Learning Molecular Representations from SMILES for Acute Toxicity Prediction

Simona Cholakova
University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies (UP FAMNIT)
Koper, Slovenia
simonacholakova04@gmail.com

Domen Vake
University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies (UP FAMNIT)
Koper, Slovenia
domen.vake@famnit.upr.si

Aleksandar Tošić
University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies (UP FAMNIT)
Koper, Slovenia
aleksandar.tosic@upr.si

Abstract

Predicting molecular toxicity from structure can reduce the cost and risk of experimental screening, but performance depends strongly on the molecular representation supplied to the predictor. We study whether embedding representations learned directly from SMILES strings, text-based descriptions of a molecule's chemical structure, can support acute toxicity prediction. A custom byte-pair tokenizer and six-layer Transformer encoder were pretrained by masked language modelling on five million unlabeled PubChem molecules, then fine-tuned independently on seven rat and mouse LD50 endpoints from TOXRIC. The resulting molecule-level embeddings were evaluated with the encoder's own regression head, and separately with an XGBoost regressor trained on the embeddings alone or concatenated with Morgan fingerprints. XGBoost achieved lower RMSE than the neural regression head on all seven endpoints, with reductions of 4.4–7.5%. Morgan fingerprints improved R^2 on six endpoints, although only modestly, and randomized SMILES augmentation improved validation loss in every task. The results show that self-supervised SMILES embeddings are useful features for toxicity modelling. Qualitative error analysis further suggests potential limitations for rare atoms, complex ring systems, and long sequences.

Keywords

acute toxicity, molecular representation, SMILES, self-supervised learning, Transformer, XGBoost

1 Introduction

Experimental toxicity testing is costly, slow, and potentially harmful. Quantitative structure-activity relationship (QSAR) models offer a complementary route: estimate a biological endpoint directly from molecular structure before experimental screening. Acute toxicity is commonly summarized by LD_{50}, the dose that is lethal to half of a test population. Because LD_{50} is a measurable endpoint available for multiple species and administration routes, it is also a useful benchmark for molecular machine learning.

The representation of a molecule largely determines what a QSAR model can learn. Classical descriptors and fingerprints encode predefined substructures efficiently, but may omit task-relevant interactions [10]. Learned representations instead optimize features from data. SMILES strings are especially attractive because they express molecular graphs as sequences and can therefore be processed by methods developed for language. Transformers [9] and masked language modelling [4] have already shown that large collections of unlabeled sequences can yield transferable representations; ChemBERTa extended this principle to molecular property prediction [3], pretraining on tens of millions of PubChem SMILES with a general-purpose subword tokenizer and evaluating its embeddings largely through an end-to-end prediction head. This leaves open whether a smaller, chemistry-aware encoder, evaluated separately from its own prediction head, produces embeddings that a downstream model can exploit as well or better, a question with direct practical value where labelled data and compute are limited. Large pretrained models do not automatically dominate fingerprint or graph baselines, and the value of their embeddings depends on both the downstream learner and the amount of labelled data.

We investigate a compact pipeline learned from raw molecular strings. A custom byte-pair encoding (BPE) tokenizer and Transformer encoder are pretrained on unlabeled PubChem SMILES, fine-tuned for seven acute toxicity endpoints, and then used as a fixed feature extractor for XGBoost. We evaluate three questions: (1) whether a tree ensemble can exploit the learned embedding better than the encoder's neural regression head; (2) whether Morgan fingerprints add complementary information; and (3) whether randomized SMILES augmentation helps when labelled datasets are limited. The study makes the following contributions:

- an end-to-end, self-supervised SMILES representation pipeline for LD_{50} regression;
- a controlled comparison of prediction heads and molecular feature combinations across seven TOXRIC endpoints; and
- an analysis of data augmentation, benchmark comparability, and systematic model failures.

2 Related Work

Molecular representations trade computational simplicity against structural detail. Fixed descriptors summarize selected physicochemical properties, while fingerprints encode the presence of local substructures in a fixed-length vector. Graph and three-dimensional methods preserve richer connectivity or geometry, but introduce additional modelling or conformer-generation costs. No representation is universally best: its value depends on the endpoint, available data, and downstream learner [10]. Our experiments therefore compare a learned sequential representation both alone and in combination with a widely used fingerprint.

Self-supervised molecular language models adapt techniques from natural language processing to chemical strings. ChemBERTa [3], for example, pretrained RoBERTa-style encoders on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.88

tens of millions of PubChem SMILES and demonstrated transferable molecular features. Its results also showed that scale alone does not guarantee superiority over graph or fingerprint baselines, but its evaluation left the learned representation and the end-to-end prediction head entangled, so it remains unclear how much of the reported performance reflects the embedding itself. We address this gap directly: our smaller, chemistry-aware encoder (a 512-token task-specific BPE vocabulary vs. ChemBERTa's general-purpose 52K-token vocabulary, pretrained on 5M vs. 77M PubChem molecules) is frozen after fine-tuning and its embeddings are handed to an independently tuned XGBoost model, isolating representation quality from the small regression head used during fine-tuning.

Acute toxicity prediction has been studied with descriptor-based QSAR and deep multi-task models. Jain et al. [5] shared information across species and exposure routes through consensus multi-task learning. TOXRIC later consolidated toxicity data and evaluated multiple model and descriptor combinations [11]. In contrast, we train one encoder per endpoint to measure representation behaviour without cross-task transfer, and use the published TOXRIC values only as contextual benchmarks because the validation protocols differ.

3 Methods

3.1 Data and tasks

Unlabeled pretraining data were drawn from PubChem [6]. We used five million canonicalized SMILES strings from the public PubChem-10M collection. No toxicity labels were exposed during this stage. Labelled data came from TOXRIC [11], a database that integrates toxicological measurements and published QSAR benchmarks. We selected all acute toxicity endpoints containing more than 2,000 compounds: oral, intravenous, and intraperitoneal LD_{50} for rats; and oral, subcutaneous, intravenous, and intraperitoneal LD_{50} for mice. Their sizes range from 2,323 to 35,299 compounds. Each endpoint was treated as an independent regression task using the log-transformed LD_{50} provided by TOXRIC.

For each endpoint, compounds were randomly divided into 70% training, 15% validation, and 15% held-out test sets with a fixed seed. Augmentation was applied only to the training split; validation and test molecules retained one canonical SMILES each.

3.2 Representation learning

Figure 1 summarizes the pipeline. Before BPE training [7], SMILES were split into chemically meaningful primitive units, including multicharacter atoms, bracketed atoms, bonds, branches, and ring closures. Frequent adjacent units were iteratively merged into a 512-token vocabulary, with reserved `[PAD]`, `[CLS]`, `[UNK]`, and `[MASK]` tokens. Sequences were prepended with `[CLS]` and padded or truncated to 128 tokens.

The encoder comprised six Transformer layers, together with a 512-dimensional hidden state, eight attention heads, a 1,024-dimensional feed-forward block, 0.15 dropout, and rotary positional embeddings [8]. During masked language model pretraining, 15% of non-special tokens were selected: 80% were replaced by `[MASK]`, 10% by a random token, and 10% left unchanged. Cross-entropy was computed only at selected positions. AdamW training used a learning rate of 10^{-4}, weight decay 0.01, batch size 256, and 15 epochs.

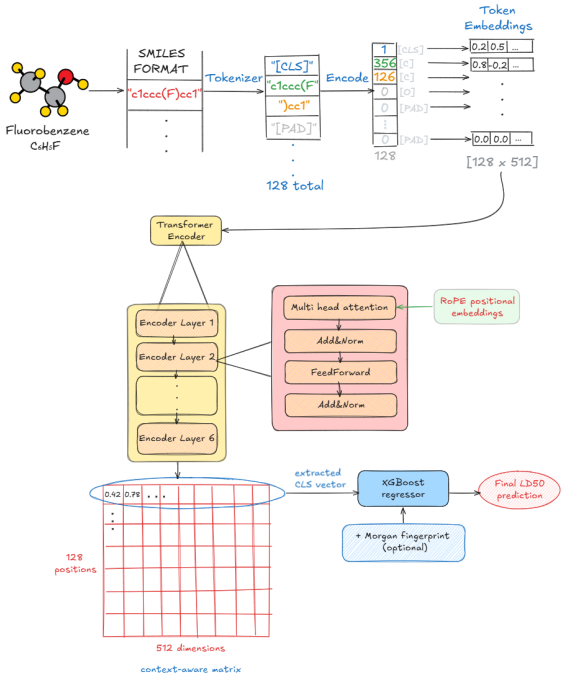


Figure 1: Pipeline from a SMILES string to the learned `[CLS]` embedding and final LD_{50} prediction. Morgan fingerprints are an optional additional feature.

3.3 Fine-tuning and prediction

The pretrained encoder was fine-tuned separately for each endpoint. A regression head mapped the 512-dimensional `[CLS]` state through a 256-unit ReLU layer and dropout to one LD_{50} value. Training minimized mean squared error with Adam, a learning rate of 10^{-4}, weight decay 5×10^{-5}, learning-rate reduction after three stagnant validation epochs, and early stopping with patience eight (maximum 50 epochs).

To increase invariance to SMILES syntax, we generated up to eight non-canonical forms per molecule using RDKit atom renumbering and `canonical=False` serialization. Identical fine-tuning runs with and without augmentation measured its effect.

After fine-tuning, the encoder was frozen; for each molecule, its canonical SMILES was passed through the encoder in evaluation mode with dropout disabled, and the final layer's [CLS] hidden state was taken as the 512-dimensional feature vector. XGBoost [2] was trained on (i) the 512 learned features and (ii) those features concatenated with a 2,048-bit radius-2 Morgan fingerprint. Training and validation features were combined; model selection used five-fold grid search over 500–2,000 estimators, depths 4–8, and learning rates 0.01–0.1. The held-out test set was evaluated once. We report root mean squared error (RMSE) and coefficient of determination (R^2).

3.4 Evaluation protocol

The neural head, XGBoost on `[CLS]`, and XGBoost on `[CLS]` plus Morgan features used identical endpoint splits. RMSE measures error on the scale of the log-transformed target and emphasizes large deviations; R^2 measures the explained test-set variance. Each result is from the untouched 15% test partition after model selection. The augmentation ablation instead compares the lowest validation mean squared error reached by otherwise identical fine-tuning runs. For external context, we compare with TOXRIC's

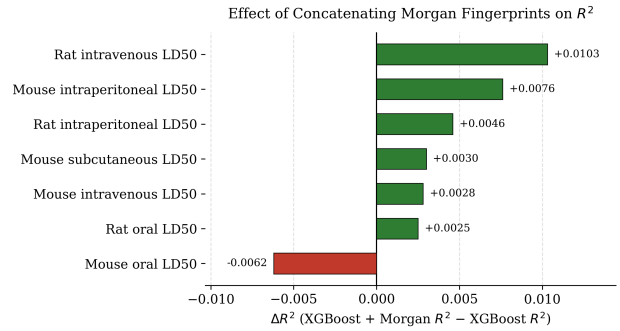


Figure 2: Change in R^2 after concatenating Morgan fingerprints with learned `[CLS]` embeddings.

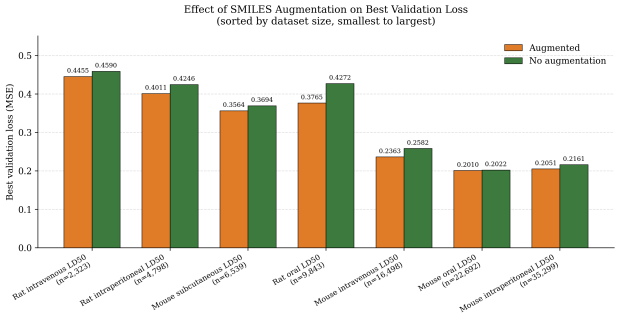


Figure 3: Best fine-tuning validation loss with and without randomized SMILES augmentation, ordered by endpoint size.

published results, but treat that comparison separately because TOXRIC reports five-fold cross-validation rather than our fixed holdout.

4 Results and Discussion

4.1 Prediction head and feature comparison

Table 1 reports held-out performance. XGBoost using learned embeddings outperformed the encoder's prediction head on all seven endpoints. Relative RMSE reductions ranged from 4.4% for mouse subcutaneous LD_{50} to 7.5% for rat oral LD_{50}. The result indicates that supervised fine-tuning produces an informative molecular embedding even when the small end-to-end regression head does not extract its full predictive value.

Adding Morgan fingerprints improved R^2 on six endpoints, but the gains were small (0.0025–0.0103). The largest increase occurred for rat intravenous LD_{50}, the smallest dataset; mouse oral LD_{50} was the sole negative case ($\Delta R^2 = -0.0062$). The largest gain occurred on the smallest endpoint, suggesting that classical fingerprints may be particularly useful as a complementary representation in lower-data settings. Learned and predefined representations capture overlapping, but not identical, structural information [10].

Dataset size alone did not determine difficulty. Rat intravenous had the highest RMSE and the fewest compounds, yet mouse oral had the lowest XGBoost R^2 despite being the second-largest endpoint. These results indicate that sample count alone does not explain task difficulty since differences in endpoint heterogeneity and chemical-space coverage may also contribute to it.

4.2 SMILES augmentation

Randomized SMILES augmentation reduced the best validation loss for every endpoint (Figure 3). Gains were generally larger with fewer labelled molecules. Rat oral improved by 11.9% (0.4272 to 0.3765) and mouse intravenous by 8.5% (0.2582 to 0.2363), whereas mouse oral improved by only 0.6% (0.2022 to 0.2010). Multiple valid serializations [1] encourage the encoder to model molecular structure rather than memorize one textual form, with greatest value when distinct labelled molecules are scarce.

4.3 Published benchmarks and failure modes

For context, we compared against TOXRIC's published XGBoost, random forest, support vector machine, and deep neural network results, which use PubChem fingerprints plus RDKit descriptors. Table 2 contrasts our best feature configuration per endpoint with the best of those four published models. Our approach achieved lower RMSE for mouse intravenous, rat oral, and rat intraperitoneal LD_{50}; the published models retained better RMSE on the other four tasks and better R^2 throughout. The R^2 gaps are particularly visible for the four mouse endpoints. This is only an approximate comparison: TOXRIC reports means from five-fold cross-validation, whereas our numbers use one held-out split. Direct claims of superiority would require identical folds and repeated fine-tuning.

Inspection of the three largest-error test molecules per endpoint revealed recurring limitations. Errors clustered around large fused-ring systems, compounds containing comparatively rare atoms such as silicon or iodine, and long SMILES strings (roughly 60–80 characters). The first two patterns suggest gaps in pretraining and endpoint coverage. Truncation was not the cause of the long-SMILES pattern in the present data: across all seven endpoints, every molecule fit within the 128-token limit (maximum 110 tokens, 0% truncated), so this difficulty likely stems from the greater structural complexity such strings tend to encode rather than from lost sequence information. However, this sequence length was sufficient only for the compounds in the current TOXRIC endpoints; larger or more structurally diverse molecular datasets could exceed it, so the 128-token limit should be revisited if the pipeline is extended to other chemical spaces. These observations motivate scaffold-aware evaluation, longer or structure-aware tokenization, and pretraining that better balances chemical space.

4.4 Limitations and validity

Three limitations constrain interpretation. First, random compound splits can place close structural analogues in both training and test sets; a scaffold-based split would more rigorously test extrapolation to unfamiliar chemistry. Second, results come from a single fixed train/validation/test split, so we cannot yet separate genuine representational gains from split-specific variance. We deliberately scope both a scaffold-aware split and repeated-seed or k-fold evaluation out of the present study, in favor of first establishing the XGBoost-versus-neural-head comparison and augmentation effects under controlled, identical conditions; we treat repeated, scaffold-aware cross-validation as the immediate next step, both to quantify this variance and to enable a fold-matched comparison with TOXRIC's published 5-fold results, rather than the current approximate one. Third, each endpoint was modelled independently, although related species and administration routes may contain transferable signal. The reported advantage of XGBoost over the neural head is controlled within

Table 1: Held-out test performance. "Head" is the fine-tuned neural regression head; "CLS" uses learned embeddings with XGBoost; "+Morgan" concatenates Morgan fingerprints. Best RMSE and R^2 within each endpoint are bold.

		Head		CLS		CLS + Morgan	
Endpoint	n	RMSE	R^2	RMSE	R^2	RMSE	R^2
Mouse oral	22,692	0.4741	0.3994	**0.4422**	**0.4775**	0.4448	0.4713
Mouse subcutaneous	6,539	0.6382	0.5098	0.6099	0.5523	**0.6078**	**0.5553**
Mouse intraperitoneal	35,299	0.4768	0.5277	0.4513	0.5768	**0.4473**	**0.5844**
Mouse intravenous	16,498	0.4811	0.5489	0.4490	0.6070	**0.4475**	**0.6098**
Rat oral	9,843	0.6227	0.5317	0.5761	0.5991	**0.5743**	**0.6016**
Rat intravenous	2,323	0.6936	0.5392	0.6481	0.5976	**0.6398**	**0.6079**
Rat intraperitoneal	4,798	0.6436	0.4538	0.6033	0.5200	**0.6004**	**0.5246**

Table 2: Contextual comparison with the best result among TOXRIC's four published models. Our result is the better of CLS and CLS + Morgan. Bold denotes the better value in each pair; protocols differ, so values are not a controlled ranking.

	RMSE (lower is better)		R^2 (higher is better)	
Endpoint	Ours	Best TOXRIC	Ours	Best TOXRIC
Mouse oral	0.4422	**0.4239**	0.4775	**0.6258**
Mouse subcutaneous	0.6078	**0.5658**	0.5553	**0.6212**
Mouse intraperitoneal	0.4473	**0.4407**	0.5844	**0.7047**
Mouse intravenous	**0.4475**	0.4557	0.6098	**0.7264**
Rat oral	**0.5743**	0.5759	0.6016	**0.6103**
Rat intravenous	0.6398	**0.6179**	0.6079	**0.6652**
Rat intraperitoneal	**0.6004**	0.6094	0.5246	**0.5481**

each split, but broader claims about representation quality require these additional evaluations. Finally, predicted LD_{50} values are screening aids, not substitutes for expert toxicological assessment; applicability-domain checks are especially important for the rare and complex molecules highlighted by the error analysis.

5 Conclusion

A Transformer trained by masked language modelling on unlabeled SMILES produces useful features for acute toxicity prediction. Across seven TOXRIC endpoints, XGBoost consistently exploited those features better than the encoder's neural regression head. Randomized SMILES augmentation reduced the best validation loss for every endpoint, while Morgan fingerprints provided small improvements on six endpoints and were most beneficial for the smallest dataset. The approach was competitive with published descriptor-based models, but did not dominate them across metrics and evaluation protocols. Future work should use scaffold-based, repeated cross-validation; multi-task learning across related species and administration routes [5]; and representations designed for long or underrepresented molecular structures. With the scaffold-aware and multi-task extensions outlined above, this pipeline could serve as a reusable foundation for broader molecular property prediction beyond acute toxicity. Code, the trained BPE tokenizer, and pretrained encoder weights (for both the CLS-only and CLS+Morgan configurations) are publicly available at https://github.com/simona-cholakova/molecular-representation-learning.

References

[1] Esben Jannik Bjerrum. 2017. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv preprint arXiv:1703.07076* (2017). doi:10.48550/arXiv.1703.07076

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. doi:10.1145/2939672.2939785

[3] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv preprint arXiv:2010.09885* (2020). doi:10.48550/arXiv.2010.09885

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186. doi:10.18653/v1/N19-1423

[5] Sankalp Jain, Vishal B. Siramshetty, Vinicius M. Alves, Eugene N. Muratov, Nicole Kleinstreuer, Alexander Tropsha, Marc C. Nicklaus, Anton Simeonov, and Alexey V. Zakharov. 2021. Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. *Journal of Chemical Information and Modeling* 61, 2 (2021), 653–663. doi:10.1021/acs.jcim.0c01174

[6] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2023. PubChem 2023 Update. *Nucleic Acids Research* 51, D1 (2023), D1373–D1380. doi:10.1093/nar/gkac956

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1715–1725. doi:10.18653/v1/P16-1162

[8] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing* 568 (2024), 127063. doi:10.1016/j.neucom.2023.127063

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[10] Daniel S. Wigh, Jonathan M. Goodman, and Alexei A. Lapkin. 2022. A Review of Molecular Representation in the Age of Machine Learning. *WIREs Computational Molecular Science* 12, 5 (2022), e1603. doi:10.1002/wcms.1603

[11] Lianlian Wu, Bowei Yan, Junshan Han, Ruijiang Li, Jian Xiao, Song He, and Xiaochen Bo. 2023. TOXRIC: A Comprehensive Database of Toxicological Data and Benchmarks. *Nucleic Acids Research* 51, D1 (2023), D1432–D1445. doi:10.1093/nar/gkac1074