molecules

*Article*

# Traditional Machine and Deep Learning for Predicting Toxicity Endpoints

**Ulf Norinder**

Department of Computer and Systems Sciences, Stockholm University, 164 07 Kista, Sweden; ulfn@dsv.su.se

**Abstract:** Molecular structure property modeling is an increasingly important tool for predicting compounds with desired properties due to the expensive and resource-intensive nature and the problem of toxicity-related attrition in late phases during drug discovery and development. Lately, the interest for applying deep learning techniques has increased considerably. This investigation compares the traditional physico-chemical descriptor and machine learning-based approaches through autoencoder generated descriptors to two different descriptor-free, Simplified Molecular Input Line Entry System (SMILES) based, deep learning architectures of Bidirectional Encoder Representations from Transformers (BERT) type using the Mondrian aggregated conformal prediction method as overarching framework. The results show for the binary CATMoS non-toxic and very-toxic datasets that for the former, almost equally balanced, dataset all methods perform equally well while for the latter dataset, with an 11-fold difference between the two classes, the MolBERT model based on a large pre-trained network performs somewhat better compared to the rest with high efficiency for both classes (0.93–0.94) as well as high values for sensitivity, specificity and balanced accuracy (0.86–0.87). The descriptor-free, SMILES-based, deep learning BERT architectures seem capable of producing well-balanced predictive models with defined applicability domains. This work also demonstrates that the class imbalance problem is gracefully handled through the use of Mondrian conformal prediction without the use of over- and/or under-sampling, weighting of classes or cost-sensitive methods.

**Keywords:** CATMoS dataset; CDDD; BERT; conformal prediction; random forest; RDKit

check for updates

**Citation:** Norinder, U. Traditional Machine and Deep Learning for Predicting Toxicity Endpoints. *Molecules* **2023**, *28*, 217. https://doi.org/10.3390/molecules28010217

Academic Editor: Igor F. Tsigelny

Received: 24 November 2022
Revised: 16 December 2022
Accepted: 21 December 2022
Published: 26 December 2022

## 1. Introduction

Drug discovery is an expensive and resource-intensive process that involves many challenges, not least within the area of toxicity [1]. A growing concern in drug development is the high attrition rate in late clinic trials due to issues related to toxicity [2]. Computational methods, e.g., computer-aided drug design, have thus become standard tools in order to improve the efficiency of the drug discovery process and to mitigate undesirable toxicity effects [3–6].

Methods such as molecular structure property modeling have been used for decades in the pharmaceutical field in order to predict important properties, e.g., biological activity, solubility and toxicity, for prioritization of compounds with respect to potential toxicity issues and experimental testing [7]. For a recent review on computational tools, see reference [8]. The increasing focus on identifying undesirable toxic effects for chemical structures of interest, real or virtual, is manifested by recent publications such as [9,10] and recent reviews on machine learning (ML) techniques by Dara and co-workers [11] and by Matsuzaka and Yashiro [12].

During the last few years, deep learning (DL) techniques have successfully been applied in various domains, e.g., natural language processing [13] as well as image analysis [14], and have increased the interest for applying such methodologies also for molecular property predictions [15].

Many of the DL investigations have used the Simplified Molecular Input Line Entry System (SMILES) notation [16] as a starting point for feature generation and have shown impressive performances [17–20]. Several different architectures have been investigated such as autoencoders [18], convolutional neural networks [21], recurrent neural networks [22] and transformers [23–25]. Lately, publications using different implementations of Bidirectional Encoder Representations from Transformers (BERT) have been published [26,27].

The Mistra SafeChem research program, financed by Mistra (The Swedish Foundation for Strategic Environmental Research, Stockholm, Sweden), has the overarching aims to create a sustainable chemical industry and reduce exposure to hazardous substances [28]. One activity, among many other objectives and activities, is to develop and use *in silico* predictive models for early prediction and verification of hazardous properties of new molecules or materials. This includes the development of molecular property models for predicting toxicity, e.g., acute toxicity.

The well-known CATMoS benchmark dataset [29] has been used in this study for modeling acute toxicity and the results from both traditional machine learning as well as deep leaning are presented.

## 2. Materials and Methods

### 2.1. CATMoS Datasets

The dataset can be downloaded from reference [29].

The originally defined training and evaluation set in [29] were used.

The training, validation and conformal prediction (CP) calibration sets were randomly selected from the original training set.

Two different binary classification sets were investigated:

The very toxic (VT; catmos_vt) and the non-toxic (NT; catmos_nt).

Table 1 shows the number of compounds in each set after standardization.

**Table 1.** Number of compounds in each of the sets.

| Dataset | Training Set | DL Validation Set | Evaluation Set | CP Calibration Set [a] |
|---|---|---|---|---|
| catmos_nt | 6004 | 662 | 2776 | 1670 |
| catmos_vt | 6449 | 717 | 2985 | 1789 |

[a]: Conformal prediction calibration set.

### 2.2. Feature Generation

#### 2.2.1. Structure Standardization

The structures, represented as SMILES, were standardized using the “remove_salt_stereo” and “organic_filter” functions of the “preprocessing.py” script found in the Continuous and Data-Driven Descriptors (CDDD) GitHub repository [30], neutralized, followed by RDKiT smiles tautomer standardization [31]. Additional structures were excluded due to MolBERT sequence length errors.

#### 2.2.2. RDKit Descriptors

A total of 96 different physiochemical descriptors were calculated using MolecularDescriptorCalculator in RDKit [31] (list of calculated descriptors is available in Supplementary Materials S2).

#### 2.2.3. CDDD Descriptors

In total, 512 CDDD descriptors of length 512 were calculated using the CDDD GitHub repository code and model [30]. The RNN architecture-based translation model translating a SMILES string into its canonical SMILES was used.

### *2.3. Traditional Machine Learning*

The binary classification models using RKDit and CDDD descriptors, respectively, were built using the Scikitlearn RandomForestClassifier [32] with default options (as part of the conformal prediction model generation, see Section 2.5 for details) [31].

### *2.4. Deep Learning*

Two different approaches were used for generating BERT models.

#### 2.4.1. MolBERT

The MolBERT binary classification models were fine-tuned using the MolBERT GitHub repository code [33] and the pre-trained model available at reference [34]. The fine-tuning network consisted of a single linear layer connected to the pooled transformer output. Default settings were used for fine-tuning. A validation set was used to avoid over-fitting of the model as well as check-pointing for saving the "best" model according to the validation set. Ten models were built using different initialization seeds.

A second set of models where fine-tuned using a pre-trained model on 500 k randomly selected PubChem compounds with a validation set of 50 k compounds.

#### 2.4.2. Molecular-Graph-BERT

The Molecular-graph-BERT binary classification models were fine-tuned using the Molecular-graph-BERT GitHub repository code [35] and a Molecular-graph-BERT pre-trained 500 k PubChem model. The fine-tuning network consisted of a two-layer fully connected neural network attached to the transformer encoder layer output. Default settings were used for fine-tuning. A validation set was used to avoid over-fitting of the model as well as check-pointing for saving the "best" model according to the validation set. Ten models were built using different initialization seeds.

### *2.5. Conformal Prediction*

A conformal predictor (CP) is a member of a family called confidence predictors [36]. These predictors have several useful properties for prediction tasks in biomedical research [37]. A particularly useful property of conformal prediction is that the method will result in valid predictions based on a user-defined significance level, i.e., a level of acceptable percentage of errors, given that the data is exchangeable. This property of validity is based on a mathematical proof by Volk and co-workers [36]. In this investigation, we used Mondrian (inductive) conformal prediction that guarantees validity for each of the two classes independently and finally median aggregated the conformal prediction outcomes from the 10 developed models for each compound and each class for final class assignment [38].

The nonconformist package [39], where scikit-learn algorithms such as the RandomForestClassifier serve as a base classifier, was employed that provide the results from conformal prediction. The following expression was used in the ICP function (Condition = lambda x: x [1]) in order to enable Mondrian conformal prediction in the nonconformist package.

An in-house script was employed to perform the conversion of predictions (scoring) from the models using BERT algorithms into results from conformal prediction. This conversion involves a calibration of the output for each compound and each label (class) in the evaluation set in relation to the output predictions for the compounds of the corresponding class in the calibration set.

A CP binary classification problem can have four possible outcomes. A test (evaluation) compound can be assigned a label for either of the two classes, assigned both labels (both classification) or none of the labels (empty classification). For a detailed description on how this calibration is performed, see Norinder and co-workers [40].

A flow chart overview of the employed machine learning approaches and calibration is depicted in Figure 1.

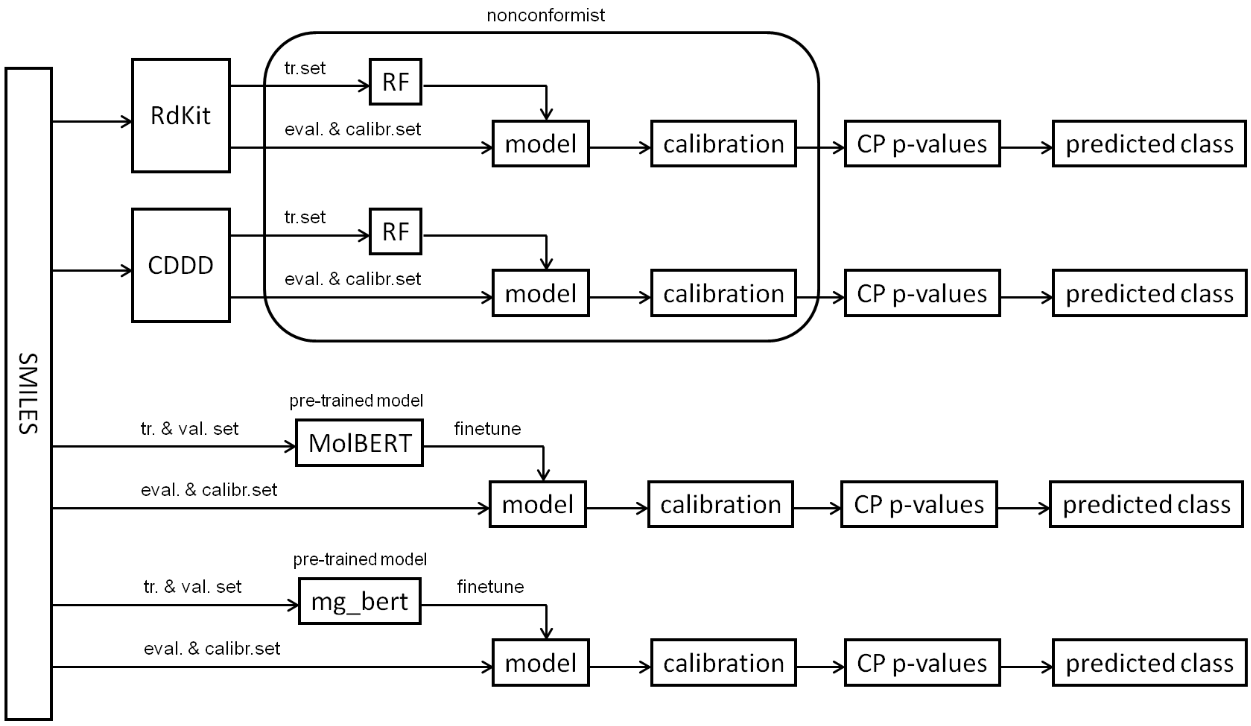

**Figure 1.** A flow chart overview depiction of the employed machine learning approaches. RdKit and CDDD = RdKit and CDDD descriptor calculation, tr. and val. set. = training and validation set, respectively, eval. and calibr. set. = evaluation and CP calibration set, respectively.

Validity and efficiency are two key measures in conformal prediction. Validity, for each class, is the percentage of correct predictions in Mondrian conformal prediction at a given significance level where the prediction contains the correct class. Thus, in binary classification the both classification is always correct (contains both available labels) while the empty classification is always erroneous (contains no labels). Models were considered valid when the resulting error rate does not exceed the set error rate (significance level) by more the 2.5%.

Efficiency, for each class, is defined as the percentage of single label predictions (only one label), regardless of whether the prediction is correct or not, at a given significance level. High efficiency is therefore desirable since a larger number of predictions are single label predictions and more informative as they contain only one class.

## 3. Results and Discussion

The aim of this study is to investigate how different molecular representations and algorithmic approaches may affect the predictive performance of the derived models. The present study therefore compares results from the traditionally used Random Forest/physico-chemical descriptor approach through an intermediate Random Forest/auto encoder representation to deep learning BERT/molecular-graph-based approaches.

The results from the study are shown in Table 2 and Supplementary Materials Table S1 and depicted in Figures 2–5.

**Table 2.** Aggregated conformal prediction results at significance level 0.2.

| Dataset | Method [a] | Significance Level [b] | Validity Minority Class 1 | Validity Majority Class 0 | Efficiency Minority Class 1 | Efficiency Majority Class 0 | Sensitivity (SE) | Specificity (SP) | Balanced Accuracy (BA) |
|---|---|---|---|---|---|---|---|---|---|
| catmos_nt | cddd | 0.2 | 0.802 | 0.824 | 0.855 | 0.879 | 0.769 | 0.800 | 0.785 |
| catmos_nt | mg_bert | 0.2 | 0.798 | 0.848 | 0.814 | 0.845 | 0.751 | 0.821 | 0.786 |
| catmos_nt | molbert_p | 0.2 | 0.791 | 0.830 | 0.856 | 0.864 | 0.756 | 0.803 | 0.779 |
| catmos_nt | molbert | 0.2 | 0.797 | 0.830 | 0.873 | 0.892 | 0.767 | 0.810 | 0.788 |
| catmos_nt | rdkit | 0.2 | 0.800 | 0.805 | 0.909 | 0.912 | 0.780 | 0.786 | 0.783 |
| catmos_vt | cddd | 0.2 | 0.770 | 0.817 | | | | | |
| catmos_vt | mg_bert | 0.2 | 0.843 | 0.829 | 0.923 | 0.900 | 0.830 | 0.810 | 0.820 |
| catmos_vt | molbert_p | 0.2 | 0.798 | 0.819 | 0.996 | 0.991 | 0.798 | 0.820 | 0.809 |
| catmos_vt | molbert | 0.2 | 0.815 | 0.818 | 0.944 | 0.931 | 0.863 | 0.878 | 0.871 |
| catmos_vt | rdkit | 0.2 | 0.819 | 0.821 | 0.996 | 0.979 | 0.822 | 0.839 | 0.830 |

[a]: cddd = RF/cddd 10 models, mg_bert = Molecular-graph-BERT/smiles 10 models, molbert = MolBERT/smiles 10 models, molbert_p = MolBERT/smiles 10 models with PubChem pre-trained model, rdkit = RF/rdkit 10 models, [b] = CP significance level.

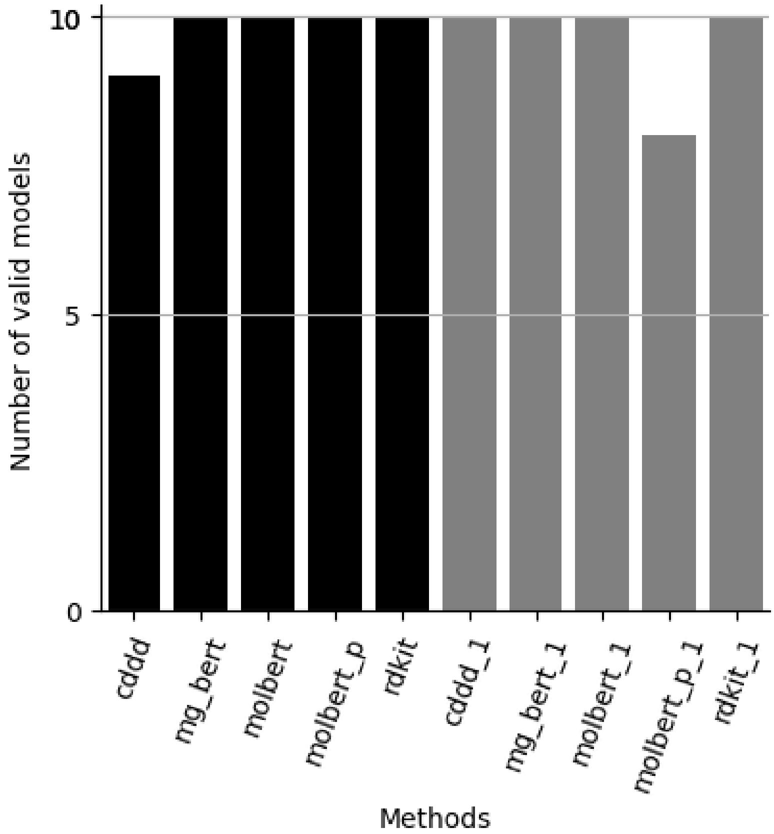

**Figure 2.** Number of valid evaluation set models (maximum 10) for each method type. Methods: cddd = RF/cddd 10 models, mg_bert = Molecular-graph-BERT/smiles 10 models, molbert = MolBERT/smiles 10 models, molbert_p = MolBERT/smiles 10 models with PubChem pre-trained model, rdkit = RF/rdkit 10 models, xxx_1 is the corresponding approach based on only 1 model.

From Figure 2, it can be noted that most methods, both ensemble and single models, produce valid models for significance level 0.1–0.3 with the exception of one CDDD ensemble model and two single MolBERT models based on the PubChem pre-trained model.

Figure 3 shows that both RDKit approaches and the BERT ensemble approaches as well as many single models result in valid models for both datasets at significance levels of primary interest, e.g., with error levels (significance levels) set at 10, 15 and 20%.

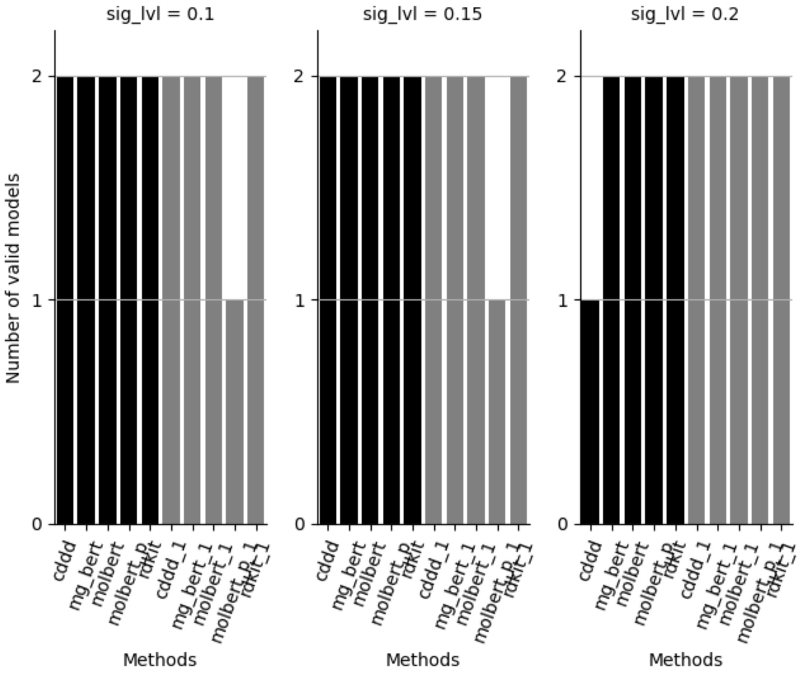

**Figure 3.** Number of valid evaluation set models, at significance levels 0.1, 0.15 and 0.2, for each method (maximum 2). Methods: cddd = RF/cddd 10 models, mg_bert = Molecular-graph-BERT/smiles 10 models, molbert = MolBERT/smiles 10 models, molbert_p = MolBERT/smiles 10 models with PubChem pre-trained model, rdkit = RF/rdkit 10 models, xxx_1 is the corresponding approach based on only 1 model.

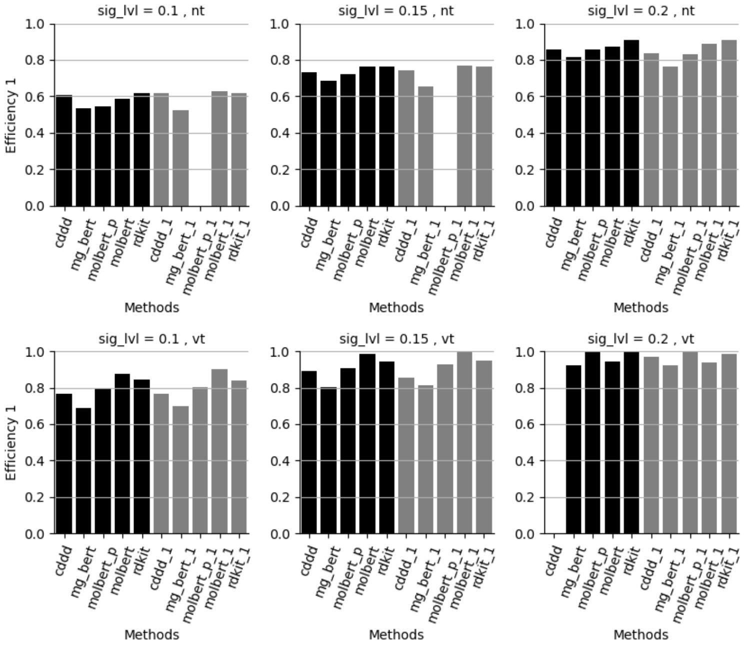

**Figure 4.** Evaluation set efficiency for class "1" for the 2 datasets (NT model upper row, VT model lower row), at significance levels 0.1–0.2, for each method. Class "1": non-toxic class and very toxic class for the 2 datasets nt and vt, respectively. Methods: cddd = RF/cddd 10 models, mg_bert = Molecular-graph-BERT/smiles 10 models, molbert = MolBERT/smiles 10 models, molbert_p = MolBERT/smiles 10 models with PubChem pre-trained model, rdkit = RF/rdkit 10 models, xxx_1 is the corresponding approach based on only 1 model.

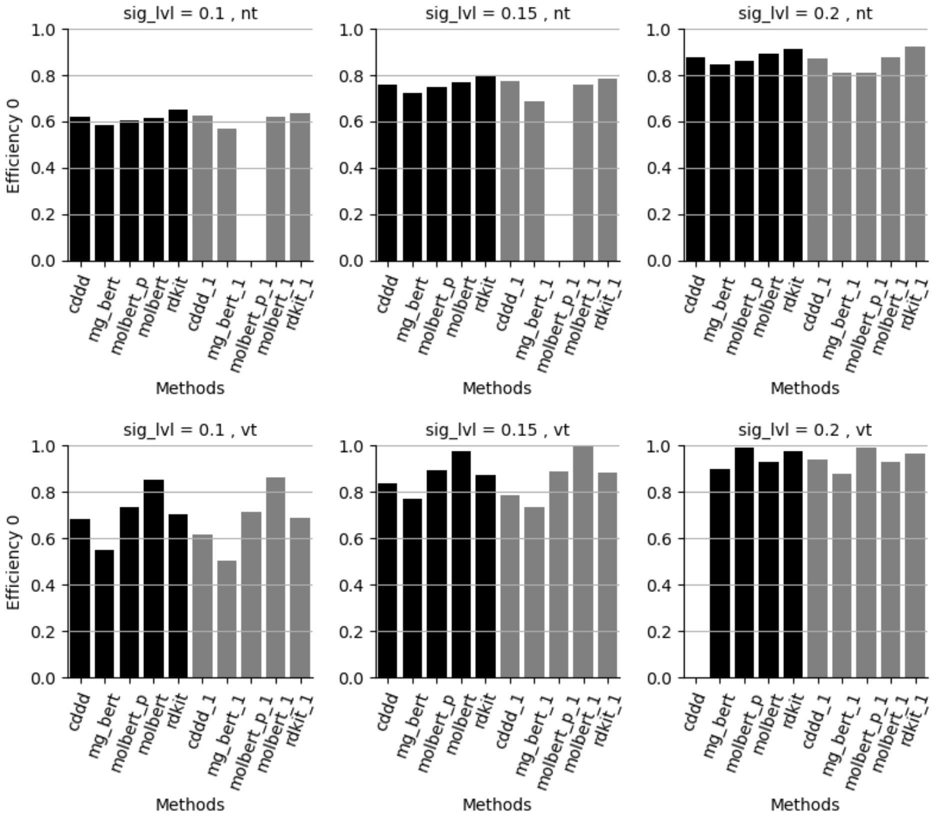

**Figure 5.** Evaluation set efficiency for class "0" for the 2 datasets (NT model upper row, VT model lower row), at significance levels 0.1–0.2, for each method. Class "0": the other binary class for each dataset as compared to Figure 4. Methods: cddd = RF/cddd 10 models, mg_bert = Molecular-graph-BERT/smiles 10 models, molbert = MolBERT/smiles 10 models, molbert_p = MolBERT/smiles 10 models with PubChem pre-trained model, rdkit = RF/rdkit 10 models, xxx_1 is the corresponding approach based on only 1 model.

Figures 4 and 5 show that the efficiencies, i.e., the percentage of single label predictions, for both class 1 and 0 at an acceptable error level of 10% are, on average, only 60–70%, which, for most cases, cannot be considered sufficient. At 15% acceptable error rate both the MolBERT and RDKit approaches show efficiencies for both class 1 and 0 close to or above 80% which ensures a large portion of single-label predictions. The efficiency is further increased at 20% acceptable error rate where all approaches have efficiencies well above 80% for both class 1 and 0. The performance is generally somewhat better for the catmos_vt dataset compared to the catmos_nt dataset.

The class distribution (class "0"/class "1") is ~1.6:1 for catmos_nt and ~11:1 for catmos_vt which means that the former data set is rather balanced while the latter is rather unbalanced. This imbalance may, in turn, cause some issues for ML algorithms to properly handle the minority class [41–43].

From the catmos_nt results presented in Table 2 it can be noted that all of the developed models in this study performs similarly with respect to the SE/SP balance with an absolute average difference of 0.039.

The catmos_vt results presented in Table 2 show the graceful handling of the class imbalance in this dataset, by an absolute average difference between SP and SE for the four models (cddd not valid for class 1) in this investigation of 0.019, by using Mondrian conformal prediction. Furthermore, this balanced performance was the result of running the Mondrian CP framework in default mode and without the use of over- and/or under-sampling, weighting of classes or cost-sensitive measures.

The more balanced results from this investigation with respect to SE and SP are due to the independent handling of the two classes as part of the Mondrian conformal prediction calibration procedure; see Section 2.5 for more details.

The well-balanced SE/SP performance is of importance from the point of safety, i.e., not to err on the false negative side, when predicting a toxic compound to be non-toxic for the catmos_vt model. It is less of a problem for the catmos_nt model if a few more non-toxic compounds are predicted to be toxic from a safety perspective.

All methods in Table 2 are performing equally well on the balanced catmos_nt dataset while MolBERT, based on the larger pre-trained model (method "molbert" in Table 2), seems to be performing somewhat better than the other methods for the catmos_vt dataset with respect to SE and SP and the balance between them based on BA results from the 10 individual models (95% confidence, Mann–Whitney U test with Bonferroni correction for multiple testing) constituting the molbert ensemble results.

Advantages of the BERT approaches over the RDKit descriptor-based approach is that they are descriptor-free in that SMILES are used as input without the need for explicit descriptor generation prior to modeling and that the results from the models can be projected back onto the atoms of the molecules through their attention mechanisms. The advantage of the RDKit and CDDD approaches is shorter computation costs and that smaller datasets can be modeled with acceptable outcomes compared to using deep learning in the form of BERT that usually requires larger training sets.

## 4. Conclusions

The results show for the binary CATMoS non-toxic, almost equally balanced dataset that all methods perform equally well. The results also show that the MolBERT model based on a larger pre-trained network performs somewhat better compared to the rest for the binary CATMoS very-toxic dataset with an 11-fold difference between the two classes. The descriptor-free, SMILES based, deep learning BERT architectures seem capable of producing well-balanced predictive models with defined applicability domains. This work also demonstrates that the class imbalance problem is gracefully handled through the use of Mondrian conformal prediction without the use of over- and/or under-sampling, weighting of classes or cost-sensitive methods.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules28010217/s1, Table S1: Conformal predictions results at significance levels 0.1–0.3.; list of calculated RDKit descriptors Materials S2: list of calculated RDKit descriptors.

**Funding:** The author thank the Swedish Foundation for Strategic Environmental Research (Mistra: project Mistra SafeChem, project number 2018/11) for financial support.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used in this study are publicly available in reference [29].

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33. [CrossRef] [PubMed]
2. Hwang, T.J.; Carpenter, D.; Lauffenburger, J.; Wang, B.; Franklin, J.M.; Kesselheim, A. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern. Med.* **2016**, *176*, 1826–1833. [CrossRef] [PubMed]
3. Schaduangrat, N.; Lampa, S.; Simeon, S.; Gleeson, M.P.; Spjuth, O.; Nantasenamat, C. Towards reproducible computational drug discovery. *J. Cheminform.* **2020**, *12*, 9. [CrossRef]
4. Sabe, V.T.; Ntombela, T.; Jhamba, L.A.; Maguire, G.E.; Govender, T.; Naicker, T.; Kruger, H.G. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **2021**, *224*, 113705. [CrossRef] [PubMed]

5. Lin, X.; Li, X.; Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **2020**, *25*, 1375. [CrossRef]
6. Brogi, S.; Ramalho, T.C.; Kuca, K.; Medina-Franco, J.L.; Valko, M. Editorial: In silico Methods for Drug Design and Discovery. *Front. Chem.* **2020**, *8*, 612. [CrossRef] [PubMed]
7. Muratov, E.N.; Bajorath, J.; Sheridan, R.P.; Tetko, I.V.; Filimonov, D.; Poroikov, V.; Oprea, T.I.; Baskin, I.I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564. [CrossRef]
8. Cox, P.B.; Gupta, R. Contemporary Computational Applications and Tools in Drug Discovery. *ACS Med. Chem. Lett.* **2022**, *13*, 1016–1029. [CrossRef]
9. Idakwo, G.; Luttrell, J.; Chen, M.; Hong, H.; Zhou, Z.; Gong, P.; Zhang, C. A review on machine learning methods for in silico toxicity prediction. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **2018**, *36*, 169–191. [CrossRef]
10. Pérez, S.E.; Rodríguez, S.R.; González, G.M.; Del Mar García Suárez, M.; Díaz, G.D.B.; Cabal, M.D.C.; Rojas, J.M.M.; López Sánchez, J.I. Toxicity prediction based on artificial intelligence: A multidisciplinary overview. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1516.
11. Dara, S.; Dhamercherla, S.; Jadav, S.S.; Babu, C.M.; Ahsan, M.J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **2021**, *55*, 1947–1999. [CrossRef] [PubMed]
12. Matsuzaka, Y.; Yashiro, R. Applications of Deep Learning for Drug Discovery Systems with BigData. *Biomedinformatics* **2022**, *2*, 603–624. [CrossRef]
13. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
14. Yang, B.; Xu, S.; Chen, H.; Zheng, W.; Liu, C. Reconstruct Dynamic Soft-Tissue With Stereo Endoscope Based on a Single-Layer Network. *IEEE Trans. Image Process.* **2022**, *31*, 5828–5840. [CrossRef] [PubMed]
15. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [CrossRef] [PubMed]
16. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [CrossRef]
17. Jastrzebski, S.; Lesniak, D.; Czarnecki, W.M. Learning to SMILE(S). *arXiv* **2016**, arXiv:1602.06289.
18. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2018**, *10*, 1692–1701. [CrossRef]
19. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. [CrossRef]
20. Goh, G.B.; Siegel, C.M.; Vishnu, A.; Hodas, N.O. Using Rule-Based Models for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19 August 2018; pp. 302–310.
21. Goh, G.B.; Siegel, C.M.; Vishnu, A.; Hodas, N.O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. *arXiv* **2017**, arXiv:1706.06689.
22. Li, X.; Fourches, D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, *12*, 27. [CrossRef] [PubMed]
23. Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzebski, S. Molecule Attention Transformer. *arXiv* **2020**, arXiv:2002.08264.
24. Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, New York, NY, USA, 7 September 2019; pp. 429–436.
25. Maziarka, Ł.; Majchrowski, D.; Danel, T.; Gaiński, P.; Tabor, J.; Podolak, I.; Morkisz, P.; Jastrzębski, S. Relative Molecule Self-Attention Transformer. *arXiv* **2021**, arXiv:2110.05841.
26. Zhang, X.-C.; Wu, C.-K.; Yang, Z.-J.; Wu, Z.-X.; Yi, J.-C.; Hsieh, C.-Y.; Hou, T.-J.; Cao, D.-S. MG-BERT: Leveraging unsupervised atomic representation learning for molecular property prediction. *Brief. Bioinform.* **2021**, *22*, bbab152. [CrossRef] [PubMed]
27. Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M.H.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv* **2020**, arXiv:2011.13230.
28. Mistra SafeChem. Available online: https://www.ivl.se/projektwebbar/mistra-safechem.html (accessed on 26 October 2022).
29. Mansouri, K.; Karmaus, A.L.; Fitzpatrick, J.; Patlewicz, G.; Pradeep, P.; Alberga, D.; Alepee, N.; Allen, T.E.H.; Allen, D.; Alves, V.M.; et al. CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environ. Health Perspect.* **2021**, *129*, 47013. [CrossRef]
30. Continuous and Data-Driven Descriptors (CDDD). Available online: https://github.com/jrwnter/cddd (accessed on 11 August 2019).
31. RDKit: Open-Source Cheminformatics. version 2020.09.1.0. Available online: https://www.rdkit.org (accessed on 28 January 2021).
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830. Available online: https://scikit-learn.org (accessed on 28 January 2021).
33. MolBERT. Available online: https://github.com/BenevolentAI/MolBERT (accessed on 21 August 2022).
34. MolBERT Pre-Trained Model. Available online: https://ndownloader.figshare.com/files/25611290 (accessed on 21 August 2022).

35. Molecular-Graph-BERT. Available online: https://github.com/zhang-xuan1314/Molecular-graph-BERT (accessed on 21 August 2022).
36. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005; pp. 1–324.
37. Cortés-Ciriano, I.; Bender, A. Concepts and applications of conformal prediction in computational drug discovery. In *Artificial Intelligence in Drug Discovery*; Nathan, B., Ed.; The Royal Society of Chemistry: Cambridge, UK, 2021; pp. 63–101.
38. Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. In *Artificial Intelligence Applications and Innovations. AIAI 2014. IFIP Advances in Information and Communication Technology*; Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S., Makris, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 437, pp. 231–240.
39. Nonconformist. Available online: https://github.com/donlnz/nonconformist (accessed on 28 January 2021).
40. Norinder, U.; Myatt, G.; Ahlberg, E. Predicting Aromatic Amine Mutagenicity with Confidence: A Case Study Using Conformal Prediction. *Biomolecules* **2018**, *8*, 85. [CrossRef]
41. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
42. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
43. Korkmaz, S. Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 4180–4190. [CrossRef] [PubMed]

**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.