the performance of the discussed methods. For example, in the EU-ADR dataset, negatives are defined as pairs without MEDLINE citations and AERs in VigiBase, making it the easiest dataset to predict, as also seen in the associated AUROC values (ROR: 0.9094, aer2vec: 0.9120, T-HARE: 0.9230, and NSG-reaction2all: 0.9213). The OMOP dataset, on the other hand, was sourced via rigorous literature review and specifically aimed at providing a broad set of positive and negative controls. This is well reflected in the AUROC values (ROR: 0.7096, aer2vec: 0.8210, T-HARE: 0.8430, and NSG-reaction2all: 0.9098) in which our model achieved a significant improvement compared to the other methods, potentially due to its training data being independent from AERs. Finally, in the FDA-rev dataset, negative controls were the result of a nearly identical negative sampling approach to what we used for generating our training data: aiming at a balanced distribution of drugs and reactions on both the positive and negative side while adhering to known associations. This led to the hardest negative samples and arguably the most difficult benchmark dataset, on which even our best embedding

was slightly outperformed by the ROR method (AUROC: 0.7544 versus 0.7592). Low performance might also be attributed to the FDA-rev positive pairs containing not just validated reactions but “warnings and precautions” as well, which are significant safety hazards with or without established causality (see Supplementary Table S11 for our assessment). There are no performance measures available on FDA-rev data for the other methods.

Conclusions

In conclusion, we advance machine-assisted pharmacovigilance by offering a versatile and interpretable vector representation of adverse events as a resource that could enable the development of a wide array of machine learning applications to support decision-making and improve patient safety.

Key Points

- Assessing the potential association between drugs and adverse events is a central, yet laborious task of drug safety. However, supporting the decision-making process of experts by machine learning during this causality assessment is hindered by the text-based nature of the adverse event reports (AERs) and the unreliability of artificial intelligence models that operate purely by linguistic information, such as large language models (LLMs).
- Here, we present a representation learning framework utilizing contrastive learning algorithms that generate adverse event embeddings from structured AERs, reflecting the real-world drug safety profile of the adverse events rather than their linguistic characteristics as inferred by LLMs from natural text.
- Through various interpretability analysis methods, like visualization, clustering, semantic analysis, and statistical evaluations, we demonstrate the preserved drug-safety-related information and emerging pathophysiological patterns of the adverse event embeddings. By employing the embeddings as input features in a drug–event association prediction model, we demonstrate their real-life applicability as a general resource for downstream tasks, outperforming LLM-based embeddings.
- We offer publicly available code and the accompanying data, both input and output, of our representation learning framework at deepvigilance.com.

Acknowledgments

We thank Dóra Bihary for providing valuable feedback about the grammatical correctness and language of the manuscript.

Author contributions

Olivér M. Balogh (Conceptualization [equal], Methodology design [equal], Coding [equal], Model implementation, Data analysis [equal], Visualization [equal], Paper preparation—original draft [equal]), Mátyás Pétervári (Conceptualization [equal], Data analysis [equal], Domain expertise [equal], Paper preparation—revision [equal]), Áron M. Csernák (Coding [equal], Data analysis [equal], Domain expertise [equal], Visualization [equal], Paper preparation—original draft [equal]), Eszter Puhl (Data analysis [equal], Domain expertise

[equal], Paper preparation—revision [equal]), András Horváth (Methodology design [equal], Paper preparation—revision [equal]), Péter Ferdinandy (Project supervision [equal], Domain expertise [equal], Paper preparation—revision [equal]), and Bence Ágg (Project supervision [equal], Conceptualization [equal], Methodology design [equal], Paper preparation—revision [equal])

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

P.F. is the founder and CEO of, and B.Á. and O.B.M. are employed by Pharmahungary Group, a group of R&D companies. M.P. is the founder and CEO of Sanovigado Kft, a pharmaceutical consultancy and R&D company.

Funding

Project No. RRF-2.3.1-21-2022-00003 has been implemented with the support provided by the European Commission, European Union. This project has received funding from the HUN-REN Hungarian Research Network and the National Research, Development and Innovation Office of Hungary: 2020-1.1.6-JÖVŐ-2021-00013. The research was financed by the Thematic Excellence Program (2020-4.1.1.-TKP2020) of the Ministry for Innovation and Technology in Hungary, within the framework of the Therapeutic Development and Bioimaging thematic programs of the Semmelweis University. This study was supported by the Semmelweis 250+ Excellence Fellowship (O.M.B. and E.P.). This study was supported by the 2024-1.2.3.-HU-RIZONT-2024-00026 (P.F.), the 2024-2.1.1-EKÖP-2024-00023, and the 2025-2.1.1-EKÖP-2025-00014 New National Excellence Program (O.M.B.) of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. This project has been implemented with the support provided by the European Union and co-financed by the Hungarian Government under the Digital Renewal Operational Program Plus (DROP Plus), project ID: DIMOP_PLUSZ-1.1.2/B-24-2025-00130.

Data availability

All the data files for training our embedding models and the drug-event association prediction model, as well as for performing clustering is available at deepvigilance.com, namely the preprocessed report corpus made from FAERS AERs, the derived drug and reaction frequencies, the preprocessed validated and benchmark drug-ADR data, the collected drug feature vectors, the presented reaction embedding vectors, and their 2D *t*-SNE coordinates. In addition, the complete results of the HDBSCAN clustering, the raw top-neighbor tables for “Myocardial Ischemia,” and the prediction results made by the ROR-based signal detection method and our best classifier on the benchmark sets were also made available, along with the best performing classifier model itself. Interactive HTML visualizations are provided for the *t*-SNE and clustering results, enabling manual data analysis and browsing. Due to licensing, MedDRA® terminology data are not provided, which are needed for top 10 closest neighbor and dispersion re-calculations with the hierarchical semantic similarity measures.

Other supporting data files are available upon reasonable request to the authors.

Code availability

All code necessary for replicating the results of this study from the provided preprocessed adverse event reporting data, i.e. Python code for generating, analyzing, and using the reaction embeddings, is available under the MIT license at github.com/semmelweis-pharmacology/deepvigilance.

References

1. European Medicines Agency. *Guideline on Good Pharmacovigilance Practices (GVP). Annex I-Definitions (Rev 4)*. 2017.
2. EudraVigilance | European Medicines Agency (EMA).
3. FDA Adverse Event Reporting System (FAERS) Database | FDA.
4. Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. *Drug Inf J* 2008;**42**:409–19. <https://doi.org/10.1177/009286150804200501>
5. Behera SK, Das S, Xavier AS *et al*. Comparison of different methods for causality assessment of adverse drug reactions. *Int J Clin Pharm* 2018;**40**:903–10. <https://doi.org/10.1007/s11096-018-0694-9>
6. FDA Adverse Events Reporting System (FAERS) Public Dashboard | Sheet - Qlik Sense.
7. Danysz K, Cicirello S, Mingle E *et al*. Artificial intelligence and the future of the drug safety professional. *Drug Saf* 2019;**42**:491–7. <https://doi.org/10.1007/s40264-018-0746-z>
8. Al-Azzawi F, Mahmoud I, Haguinet F *et al*. Developing an artificial intelligence-guided signal detection in the Food and Drug Administration Adverse Event Reporting System (FAERS): a proof-of-concept study using galcanezumab and simulated data. *Drug Saf* 2023;**46**:743–51. <https://doi.org/10.1007/s40264-023-01317-0>
9. Cherkas Y, Ide J, van Stekelenborg J. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. *Drug Saf* 2022;**45**:571–82. <https://doi.org/10.1007/s40264-022-01163-6>
10. De Abreu FR, Zhong S, Moureaud C *et al*. A pilot, predictive surveillance model in pharmacovigilance using machine learning approaches. *Adv Ther* 2024;**41**:2435–45. <https://doi.org/10.1007/s12325-024-02870-5>
11. Bae JH, Baek YH, Lee JE *et al*. Machine learning for detection of safety signals from spontaneous reporting system data: example of nivolumab and docetaxel. *Front Pharmacol* 2021;**11**:602365.
12. Wishart DS, Feunang YD, Guo AC *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82. <https://doi.org/10.1093/nar/gkx1037>
13. Kim S, Bolton EE. PubChem: a large-scale public chemical database for drug discovery. In: Daina A, Przewosny M, Zoete V (eds.), *Open Access Databases and Datasets for Drug Discovery*, pp. 41–66. Wiley-VCH, Weinheim, Germany: John Wiley & Sons, Ltd, 2023.
14. Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. *J Am Med Inform Assoc* 2018;**25**:1339–50.
15. Gattepaille LM. Using the WHO database of spontaneous reports to build joint vector representations of drugs and adverse drug reactions, a promising avenue for pharmacovigilance. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. USA: Institute of Electrical and Electronics Engineers Inc., 2019.
16. Beam AL, Kompa B, Schmaltz A *et al*. Clinical concept embeddings learned from massive sources of multimodal medical data. *Pac Symp Biocomput* 2020;**25**:295–306.

17. Choi Y, Chiu CY-I, Sontag D. Learning low-dimensional representations of medical concepts. *AMIA Jt Summits Transl Sci Proc* 2016;**2016**:41–50.
18. Portanova J, Murray N, Mower J *et al.* aer2vec: distributed representations of adverse event reporting system data as a means to identify drug/side-effect associations. *AMIA Annu Symp Proc* 2019;**2019**:717–26.
19. Nikfarjam A, Sarker A, O'Connor K *et al.* Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015;**22**:671.
20. Chen Q, Lee K, Yan S *et al.* BioConceptVec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput Biol* 2020;**16**:e1007617. <https://doi.org/10.1371/journal.pcbi.1007617>
21. Erlanson N, China JF, Taavola H *et al.* Clinical relatedness and stability of vigiVec semantic vector representations of adverse events and drugs in pharmacovigilance. *Drug Saf* 2025;**48**:401. <https://doi.org/10.1007/s40264-024-01509-2>
22. Zhang Y, Chen Q, Yang Z *et al.* BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019;**6**:52.
23. Zhao H, Ni P, Zhao Q *et al.* Identifying the serious clinical outcomes of adverse reactions to drugs by a multi-task deep learning framework. *Commun Biol* 2023;**6**:1–14.
24. Zhang F, Sun B, Diao X *et al.* Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Med Inform Decis Mak* 2021;**21**:1–11.
25. Joopudi V, Dandala B, Tsou C-H *et al.* Hierarchy-aware adverse reaction embeddings for signal detection. *AMIA Annu Symp Proc* 2022;**2022**:596.
26. Mao Y, Fung KW. Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. *J Am Med Inform Assoc* 2020;**27**:1538–46.
27. Alsentzer E, Murphy J, Boag W *et al.* Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–8. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019.
28. Park J, Djelassi M, Chima D *et al.* Validation of a natural language machine learning model for safety literature surveillance. *Drug Saf* 2024;**47**:71–80. <https://doi.org/10.1007/s40264-023-01367-4>
29. Portelli B, Lenzi E, Chersoni E *et al.* BERT prescriptions to avoid unwanted headaches: a comparison of transformer architectures for adverse drug event detection. In: *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 1740–7. USA: Association for Computational Linguistics (ACL), 2021.
30. Sounack T, Davis J, Durieux B *et al.* BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP. 2025.
31. Lee SA, Wu A, Chiang JN. *Clinical ModernBERT: An Efficient and Long Context Encoder for Biomedical Text*. 2025.
32. Gu Y, Tinn R, Cheng H *et al.* Domain-specific language model pre-training for biomedical natural language processing. *ACM Trans Comput Healthc* 2021;**3**:24.
33. Gururangan S, Marasovic A, Swayamdipta S *et al.* Don't stop pretraining: adapt language models to domains and tasks. *Proc Annu Meet Assoc Comput Linguist* 2020; 58th proceedings, pages 8342–60.
34. Mozzicato P. MedDRA: an overview of the medical dictionary for regulatory activities. *Pharmaceut Med* 2009;**23**:65–75.
35. Bommasani R, Davis K, Cardie C. Interpreting pretrained contextualized representations via reductions to static embeddings. *Proc Annu Meet Assoc Comput Linguist* 2020; 58th proceedings, pages 4758–81.
36. Gupta P, Jaggi M. Obtaining better static word embeddings using contextual embedding models. In: *ACL-IJCNLP 2021 - 59th Annu Meet Assoc Comput Linguist 11th Int Jt Conf Nat Lang Process Proc Conf*, Vol. 1, pp. 5241–53. USA: Association for Computational Linguistics, 2021.
37. Chen T, Kornblith S, Norouzi M *et al.* A simple framework for contrastive learning of visual representations. In: HD III, Singh A (eds.), *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119, pp. 1597–607. USA: PMLR, 2020.
38. Mikolov T, Sutskever I, Chen K *et al.* *Distributed Representations of Words and Phrases and their Compositionality*. Neural Information Processing Systems Foundation. USA: Advances in Neural Information Processing Systems, 2013.
39. Gutmann M, Hyvärinen A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Teh YW, Titterton M (eds.), *J Mach Learn Res*. USA, 2010;**9**:297–304.
40. Nelson SJ, Zeng K, Kilbourne J *et al.* Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;**18**:441–8. <https://doi.org/10.1136/amiajnl-2011-000116>
41. WHO Collaborating Centre for Drug Statistics Methodology. *Guidelines for ATC Classification and DDD Assignment*. Oslo, Norway: Norwegian Institute of Public Health, Norwegian Institute of Public Health, 2023.
42. Kuhn M, Letunic I, Jensen LJ *et al.* The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075. <https://doi.org/10.1093/nar/gkv1075>
43. Wisniewski AFZ, Bate A, Bousquet C *et al.* Good signal detection practices: evidence from IMI PROTECT. *Drug Saf* 2016;**39**:469–90. <https://doi.org/10.1007/s40264-016-0405-1>
44. Kontsioti E, Maskell S, Dutta B *et al.* A reference set of clinically relevant adverse drug-drug interactions. *Sci Data* 2022;**9**:1–9.
45. Coloma PM, Avillach P, Salvo F *et al.* A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Saf* 2013;**36**:13–23. <https://doi.org/10.1007/s40264-012-0002-x>
46. Ryan PB, Schuemie MJ, Welebob E *et al.* Defining a reference set to support methodological research in drug safety. *Drug Saf* 2013;**36**:33–47. <https://doi.org/10.1007/s40264-013-0097-8>
47. Harpaz R, Odgers D, Gaskin G *et al.* A time-indexed reference standard of adverse drug reactions. *Sci Data* 2014;**1**:1–10.
48. Le-Khac PH, Healy G, Smeaton AF. Contrastive representation learning: a framework and review. *IEEE Access* 2020;**8**:193907–34. <https://doi.org/10.1109/ACCESS.2020.3031549>
49. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
50. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L *et al.* (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 160–72. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.
51. Linderman GC, Steinerberger S. Clustering with t-SNE, provably. *SIAM J Math Data Sci* 2019;**1**:313–32. <https://doi.org/10.1137/18M1216134>
52. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 2006;**62**:245–53. <https://doi.org/10.1111/j.1541-0420.2005.00440.x>
53. Gijbels I, Omelka M. Testing for homogeneity of multivariate dispersions using dissimilarity measures. *Biometrics* 2013;**69**:137–45. <https://doi.org/10.1111/j.1541-0420.2012.01797.x>

54. Spector J, Aldana A, Sebek M *et al.* Transformers enhance the predictive power of network medicine. *medRxiv* 2025; 2025.01.27.25321204.
55. Wang CS, Lin PJ, Cheng CL *et al.* Detecting potential adverse drug reactions using a deep neural network model. *J Med Internet Res* 2019;**21**:e11016. <https://doi.org/10.2196/11016>
56. Chollet F. *Others. Keras*. 2015.
57. Pedregosa F, Weiss R, Brucher M *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
58. Inc. PT. Collaborative Data Science. 2015.
59. Wolf T, Debut L, Sanh V *et al.* Transformers: state-of-the-art natural language processing. In: Liu Q, Schlangen D (eds.), *EMNLP 2020 - Conf Empir Methods Nat Lang Process Proc Syst Demonstr*, pp. 38–45. USA: Association for Computational Linguistics, 2020.
60. Paszke A, Gross S, Massa F *et al.* PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019;**32**.
61. Roth GA, Johnson C, Abajobir A *et al.* Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol* 2017;**70**:1–25.
62. Rezende PC, Ribas FF, Serrano CV *et al.* Clinical significance of chronic myocardial ischemia in coronary artery disease patients. *J Thorac Dis* 2019;**11**:1005. <https://doi.org/10.21037/jtd.2019.02.85>
63. Shimokawa H, Yasuda S. Myocardial ischemia: current concepts and future perspectives. *J Cardiol* 2008;**52**:67–78. <https://doi.org/10.1016/j.jjcc.2008.07.016>
64. Szabados T, Molnár A, Kenyeres É *et al.* Identification of new, translatable ProtectomiRs against myocardial ischemia/reperfusion injury and oxidative stress: the role of MMP/Biglycan signaling pathways. *Antioxidants* 2024;**13**:674. <https://doi.org/10.3390/antiox13060674>
65. Nagy RN, Makkos A, Baranyai T *et al.* Cardioprotective microRNAs (protectomiRs) in a pig model of acute myocardial infarction and cardioprotection by ischaemic conditioning: MiR-450a. *Br J Pharmacol* 2024;**182**:396–416.
66. Bill RW, Liu Y, McInnes BT *et al.* Evaluating semantic relatedness and similarity measures with standardized MedDRA queries. *AMIA Annu Symp Proc* 2012;**2012**:43–50.
67. European Medicines Agency. *Screening for Adverse Reactions in EudraVigilance*. Data Standardisation and Analytics (EMA), 2016.
68. Pétervári M, Benczik B, Balogh OM *et al.* Network analysis for signal detection in spontaneous adverse event reporting database: application of network weighting normalization to characterize cardiovascular drug safety. *Drug Saf* 2022;**45**:1423–38. <https://doi.org/10.1007/s40264-022-01225-9>
69. Barbieri MA, Abate A, Balogh OM *et al.* Network analysis and machine learning for signal detection and prioritization using electronic healthcare records and administrative databases: a proof of concept in drug-induced acute myocardial infarction. *Drug Saf* 2025;**48**:1–14. <https://doi.org/10.1007/s40264-025-01515-y>
70. CIOMS Working Group XIV. *Artificial Intelligence in Pharmacovigilance*. Geneva: Council for International Organizations of Medical Sciences (CIOMS), 2026.
71. Caster O, Dietrich J, Kürzinger ML *et al.* Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project. *Drug Saf* 2018;**41**:1355. <https://doi.org/10.1007/s40264-018-0699-2>
72. Osokogu OU, Fregonese F, Ferrajolo C *et al.* Pediatric drug safety signal detection: a new drug–event reference set for performance testing of data-mining methods and systems. *Drug Saf* 2015;**38**:207. <https://doi.org/10.1007/s40264-015-0265-0>
73. Ito S, Narukawa M. Development of a drug safety signal detection reference set using Japanese safety information. *Ther Innov Regul Sci* 2025;**59**:288–94. <https://doi.org/10.1007/s43441-024-00729-z>