International Journal of
Molecular Sciences

Article

Advancing Adverse Drug Reaction Prediction with Deep Chemical Language Model for Drug Safety Evaluation

Jinzhu Lin [1], Yujie He [1], Chengxiang Ru [1], Wulin Long [1], Menglong Li [1,*] and Zhining Wen [1,2,*]

[1] College of Chemistry, Sichuan University, Chengdu 610064, China
[2] Medical Big Data Center, Sichuan University, Chengdu 610064, China
* Correspondence: liml@scu.edu.cn (M.L.); w_zhining@scu.edu.cn (Z.W.)

Abstract: The accurate prediction of adverse drug reactions (ADRs) is essential for comprehensive drug safety evaluation. Pre-trained deep chemical language models have emerged as powerful tools capable of automatically learning molecular structural features from large-scale datasets, showing promising capabilities for the downstream prediction of molecular properties. However, the performance of pre-trained chemical language models in predicting ADRs, especially idiosyncratic ADRs induced by marketed drugs, remains largely unexplored. In this study, we propose MoLFormer-XL, a pre-trained model for encoding molecular features from canonical SMILES, in conjunction with a CNN-based model to predict drug-induced QT interval prolongation (DIQT), drug-induced teratogenicity (DIT), and drug-induced rhabdomyolysis (DIR). Our results demonstrate that the proposed model outperforms conventional models applied in previous studies for predicting DIQT, DIT, and DIR. Notably, an analysis of the learned linear attention maps highlights amines, alcohol, ethers, and aromatic halogen compounds as strongly associated with the three types of ADRs. These findings hold promise for enhancing drug discovery pipelines and reducing the drug attrition rate due to safety concerns.

Keywords: adverse drug reactions; drug safety evaluation; deep chemical language model; structural alerts; deep learning

Citation: Lin, J.; He, Y.; Ru, C.; Long, W.; Li, M.; Wen, Z. Advancing Adverse Drug Reaction Prediction with Deep Chemical Language Model for Drug Safety Evaluation. *Int. J. Mol. Sci.* **2024**, *25*, 4516. https://doi.org/10.3390/ijms25084516

Academic Editor: Giulio Vistoli

Received: 26 March 2024
Revised: 15 April 2024
Accepted: 16 April 2024
Published: 20 April 2024

1. Introduction

Market drugs frequently lead to unexpected adverse drug reactions (ADRs) during clinical use, presenting significant challenges to patient safety [1,2]. Among these, idiosyncratic ADRs, such as drug-induced rhabdomyolysis, pose a particular risk due to their life-threatening nature and intricate pathogenesis, compounded by their low incidence rates [3–5]. Despite efforts to identify clinical biomarkers and monitoring modalities for timely diagnosis during medication intake, the complexity of these ADRs and substantial inter-individual variability among patients impede the diagnostic efficacy of these clinical parameters [1,6,7]. Consequently, the accurate prediction of the risk associated with serious ADRs induced by marketed drugs becomes imperative in drug safety evaluation. Such predictions not only facilitate the prompt adjustment of medication strategies to ensure patient safety and efficacy but also serve as early detection mechanisms within the drug discovery process. Structure–Activity Relationship (SAR) methodologies, which are based on the chemical structures and properties of compounds, utilize various molecular descriptors as features and leverage machine-learning techniques to predict or interpret the relationship between the chemical structure of compounds and its biological activity or toxicological characteristics [8]. These methodologies offer pivotal approaches to correlate the chemical structure of drugs with their propensity for ADRs, representing indispensable tools for predicting drug-induced adverse reactions [9,10]. Previous research has demonstrated the utility of SAR models employing conventional machine-learning methods, including Support Vector Machine (SVM) [11–14], Random Forest (RF) [15–17], and Decision Forest

(DF) [18,19]. Notably, these models have exhibited promising performance in predicting liver toxicity [15,18–20], cardiotoxicity [13,14,17,21], and other adverse reactions. In our previous investigation, we constructed datasets encompassing three distinct ADRs based on FDA-approved drug labeling information: the drug-induced QT prolongation atlas (DIQTA) [22], drug-induced teratogenicity dataset (DITD) [12], and drug-induced rhabdomyolysis atlas (DIRA) [23]. Leveraging conventional machine-learning algorithms in conjunction with molecular descriptors, we predicted the risk associated with these three types of ADRs [12,13,16].

Conventional machine-learning models, which utilize various types of molecular descriptors as features, have exhibited limitations in predicting adverse drug reactions (ADRs) [24,25]. The selection of molecular descriptors for prediction tends to be somewhat arbitrary, typically relying on experiential knowledge or the incorporation of as many descriptors as feasible, followed by an assessment of feature importance. Despite these efforts, effectively translating these descriptors into structural alerts (SAs) for analyzing structural features closely associated with ADRs during mechanistic interpretation remains challenging. Furthermore, the reliability of these models may be compromised by limited sample sizes, particularly for drugs that induce idiosyncratic ADRs. In recent years, several studies have investigated the development of chemical language models employing deep learning algorithms [26–30]. These models can be trained on extensive datasets of molecules to learn structural information from molecular representations, such as the Simplified Molecular Input Line Entry System (SMILES), and subsequently establish pre-trained models for predicting molecular properties. Li et al. introduced SMILES Pair Encoding (SPE) [31], which learns the vocabulary of high-frequency SMILES substrings from large chemical datasets and then tokenizes SMILES. This approach has demonstrated promise in molecular generation and Quantitative Structure–Activity Relationship (QSAR) prediction tasks. Wu et al. employed the BiLSTM (Bidirectional Long Short-Term Memory) attention network to extract key features from SMILES strings [32], achieving superior results across eleven tasks related to absorption, distribution, toxicity, and others. Ucak et al. proposed the atom in the SMILES tokenization scheme [33], which incorporates environmental information and resolves ambiguities in the general nature of SMILES tokens. These studies utilize deep learning algorithms to decipher the intricate chemical information concealed within SMILES and underscore the significant impact of appropriate tokenization on multiple chemical translation and molecular property prediction tasks. However, the performance of pre-trained chemical language models combined with deep learning methods in predicting ADRs, particularly idiosyncratic ADRs induced by marketed drugs, remains largely unexplored.

In this study, we employ MoLFormer-XL [34], a pre-trained model trained on large-scale molecular datasets, to encode molecules. This is coupled with convolutional neural networks (CNNs) to predict the risk of drug-induced QT prolongation (DIQT), drug-induced teratogenicity (DIT), and drug-induced rhabdomyolysis (DIR), exploring the performance of deep chemical language models in predicting marketed drug-induced adverse reactions. We compare the results with conventional machine-learning methods employed in previous studies to evaluate the effectiveness of the model. Furthermore, by scrutinizing the structures emphasized in the linear attention maps of drugs associated with high-risk concerns, we discerned structural alerts, such as amines, alcohols, ethers, and aromatic halogens, which may correlate with ADRs. In summary, our investigation underscores that deep chemical language models effectively capture molecular structural information, proving instrumental in predicting ADRs induced by marketed drugs. This presents a viable alternative approach to bolstering drug safety evaluation and facilitating the early detection of ADRs throughout the drug development process.

2. Results

2.1. Model Performance on Three Curated Datasets

Three meticulously curated datasets, namely DIQTA, DITD, and DIRA, were employed to assess the performance of our proposed model, MoLFormer-XL-CNN. This model combines a pre-trained deep chemical language model with CNN to predict ADRs induced by marketed drugs. Performance metrics, including accuracy, recall rate, precision, Matthew's correlation coefficient (MCC), balanced accuracy score (BACC), F1 score, the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), and specificity, achieved by our model in predicting the risk of drug-induced QT prolongation, teratogenicity, and rhabdomyolysis, are listed in Table 1. The MCCs and recall rates achieved on the three datasets are over 0.50 and 0.85, respectively, indicating satisfactory performance in predicting these ADRs. Particularly noteworthy are the high recall rates of 0.942 and 0.974 achieved on the DIQTA and DIRA datasets, respectively, illustrating the model's effectiveness in identifying positive samples at risk of inducing QT interval prolongation and rhabdomyolysis. However, it is noteworthy that the specificity for the DIRA dataset is 0.432, indicating a higher false-positive rate in predicting drug-induced rhabdomyolysis compared to the other two ADRs. We compared the performance of our proposed MoLFormer-XL-CNN model with state-of-the-art methods reported in previous studies (Figure 1). In previous studies, Wulin Long et al. indicated that SVM achieved the highest MCC value and recall rate (mean MCC = 0.591, mean recall rate = 0.870), predicting the high risk of inducing QT interval prolongation of marketed drugs [13]. Liyuan Kang et al. indicated that SVM performed better in detecting the marketed drugs with high teratogenic risk (mean MCC = 0.312) [12]. Yifan Zhou et al. proposed that the RF model performed the best (mean MCC = 0.46) in predicting the DIR severity of the marketed drugs [16]. The three previous studies all leveraged conventional machine-learning algorithms in conjunction with molecular descriptors. The results demonstrated the superior performance of our model in terms of both MCCs and recall rates, highlighting the effectiveness of leveraging pre-trained deep chemical language models for ADR prediction tasks. Significant improvements in MCCs, especially for the DITD dataset, were observed where the MCC increased from 0.312 to 0.503. Additionally, our model exhibited improved recall rates compared to the supervised baselines. These results demonstrate that MoLFormer-XL effectively captures the molecular features of drugs, thereby facilitating subsequent prediction tasks.

Table 1. Prediction results of MoLFormer-XL-CNN model for three datasets.

	DIQTA	DITD	DIRA
Accuracy	0.859 ± 0.029	0.750 ± 0.037	0.866 ± 0.034
Recall rate	0.942 ± 0.027	0.862 ± 0.160	0.974 ± 0.027
Precision	0.845 ± 0.029	0.771 ± 0.010	0.872 ± 0.029
MCC	0.702 ± 0.062	0.503 ± 0.047	0.535 ± 0.095
BACC	0.835 ± 0.033	0.719 ± 0.058	0.703 ± 0.042
F1 score	0.891 ± 0.022	0.800 ± 0.052	0.920 ± 0.023
AUROC	0.829 ± 0.060	0.702 ± 0.048	0.703 ± 0.042
AUPRC	0.822 ± 0.079	0.742 ± 0.101	0.832 ± 0.062
Specificity	0.747 ± 0.036	0.576 ± 0.260	0.432 ± 0.088

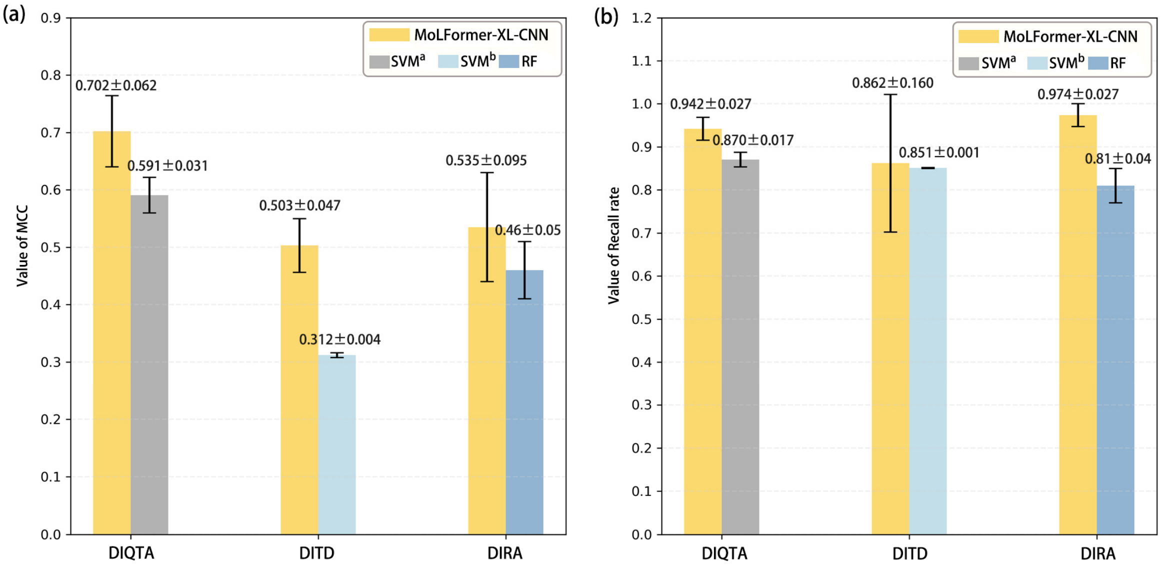


Figure 1. The comparison of the predictive results with three previous studies. SVM^a: Wulin Long et al. established an SVM model to predict DIQT based on the DIQTA dataset [13]. SVM^b: Liyuan Kang et al. established an SVM model to predict DIT based on the DITD dataset [12]. RF: Yifan Zhou et al. established an RF model to predict DIR based on the DIRA dataset [16]. (**a**) The comparison of the value of MCC. (**b**) Comparison of the value of recall rate.

2.2. *Attention Analysis for Drugs with a High Risk of DIQT*

MoLFormer-XL's average-pooled attention metrics encode chemical information within a molecule during the encoding of canonical SMILES, enabling the capture of spatial relations between atoms. Higher values in the average-pooled attention metrics indicate the importance of atoms or groups of atoms (structural fractions) in the prediction task. By analyzing the average-pooled attention matrices, we extracted structural information considered important by the predictive model for ADR predictions. Table 2 lists five drugs identified as high risk for inducing QT interval prolongation by the Comprehensive in vitro Proarrhythmia Assay (CiPA) initiative [35]. The structural subunits with high attention values in the MoLFormer-XL framework were circled and compared with established structural alerts (SAs) obtained from ToxAlert in the Online Chemical Database (OCHEM, https://ochem.eu/home/show.do, accessed on 6 February 2024) [36,37]. Annotations with complete attention maps are provided in Supplementary File S1. The table reveals that MoLFormer-XL was used to report SAs associated with DIQT in previous studies, such as amines, ethers, and aromatic halogens [13]. Furthermore, Aniline structures, such as pyridine in disopyramide phosphate and the benzimidazole ring in quinidine gluconate and vandetanib, were also identified as crucial structural features associated with DIQT [38–40].

Table 2. Attention analysis for drugs with a high risk of DIQT.

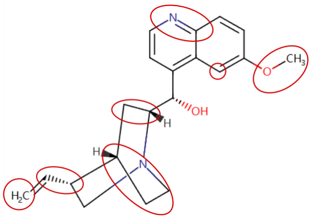
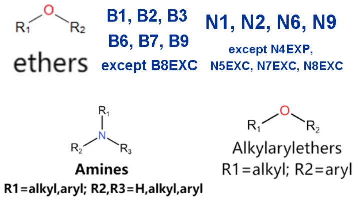
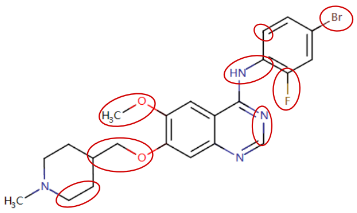
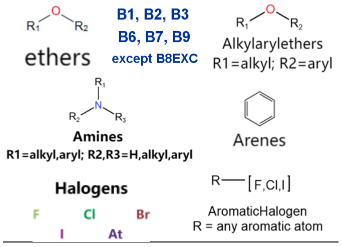
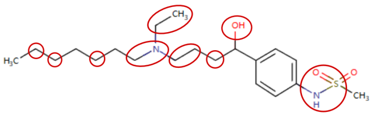
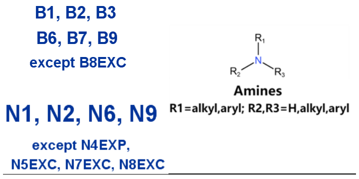
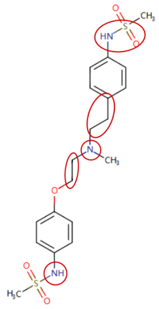
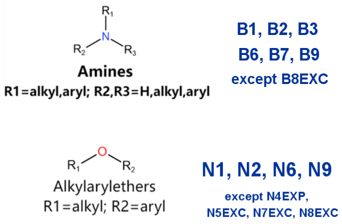
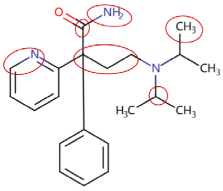
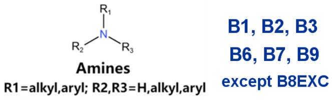
Generic/ Proper Name(s)	Canonical SMILES	Structures Extracted by Attention Map	SAs*
Quinidine gluconate	C=CC1CN2CCC1CC2C(O) c1ccnc2ccc(OC)cc12		 R₁ O R₂ ethers B1, B2, B3 B6, B7, B9 except B8EXC N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R₁ N R₂ R₃ Amines R1=alkyl,aryl; R2,R3=H,alkyl,aryl R₁ O R₂ Alkylarylethers R1=alkyl; R2=aryl

Table 2. *Cont.*

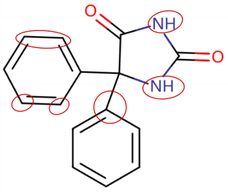
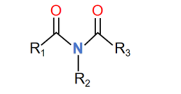
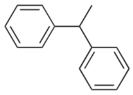
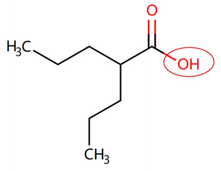
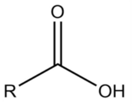
Generic/ Proper Name(s)	Canonical SMILES	Structures Extracted by Attention Map	SAs*
Vandetanib	COc1cc2c(Nc3ccc(Br)cc3F) ncnc2cc1OCC1CCN(C)CC1		 ethers B1, B2, B3 B6, B7, B9 except B8EXC Alkylarylethers R1=alkyl; R2=aryl Amines R1=alkyl,aryl; R2,R3=H,alkyl,aryl Arenes Halogens F Cl Br I At R—[F,Cl,I] AromaticHalogen R = any aromatic atom
Ibutilide fumarate	CCCCCCCN(CC)CCCC(O) c1ccc(NS(C)(=O)=O)cc1		 B1, B2, B3 B6, B7, B9 except B8EXC N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC Amines R1=alkyl,aryl; R2,R3=H,alkyl,aryl
Dofetilide	CN(CCOc1ccc(NS(C)(=O)=O) cc1)CCc1ccc(NS(C)(=O)=O)cc1		 Amines R1=alkyl,aryl; R2,R3=H,alkyl,aryl B1, B2, B3 B6, B7, B9 except B8EXC Alkylarylethers R1=alkyl; R2=aryl N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC
Disopyramide phosphate	CC(C)N(CCC(C(N)=O)(c1cc ccc1)c1ccccn1)C(C)C		 Amines R1=alkyl,aryl; R2,R3=H,alkyl,aryl B1, B2, B3 B6, B7, B9 except B8EXC

SAs*: SAs obtained from the ToxAlert platform in the Online Chemical Database (OCHEM, https://ochem.eu/home/show.do accessed on 6 February 2024).

2.3. Attention Analysis for Antiepileptic Drugs with a High Risk of DIT

Currently, teratogenic risk evaluation for antiepileptic drugs primarily relies on cohort studies from pregnancy registries, comparing the incidence and risk of congenital malformations among offspring exposed to different antiepileptic drugs. Reports from three major pregnancy registries have indicated that valproate and phenytoin are associated with a high risk of teratogenicity [41–44]. Table 3 lists the circled structures of valproate and phenytoin attended by MoLFormer-XL. Annotations with complete attention maps are provided in Supplementary File S2. For phenytoin, the unsubstituted imine nitrogen and ethane-1,1-diyldibenzene in the chemical structure were considered crucial substructures associated with teratogenicity by the model [45,46]. In the chemical structure of valproate, the carboxylic acids were highlighted and associated with teratogenicity, which is consistent with the SAs revealed in previous studies [45,47].

Table 3. Attention analysis for antiepileptic drugs with a high risk of DIT.

Generic/ Proper Name(s)	Canonical SMILES	Structures Extracted by Attention Map	SAs*	
Phenytoin	C1=CC=C(C=C1)C2(C(=O)NC(=O)N2)C3=CC=CC=C3		 Carboxylic acid imides R1, R2, R3 = H, alkyl, aryl	 SA17
Valproate	CCCC(CCC)C(=O)O		 Carboxylic acids R = any carbon atom	

SAs*: SAs obtained from the ToxAlert platform in the Online Chemical Database (OCHEM, https://ochem.eu/home/show.do accessed on 6 February 2024).

2.4. Attention Analysis for Statins

DIR is an idiosyncratic ADR, and its specific pathogenesis is still under exploration. It has been widely reported that drugs in certain therapeutic categories, such as statins, are more likely to induce rhabdomyolysis [48,49]. In the DIRA dataset, all statin drugs were categorized into the highest DIR concern level. Table 4 lists seven statin drugs in the DIRA dataset with attended substructures and reported SAs. Annotations with complete attention maps are provided in Supplementary File S3. The table reveals that nucleophilic groups in the chemical structures of each drug, particularly alcohols and amines, were attended by MoLFormer-XL. These functional groups can be converted into highly reactive electrophilic metabolites, which can covalently bind to nucleophilic sites of biological macromolecules, such as DNA, RNA, or proteins in cells, potentially leading to cell damage and other toxic reactions [39,45,50]. Additionally, aromatic halogens [51] in the chemical structures of fluvastatin, atorvastatin, rosuvastatin, and pitavastatin were also identified as contributors to inducing rhabdomyolysis.

Table 4. Attention analysis for statin.

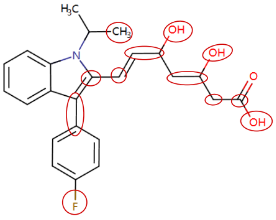
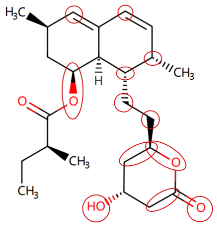
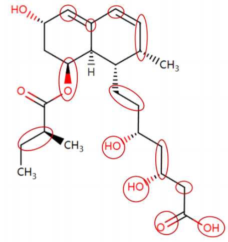
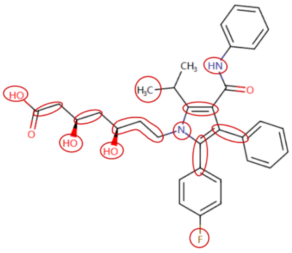
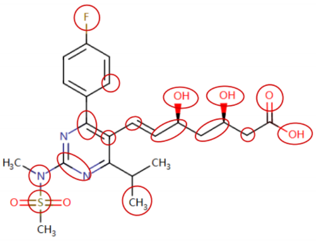
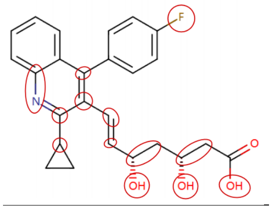
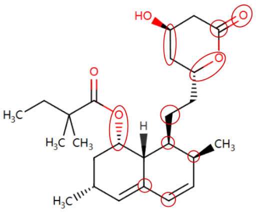
Generic/ Proper Name(s)	Canonical SMILES	Structures Extracted by Attention Map	SAs*
Fluvastatin sodium	CC(C)n1c(C=CC(O)CC(O)CC(=O)O)c(-c2ccc(F)cc2)c2ccccc21		R—[F,Cl,I] AromaticHalogen R = any aromatic atom N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments
Lovastatin	CCC(C)C(=O)OC1CC(C)C=C2C=CC(C)C(CCC3CC(O)CC(=O)O3)C21	 H_3C CH_3 H O O CH_3 CH_3 O HO O	N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments

Table 4. *Cont.*

Generic/ Proper Name(s)	Canonical SMILES	Structures Extracted by Attention Map	SAs*
Pravastatin sodium	CCC(C)C(=O)OC1CC(O)C=C2C=CC(C)C(CCC(O)CC(O)CC(=O)O)C21		N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments
Atorvastatin calcium	CC(C)c1c(C(=O)Nc2ccccc2)c(-c2ccccc2)c(-c2ccc(F)cc2)n1CCC(O)CC(O)CC(=O)O		R—[F,Cl,I] AromaticHalogen R = any aromatic atom N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments
Rosuvastatin calcium	CC(C)c1nc(N(C)S(C)(=O)=O)nc(-c2ccc(F)cc2)c1C=CC(O)CC(O)CC(=O)O		R—[F,Cl,I] AromaticHalogen R = any aromatic atom N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments
Pitavastatin calcium	O=C(O)CC(O)CC(O)C=Cc1c(C2CC2)nc2ccccc2c1-c1ccc(F)cc1		R—[F,Cl,I] AromaticHalogen R = any aromatic atom N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments
Simvastatin	CCC(C)(C)C(=O)OC1CC(C)C=C2C=CC(C)C(CCC3CC(O)CC(=O)O3)C21		N1, N2, N6, N9 except N4EXP, N5EXC, N7EXC, N8EXC R—OH 5-Alcohol R= any atom except for H and some molecular fragments

SAs*: SAs obtained from the ToxAlert platform in the Online Chemical Database (OCHEM, https://ochem.eu/home/show.do accessed on 6 February 2024).

3. Discussion

Conventional QSAR methods predominantly rely on molecular descriptors, such as physicochemical properties and structural fragments, to characterize chemical compounds and correlate them with biological activities. However, deep chemical language models represent a paradigmatic shift in ADR prediction by directly encoding molecular structures from textual representations, such as SMILES strings. These models offer several advantages over conventional QSAR models. Firstly, deep chemical language models provide a more comprehensive representation of molecular structures, capturing intricate structural features inherent in SMILES strings. By encoding the entire molecular structure, including bond connectivity and spatial arrangement, these models can potentially capture subtle but important nuances that traditional descriptors might overlook. This comprehensive representation enhances the model's ability to discern complex relationships between the molecular structure and ADRs. Secondly, deep chemical language models benefit from pre-training on large-scale molecular datasets. Leveraging extensive chemical data, these models can learn universal representations of molecular structures. Pre-training allows the model to capture generalizable patterns and features across diverse chemical compounds, facilitating transfer learning for downstream tasks such as ADR prediction. Furthermore, deep chemical language models offer flexibility and scalability in model architecture. Unlike QSAR methods, which typically rely on handcrafted features and fixed model architectures, deep chemical language models can adapt their architectures to the complexity of prediction tasks.

In this study, we introduced the MoLFormer-XL-CNN model, which combines the deep chemical language model and CNNs for predicting drug-induced adverse reactions. We utilized MoLFormer-XL to encode molecular structures, leveraging its ability to capture intricate structural features. Subsequently, we employed CNNs to predict the risk of adverse reactions. To evaluate the performance of our model, we conducted predictions on three distinct datasets: DIQTA, DITD, and DIRA. Our results demonstrated significant improvements in predictive performance compared to conventional QSAR models in previous studies. Specifically, the MoLFormer-XL-CNN model achieved MCCs of 0.702, 0.503, and 0.535 on the DIQTA, DITD, and DIRA datasets, respectively. These MCC values represent substantial enhancements over the MCCs obtained by conventional QSAR methods in previous studies (MCCs = 0.591, 0.31, and 0.46, respectively) (Figure 1). Notably, the attention mechanism embedded within the MoLFormer-XL framework provided valuable insights into the chemical substructures associated with adverse reactions. By comparing these identified substructures with established SAs, we confirmed their relevance to specific adverse reactions. For instance, the model's attention to amines, ethers, and aromatic halogens corresponded to known SAs for drug-induced QT prolongation. Furthermore, our model identified additional substructures, such as Aniline structures, which may serve as novel SAs for adverse reactions. For drugs with teratogenic or rhabdomyolysis risks, MoLFormer-XL also highlighted similar SAs identified in previous studies. For instance, imine nitrogen and ethane-1,1-diyldibenzene were associated with teratogenicity, while nucleophilic groups in chemical structures were linked to rhabdomyolysis.

Additionally, there are several caveats worth further discussion. Firstly, the stability of the MoLFormer-XL-CNN model varied across datasets, as indicated by the standard deviations of MCC and recall rates (Figure 1). Due to the small sample size of the datasets, we applied data augmentation techniques to improve the model's performance. Secondly, while deep chemical language models can identify important substructures, further research is needed to elucidate the specific associations between individual or combinations of substructures and the mechanisms underlying adverse reactions. Thirdly, the severity stratification of drugs in three distinct datasets is mainly based on the U.S. FDA-approved drug labeling information. If drug labeling from other countries, such as European Union (EU) countries, is used, differences in stratification may lead to differences in prediction results and attended chemical substructures.

4. Materials and Methods

4.1. Study Design

In this study, we aim to investigate the effectiveness of pre-trained deep chemical language models in predicting adverse drug reactions (ADRs) induced by marketed drugs. MoLFormer-XL, a pre-trained model with an efficient linear attention mechanism, was trained on a comprehensive dataset comprising 1.1 billion molecules. To construct the predictive model, we employed MoLFormer-XL to encode canonical SMILES representations of drugs, in conjunction with convolutional neural networks (CNNs), to predict the occurrence of drug-induced QT prolongation (DIQT), drug-induced teratogenicity (DIT), and drug-induced rhabdomyolysis (DIR). We evaluated the performance of our model against previous studies and analyzed the average-pooled attention metrics within the MoLFormer-XL framework. These metrics were instrumental in identifying chemical substructures associated with the ADRs. By extracting the attended substructures from the attention metrics and comparing them with reported structural alerts (SAs), we elucidate the model's ability to discern relationships between molecular structures and ADR propensity. An overview of our methodology is depicted in Figure 2.

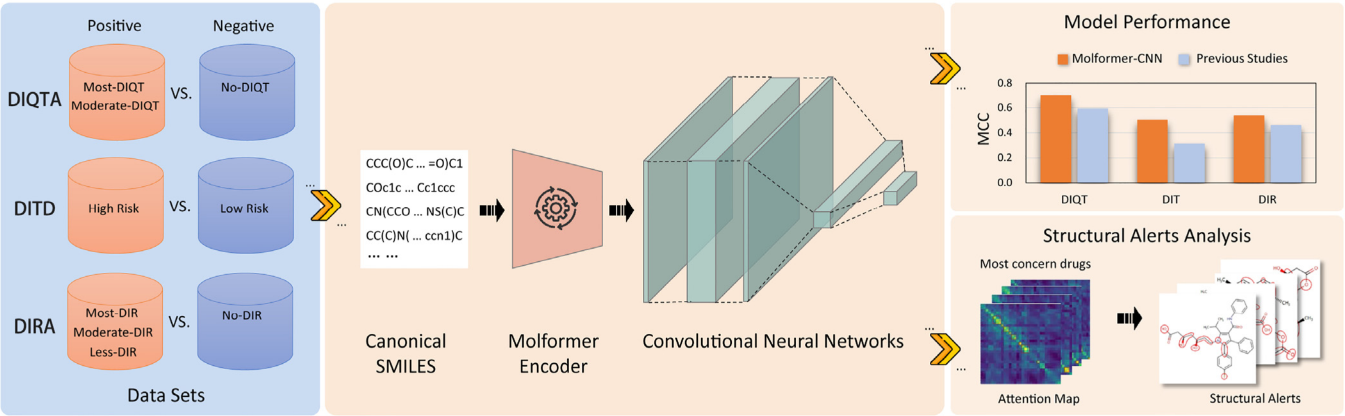


Figure 2. Workflow of our study.

4.2. Data Preparation

We utilized three meticulously curated datasets as follows: the drug-induced QT prolongation atlas (DIQTA, https://www.adratlas.com/DIQTA/download accessed on 6 February 2024) [22], the drug-induced teratogenicity dataset (DITD, https://www.frontiersin.org/articles/10.3389/fphar.2022.747935/full#supplementary-material accessed on 6 February 2024) [12], and the drug-induced rhabdomyolysis atlas (DIRA, https://www.adratlas.com/DIRA/download accessed on 6 February 2024) [23] for model construction and validation. All drugs in these datasets were classified into different concern levels based on the severity outlined in their FDA-approved drug labeling.

Within the DIQTA dataset, QT-prolonging drugs were extracted by full-text searching of the FDA-approved drug labeling with QT prolongation-related keywords. The severity score was assigned based on predefined keywords, a priori knowledge, and DIQT concerns, which were determined based on the severity scores. Marketed drugs were divided into four concern levels: the most DIQT concerns, moderate DIQT concerns, ambiguous levels, and no DIQT concerns. We selected 166 drugs from the most DIQT and moderate DIQT concern levels as positive samples and 100 drugs from the no-DIQT concern level as negative samples. Subsequently, we retrieved the SMILES from the DrugBank database via the link (https://go.drugbank.com/releases/latest accessed on 6 February 2024, release version 5.1.10) [52] using the DrugBank ID. The canonical SMILES were then generated using RDKit packages in Python. Ultimately, we obtained canonical SMILES for a total of 255 drugs, comprising 156 positive samples and 99 negative samples.

In the DITD dataset, a drug was categorized into high teratogenic risk if it was indicated in drug labeling that adequate and well-controlled studies or animal studies had shown a teratogenic risk to the fetus, while a drug was categorized into low teratogenic risk if both animal and human studies showed no risk to the fetus. Then, drugs were categorized into high teratogenic risk and low teratogenic risk groups based on the FDA's pregnancy medication classification. We utilized 67 drugs from the high teratogenic risk group as positive samples and 45 drugs from the low teratogenic risk group as negative samples. Similar to the DIQTA dataset, SMILES were obtained from the DrugBank database and converted to canonical SMILES using RDKit packages in Python. Consequently, we acquired canonical SMILES for a total of 112 drugs, including 67 positive samples and 45 negative samples.

Within the DIRA dataset, DIR drugs were extracted by full-text searching of the FDA-approved drug labeling with the keyword 'rhabdomyolysis'. The severity score was assigned based on predefined keywords and a priori knowledge, and DIR concerns were determined based on the severity scores. Marketed drugs were classified into four concern levels: the most DIR concerns, moderate DIR concerns, less DIR concerns, and no DIR concern drugs. We designated 173 drugs with DIR concerns as positive samples and 40 no-DIR concern drugs as negative samples. Similarly, SMILES was retrieved from the DrugBank database and converted to canonical SMILES using RDKit packages in Python. Finally, we obtained canonical SMILES for a total of 194 drugs, comprising 155 positive samples and 39 negative samples.

4.3. Model Construction

MoLFormer is an efficient transformer encoder model that utilizes rotary positional embeddings and the linear attention mechanism. Among various MoLFormer variants, MoLFormer-XL exhibited superior performance. Constructed by training on 1.1 billion unlabeled molecules from the PubChem and ZINC datasets, MoLFormer-XL aims to learn meaningful and universal representations of chemical molecules from large-scale chemical representation data, subsequently fine-tuning for various downstream molecular property prediction tasks. In our study, we leveraged the pre-trained MoLFormer-XL to encode molecules, enhancing the capture of molecular features of marketed drugs. Subsequently, we employed CNN for further prediction tasks.

MoLFormer-XL encoded each canonical SMILES, with the resulting embedding serving as an input to the CNN model. Our CNN framework comprises two convolutional layers and two fully connected layers. The first convolutional layer expands the dimensionality of the embedding threefold to capture richer information, utilizing a kernel size of 3 × 1 and padding of 1 to maintain the spatial dimensionality of features. The second convolutional layer compresses the feature space back to the original embedding dimension for subsequent fully connected operations, with kernel size and padding remaining unchanged. Both convolutional layers incorporate batch normalization, aiming to re-learn the original encoded features at a deeper level, thereby enhancing subsequent classification and predictive accuracy. In the two fully connected layers, we introduced a dropout layer to randomly skip some neurons to prevent overfitting. Functions for optimizer, loss, activation, and output were separately set to AdamW, CrossEntropyLoss, ReLU, and softmax.

To facilitate a comparison with the predictive outcomes in previous studies, our model's construction employed fivefold cross-validation. To enhance the model's learning capacity during training, we utilized the SMILES enumeration method [53], which generates multiple valid representations for chemical structures from a single SMILES string. This approach enriches the diversity of structural representations and has been shown in prior studies to assist in improving the accuracy of the model's predictions for molecular properties. The Python codes for the MoLFormer-XL-CNN model were uploaded to GitHub (https://github.com/LiSH7450/MoLFormer-XL-CNN_model accessed on 6 February 2024).

4.4. Evaluation of Model Performance

To evaluate the performance of the model, we used nine performance metrics in this study, including accuracy, recall rate, precision, Matthew's correlation coefficient (MCC), balanced accuracy score (BACC), F1 score, the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC) and specificity.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{recallrate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

$$\text{BACC} = \frac{(\text{TPR} + \text{TNR})}{2} \tag{5}$$

$$\text{F1score} = \frac{2 \times (\text{precision} \times \text{recallrate})}{\text{precision} + \text{recallrate}} \tag{6}$$

$$\text{AUROC} = \int_{x=0}^{1} \text{TPR}\left(\text{FPR}^{-1}(x)\right)dx \tag{7}$$

$$\text{AUPRC} = \int_{-\infty}^{+\infty} \text{precision}(x)dP[Y \leq x] \tag{8}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

5. Conclusions

In conclusion, our study underscores the potential of deep chemical language models to predict ADRs induced by marketed drugs. Our MoLFormer-XL-CNN model outperformed conventional QSAR approaches across multiple datasets, showcasing its efficacy in ADR prediction with enhanced structural representation and predictive accuracy. These findings emphasize the transformative impact of deep chemical language models, not only in pharmacovigilance but also in facilitating early ADR detection during drug development.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25084516/s1.

Author Contributions: Conceptualization, Z.W.; methodology, J.L. and Y.H.; validation, J.L., Y.H. and W.L.; investigation, J.L., W.L. and C.R.; writing—original draft preparation, J.L. and Z.W.; writing—review and editing, J.L., M.L. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 22173065 and 21575094. The APC was funded by 22173065.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://github.com/LiSH7450/MoLFormer-XL-CNN_model accessed on 6 February 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tan, Y.; Hu, Y.; Liu, X.; Yin, Z.; Chen, X.-W.; Liu, M. Improving drug safety: From adverse drug reaction knowledge discovery to clinical implementation. *Methods* **2016**, *110*, 14–25. [CrossRef] [PubMed]
2. Poudel, D.R.; Acharya, P.; Ghimire, S.; Dhital, R.; Bharati, R. Burden of hospitalizations related to adverse drug events in the USA: A retrospective analysis from large inpatient database. *Pharmacoepidemiol. Drug Saf.* **2017**, *26*, 635–641. [CrossRef] [PubMed]
3. Hohenegger, M. Drug induced rhabdomyolysis. *Curr. Opin. Pharmacol.* **2012**, *12*, 335–339. [CrossRef] [PubMed]
4. Hur, J.; Liu, Z.; Tong, W.; Laaksonen, R.; Bai, J.P.F. Drug-Induced Rhabdomyolysis: From Systems Pharmacology Analysis to Biochemical Flux. *Chem. Res. Toxicol.* **2014**, *27*, 421–432. [CrossRef] [PubMed]
5. Stahl, K.; Rastelli, E.; Schoser, B. A systematic review on the definition of rhabdomyolysis. *J. Neurol.* **2019**, *267*, 877–882. [CrossRef] [PubMed]

6. Griffin, M.R.; Stein, C.M.; Ray, W.A. Postmarketing surveillance for drug safety: Surely we can do better. *Clin. Pharmacol. Ther.* **2004**, *75*, 491–494. [CrossRef] [PubMed]
7. Alomar, M.; Tawfiq, A.M.; Hassan, N.; Palaian, S. Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: Current status, challenges and the future. *Ther. Adv. Drug Saf.* **2020**, *11*, 2042098620938595. [CrossRef] [PubMed]
8. Polishchuk, P.A.-O. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639. [CrossRef] [PubMed]
9. Ma, X.H.; Wang, R.; Xue, Y.; Li, Z.R.; Yang, S.Y.; Wei, Y.Q.; Chen, Y.Z. Advances in machine learning prediction of toxicological properties and adverse drug reactions of pharmaceutical agents. *Curr. Drug Saf.* **2008**, *3*, 100–114. [CrossRef]
10. Patlewicz, G.; Fitzpatrick, J.M. Current and Future Perspectives on the Development, Evaluation, and Application of in Silico Approaches for Predicting Toxicity. *Chem. Res. Toxicol.* **2016**, *29*, 438–451. [CrossRef]
11. Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Mol. Inform.* **2016**, *35*, 136–144. [CrossRef] [PubMed]
12. Kang, L.; Duan, Y.; Chen, C.; Li, S.; Li, M.; Chen, L.; Wen, Z. Structure-Activity Relationship (SAR) Model for Predicting Teratogenic Risk of Antiseizure Medications in Pregnancy by Using Support Vector Machine. *Front. Pharmacol.* **2022**, *13*, 747935. [CrossRef] [PubMed]
13. Long, W.; Li, S.; He, Y.; Lin, J.; Li, M.; Wen, Z. Unraveling Structural Alerts in Marketed Drugs for Improving Adverse Outcome Pathway Framework of Drug-Induced QT Prolongation. *Int. J. Mol. Sci.* **2023**, *24*, 6771. [CrossRef] [PubMed]
14. Yap, C.W.; Cai, C.Z.; Xue, Y.; Chen, Y.Z. Prediction of Torsade-Causing Potential of Drugs by Support Vector Machine Approach No funding was used to assist in conducting the study and the authors do not have any conflicts of interest directly relevant to the contents of the manuscript. *Toxicol. Sci.* **2004**, *79*, 170–177. [CrossRef] [PubMed]
15. Zhu XW, L.S. In Silico Prediction of Drug-Induced Liver Injury Based on Adverse Drug Reaction Report. *Toxicol. Sci.* **2017**, *158*, 391–400. [CrossRef]
16. Zhou, Y.; Li, S.; Zhao, Y.; Guo, M.; Liu, Y.; Li, M.; Wen, Z. Quantitative Structure–Activity Relationship (QSAR) Model for the Severity Prediction of Drug-Induced Rhabdomyolysis by Using Random Forest. *Chem. Res. Toxicol.* **2021**, *34*, 514–521. [CrossRef]
17. Jamal, S.; Ali, W.; Nagpal, P.; Grover, S.; Grover, A. Computational models for the prediction of adverse cardiovascular drug reactions. *J. Transl. Med.* **2019**, *17*, 171. [CrossRef]
18. Hong, H.; Thakkar, S.; Chen, M.; Tong, W. Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. *Sci. Rep.* **2017**, *7*, 17311. [CrossRef]
19. Chen, M.; Hong, H.; Fang, H.; Kelly, R.; Zhou, G.; Borlak, J.; Tong, W. Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs. *Toxicol. Sci.* **2013**, *136*, 242–249. [CrossRef]
20. Minerali, E.; Foil, D.H.; Zorn, K.M.; Lane, T.R.; Ekins, S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Mol. Pharm.* **2020**, *17*, 2628–2637. [CrossRef]
21. Frid, A.A.; Matthews, E.J. Prediction of drug-related cardiac adverse effects in humans-B: Use of QSAR programs for early detection of drug-induced cardiac toxicities. *Regul. Toxicol. Pharmacol.* **2010**, *56*, 276–289. [CrossRef] [PubMed]
22. Li, S.; Xu, Z.; Guo, M.; Li, M.; Wen, Z. Drug-induced QT Prolongation Atlas (DIQTA) for enhancing cardiotoxicity management. *Drug Discov. Today* **2022**, *27*, 831–837. [CrossRef] [PubMed]
23. Wen, Z.; Liang, Y.; Hao, Y.; Delavan, B.; Huang, R.; Mikailov, M.; Tong, W.; Li, M.; Liu, Z. Drug-Induced Rhabdomyolysis Atlas (DIRA) for idiosyncratic adverse drug reaction management. *Drug Discov. Today* **2019**, *24*, 9–15. [CrossRef] [PubMed]
24. Lee, C.Y.; Chen, Y.-P.P. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Brief. Bioinform.* **2021**, *22*, 1884–1901. [CrossRef] [PubMed]
25. Mostafa, F.; Chen, M. Computational models for predicting liver toxicity in the deep learning era. *Front. Toxicol.* **2024**, *5*, 1340860. [CrossRef] [PubMed]
26. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef] [PubMed]
27. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692–1701. [CrossRef] [PubMed]
28. Le, T.; Winter, R.; Noé, F.; Clevert, D.-A. Neuraldecipher–reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chem. Sci.* **2020**, *11*, 10378–10389. [CrossRef] [PubMed]
29. Clevert, D.-A.; Le, T.; Winter, R.; Montanari, F. Img2Mol–accurate SMILES recognition from molecular graphical depictions. *Chem. Sci.* **2021**, *12*, 14174–14181. [CrossRef]
30. Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration. *Research* **2022**, *2022*, 4. [CrossRef]
31. Li, X.; Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1560–1569. [CrossRef]

32. Wu, C.-K.; Wu, C.-K.; Zhang, X.-C.; Zhang, X.-C.; Yang, Z.-J.; Yang, Z.-J.; Lu, A.-P.; Lu, A.-P.; Hou, T.-J.; Hou, T.-J.; et al. Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules. *Brief. Bioinform.* **2021**, *22*, bbab327. [CrossRef] [PubMed]
33. Ucak, U.V.; Ashyrmamatov, I.; Lee, J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *J. Cheminformatics* **2023**, *15*, 55. [CrossRef]
34. Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **2022**, *4*, 1256–1264. [CrossRef]
35. Colatsky, T.; Fermini, B.; Gintant, G.; Pierson, J.B.; Sager, P.; Sekino, Y.; Strauss, D.G.; Stockbridge, N. The Comprehensive in vitro Proarrhythmia Assay (CiPA) initiative—Update on progress. *J. Pharmacol. Toxicol. Methods* **2016**, *81*, 15–20. [CrossRef]
36. Sushko, I.; Salmina, E.; Potemkin, V.A.; Poda, G.; Tetko, I.V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316. [CrossRef]
37. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V.V.; Tanchuk, V.Y.; et al. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554. [CrossRef] [PubMed]
38. Kalgutkar, A.; Dalvie, D.; O'Donnell, J.; Taylor, T.; Sahakian, D. On the Diversity of Oxidative Bioactivation Reactions on Nitrogen-Containing Xenobiotics. *Curr. Drug Metab.* **2002**, *3*, 379–424. [CrossRef] [PubMed]
39. Kalgutkar, A.S. Designing around Structural Alerts in Drug Discovery. *J. Med. Chem.* **2019**, *63*, 6276–6302. [CrossRef]
40. Islam, I.; Skibo, E.B.; Dorr, R.T.; Alberts, D.S. Structure-activity studies of antitumor agents based on pyrrolo[1,2-a]benzimidazoles: New reductive alkylating DNA cleaving agents. *J. Med. Chem.* **1991**, *34*, 2954–2961. [CrossRef]
41. Tomson, T.; Battino, D.; Bonizzoni, E.; Craig, J.; Lindhout, D.; Sabers, A.; Perucca, E.; Vajda, F. Dose-dependent risk of malformations with antiepileptic drugs: An analysis of data from the EURAP epilepsy and pregnancy registry. *Lancet Neurol.* **2011**, *10*, 609–617. [CrossRef] [PubMed]
42. Hernández-Díaz, S.; Smith, C.R.; Shen, A.; Mittendorf, R.; Hauser, W.A.; Yerby, M.; Holmes, L.B.; For the North American AED Pregnancy Registry; Cragan, J.; Hauser, A.; et al. Comparative safety of antiepileptic drugs during pregnancy. *Neurology* **2012**, *78*, 1692–1699. [CrossRef] [PubMed]
43. Campbell, E.; Kennedy, F.; Russell, A.; Smithson, W.H.; Parsons, L.; Morrison, P.J.; Liggan, B.; Irwin, B.; Delanty, N.; Hunt, S.J.; et al. Malformation risks of antiepileptic drug monotherapies in pregnancy: Updated results from the UK and Ireland Epilepsy and Pregnancy Registers. *J. Neurol. Neurosurg. Psychiatry* **2014**, *85*, 1029–1034. [CrossRef] [PubMed]
44. Tomson, T.; Battino, D.; Bonizzoni, E.; Craig, J.; Lindhout, D.; Perucca, E.; Sabers, A.; Thomas, S.V.; Vajda, F.; Faravelli, F.; et al. Comparative risk of major congenital malformations with eight different antiepileptic drugs: A prospective cohort study of the EURAP registry. *Lancet Neurol.* **2018**, *17*, 530–538. [CrossRef] [PubMed]
45. Kalgutkar, A.S.; Gardner, I.; Obach, R.S.; Shaffer, C.L.; Callegari, E.; Henne, K.R.; Mutlib, A.E.; Dalvie, D.K.; Lee, J.S.; Nakai, Y.; et al. A comprehensive listing of bioactivation pathways of organic functional groups.pdf. *Curr. Drug Metab.* **2005**, *6*, 161–225. [CrossRef]
46. Nendza, M.; Wenzel, A.; Müller, M.; Lewin, G.; Simetska, N.; Stock, F.; Arning, J. Screening for potential endocrine disruptors in fish: Evidence from structural alerts and in vitro and in vivo toxicological assays. *Environ. Sci. Eur.* **2016**, *28*, 26. [CrossRef] [PubMed]
47. Stepan, A.F.; Walker, D.P.; Bauman, J.; Price, D.A.; Baillie, T.A.; Kalgutkar, A.S.; Aleo, M.D. Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24*, 1345–1410. [CrossRef]
48. Bitzur, R.; Cohen, H.; Kamari, Y.; Harats, D. Intolerance to Statins: Mechanisms and Management. *Diabetes Care* **2013**, *36*, S325–S330. [CrossRef]
49. Needham, M.; Mastaglia, F.L. Statin myotoxicity: A review of genetic susceptibility factors. *Neuromuscul. Disord.* **2014**, *24*, 4–15. [CrossRef]
50. Liebler, D.C.; Guengerich, F.P. Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discov.* **2005**, *4*, 410–420. [CrossRef]
51. Benigni, R.; Bossa, C. Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutat. Res./Rev. Mutat. Res.* **2008**, *659*, 248–261. [CrossRef] [PubMed]
52. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef] [PubMed]
53. Bjerrum, E.J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv* **2017**, arXiv:1703.07076. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.