Appendix

A1 Chemical language transformation pipeline

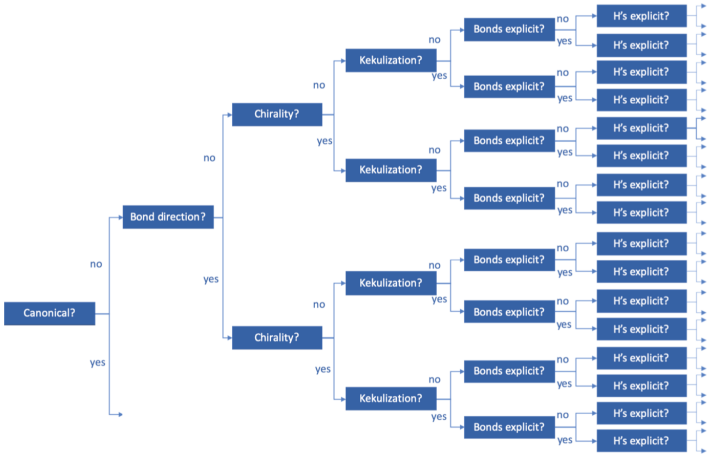


Fig. A1: **Order of transformation steps in SMILES preprocessing pipeline.** Upon the user's specification of boolean variables for each of the listed, optional SMILES transformations, the shown pipeline is executed. In a first step, the user decides whether the input SMILES should be converted into canonical SMILES or not. In the next steps, the user can specify whether the bond direction ((E) or (Z), or rather the removal of back- and forward slashes in the string) or the chirality information should be present. Then, if stated, the transformation of the aromatic moiety into a kekulized representation takes place. At last, the algorithm explicitly adds bond information and hydrogen atoms to the SMILES string, if defined by the user. Note that the same pipeline can be applied to SELFIES [39], in that case the SMILES is converted to SELFIES after the shown pipeline.

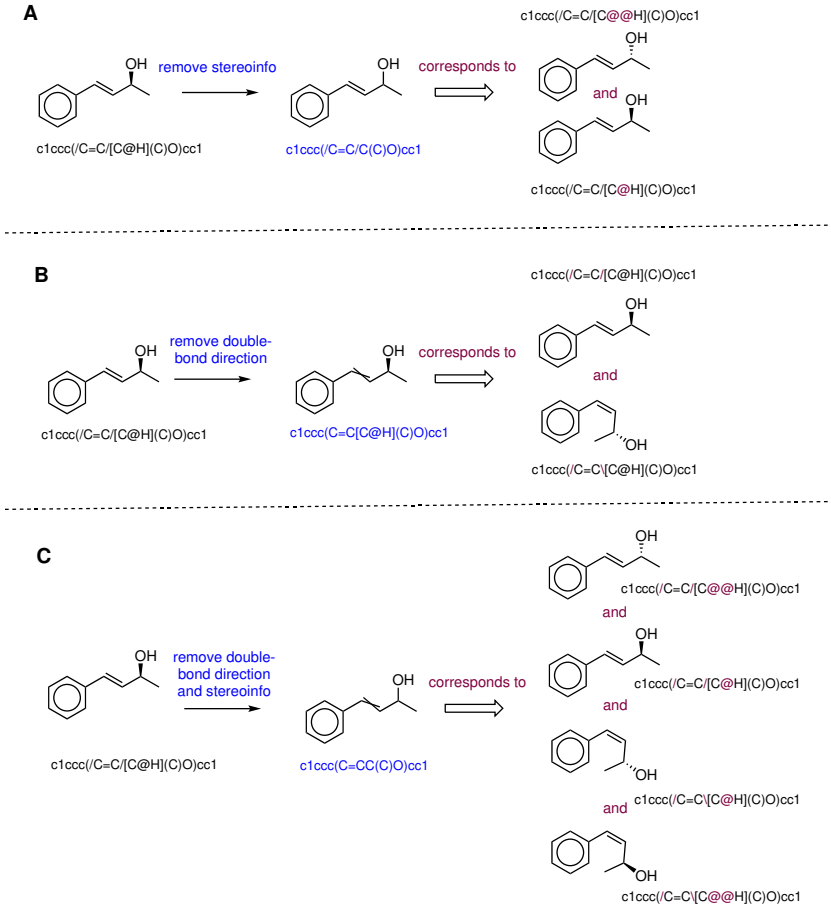


Fig. A2: **When stereoinformation is removed from the original SMILES string (*left*), then it represents multiple molecules (*right*).** This figure shows the different ways to remove information of (*S*,*E*)-4-phenylbut-3-en-2-ol. The SMILES strings are also represented by the corresponding molecules. **A:** Stereoinformation on a tetrahedral stereocenter is represented in the SMILES string by `@` and `@@`. Here, `@` lists the neighbors clockwise, whereas `@@` lists the neighbors anticlockwise. The SMILES string with the removed information corresponds to two molecules. **B:** Stereoinformation on a double bond is represented by /.../ or \...\(both corresponding to a (*Z*) double bond) and /...\or \.../ (both corresponding to an (*E*) double bond) in the SMILES string. Again, the SMILES string with the removed information corresponds to two molecules. **C:** When both, the stereoinformation on the double bond and on the stereocenter, are removed, four molecules are possible from that SMILES representation.

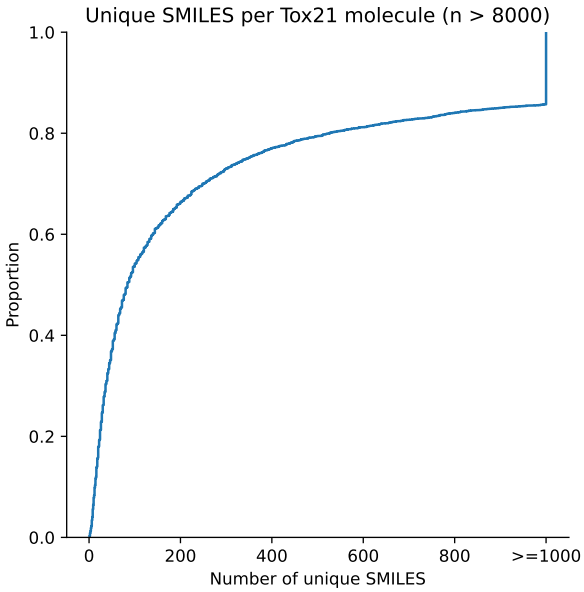


Fig. A3: **Number of unique SMILES that can be obtained per molecule from the Tox21 dataset.** For each molecule, we measured the number of unique SMILES sequences obtained through augmentation. We made two assumptions. First, if no new SMILES sequence was generated for 1000 augmentations, we assumed that all augmentations had been generated. Secondly, when $10,000$ unique SMILES were generated, we proceeded with the next molecule (to save runtime).

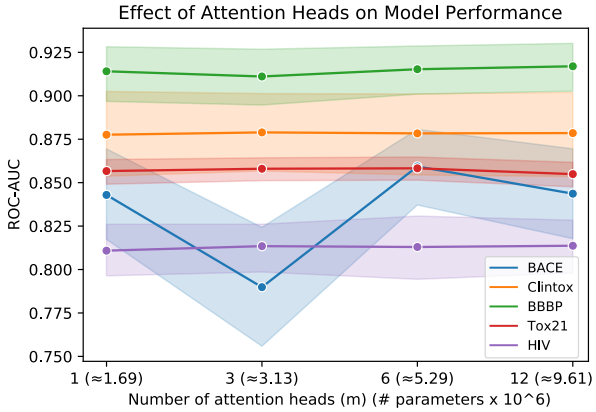
Table A1: **Differences between the original Tox21 dataset [90] and the MoleculeNet distribution [11]**

	Official Dataset	**DeepChem Dataset**
total number of entries	12707	8014
number of unique compounds	8982 (70.7%)	8011 (99.97%)
number of canonical SMILES	1831 (14.4%)	7831 (97.7%)
number of compounds with aromatic rings	8582 (67.5%)	2581 (32.3%)
number of kekulized compounds	3269 (25.7%)	0 (0.0%)
number of compounds with directed double bonds	837 (6.6%)	466 (5.8%)
number of compound with stereocenters	2247 (17.7%)	1322 (16.5%)
number of compounds with directed double bonds and stereocenters	159 (1.3%)	106 (1.3%)
number of compounds with countermolecules/-ions	2557 (20.1%)	246 (3.1%)
number of compounds with counterions	1082 (8.5%)	141 (1.8%)
number of compounds with molecule twice	0 (0.0%)	56 (0.7%)

A1.1 Effect of number of attention heads on model performance

Fig. A4: **Ablation study on the number of attention heads across MoleculeNet datasets.** In this experiment, we vary the number of attention heads (m) of our model and evaluate its performance on all datasets from MoleculeNet - BACE (1 task), Clintox (2 tasks), SIDER (27 tasks), BBBP (1 task), Tox21 (12 tasks) and HIV (1 task). We consider $m \in \{1, 3, 6, 12\}$. Each experiment is repeated 10 times.

(a) Mean ROC-AUC and its 95% CI. The performance varies little, except for BACE which shows the best performance for $m = 6$ heads. SIDER is excluded since the model performance was constant($= 0.66$).



(b) ROC-AUC is averaged across all MoleculeNet datasets for each head size.

m	**Size**	**AUC**
1	1.7M	$0.827_{\pm 0.1}$
3	3.1M	$0.819_{\pm 0.1}$
6	**5.3M**	$\mathbf{0.831}_{\pm 0.1}$
12	9.6M	$0.828_{\pm 0.1}$

A1.2 Tox21 Results

	ROC-AUC	
Embedding	**Augment: ✗**	**Augment: ✓**
one-hot	0.842 ± 0.004	**0.855** ± 0.003
learned	0.832 ± 0.005	0.851 ± 0.003
pretrained (fixed)	0.845 ± 0.005	0.853 ± 0.002
pretrained (flexible)	**0.847** ± 0.004	0.853 ± 0.003

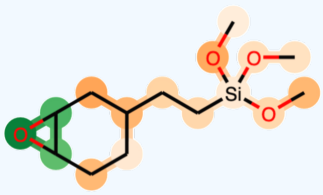
Table A2: **Ablation study on different SMILES embeddings on the Tox21 dataset [90].** For the results in the main manuscript, the learned embeddings were used. The pretrained embeddings were taken from **(author?)** [46] who trained a VAE on ChEMBL data [124]. In the *flexible* configuration the embeddings were finetuned on the Tox21 dataset. Without augmentation, the advantage of pretrained embeddings was significant, especially compared to learned embeddings.

Dataset # of tasks	BACE 1	BBBP 1	Tox21 12	Clintox 2	SIDER 27	**Average**
Ours	$0.861_{\pm 0.039}$	$\underline{0.915}_{\pm 0.023}$	$0.795_{\pm 0.050}$	$0.896_{\pm 0.006}$	$0.619_{\pm 0.037}$	$0.817_{\pm 0.031}$
TF_Robust	$0.824_{\pm 0.022}$	$0.860_{\pm 0.087}$	$0.698_{\pm 0.012}$	$0.765_{\pm 0.085}$	$0.607_{\pm 0.033}$	$0.751_{\pm 0.048}$
GraphConv	$0.854_{\pm 0.011}$	$0.877_{\pm 0.036}$	$0.772_{\pm 0.041}$	$0.845_{\pm 0.051}$	$0.593_{\pm 0.035}$	$0.788_{\pm 0.03}$
Weave	$0.791_{\pm 0.008}$	$0.837_{\pm 0.065}$	$0.741_{\pm 0.044}$	$0.823_{\pm 0.023}$	$0.543_{\pm 0.034}$	$0.747_{\pm 0.035}$
SchNet	$0.750_{\pm 0.033}$	$0.847_{\pm 0.024}$	$0.767_{\pm 0.025}$	$0.717_{\pm 0.042}$	$0.545_{\pm 0.038}$	$0.725_{\pm 0.032}$
MPNN	$0.815_{\pm 0.044}$	$0.913_{\pm 0.041}$	$0.808_{\pm 0.024}$	$0.879_{\pm 0.054}$	$0.595_{\pm 0.030}$	$0.802_{\pm 0.04}$
MGCN	$0.734_{\pm 0.030}$	$0.850_{\pm 0.064}$	$0.707_{\pm 0.016}$	$0.634_{\pm 0.042}$	$0.552_{\pm 0.018}$	$0.695_{\pm 0.034}$
AttentiveFP	$0.863_{\pm 0.015}$	$0.908_{\pm 0.050}$	$0.807_{\pm 0.020}$	$\mathbf{0.933}_{\pm 0.020}$	$0.605_{\pm 0.060}$	$\underline{0.823}_{\pm 0.033}$
N-GRAM	$\underline{0.876}_{\pm 0.035}$	$0.912_{\pm 0.013}$	$0.769_{\pm 0.027}$	$0.855_{\pm 0.037}$	$\underline{0.632}_{\pm 0.005}$	$0.808_{\pm 0.023}$
[107]	$0.851_{\pm 0.027}$	$\underline{0.915}_{\pm 0.040}$	$\underline{0.811}_{\pm 0.015}$	$0.762_{\pm 0.058}$	$0.614_{\pm 0.006}$	$0.791_{\pm 0.029}$
GROVER	$\mathbf{0.878}_{\pm 0.016}$	$\mathbf{0.936}_{\pm 0.008}$	$\mathbf{0.819}_{\pm 0.020}$	$\underline{0.925}_{\pm 0.013}$	$\mathbf{0.656}_{\pm 0.006}$	$\mathbf{0.843}_{\pm 0.013}$

Table A3: **ROC-AUC values for different algorithms evaluated on MoleculeNet datasets split using a *scaffold* splitting strategy.** For each dataset the average ROC-AUC across the tasks is reported. Results for our model were obtained by measuring test performance for 10 repeated scaffold splits. All other numbers are taken from **(author?)** [13] who trained all models on 3 repeated scaffold splits.

A1.3 Case Study on Toxicophoric Substructure

Trimethoxy(2-7-oxabicyclo[4.1.0]heptan-3-ylethyl)silane



The molecule trimethoxy(2-7-oxabicyclo[4.1.0]heptan-3-ylethyl)silane has a toxicophoric substructure (green part), the epoxide. This substructure is of dark color, meaning that the attention weight from the Tox21 model is high.
Epoxides are electrophilic structures that can be attacked by a nucleophile, for example the bases of the DNA (forming an epoxide-DNA adduct [125]) or amino acids in proteins, even active sites of proteins (such as aspartic acid [126]). The nucleophilic attack leads to a reduction of the strain caused by the small ring structure.
Epoxides have environmental effects [95, 96]. They are also known to cause skin sensitization [127] and are carcinogenic, mutagenic and genotoxic [128]. Therefore, on bottles of the compound above, the following hazard statements (H phrases) need to be written[a]:

- H317: May cause an allergic skin reaction.
- H341: Suspected of causing genetic defects.
- H351: Suspected of causing cancer.
- H412: Harmful to aquatic life with long lasting effects.

That does not mean that epoxides are not used in drug discovery: despite the hazardous properties of epoxides, or rather because of them, the main therapeutic interest is as anticancer agents. Here, they act via enzyme inhibition, induction of cell cycle arrest and apoptosis. Selected molecules are used in treatments against heart failure, infections or gastrointestinal diseases [129].

[a] `https://www.sigmaaldrich.com/DE/en/sds/ALDRICH/413321`

Fig. A5: **Case study on trimethoxy(2-7-oxabicyclo[4.1.0]heptan-3-ylethyl)silane.** The highest attention weights of the molecule are focussed on the well-known toxicophoric epoxide (cf. Figure 3 B).