MolBART: Generative Masked Language Models for Molecular Representations

Anonymous authors
Paper under double-blind review

Abstract

We discover a robust self-supervised strategy tailored towards molecular representations for generative masked language models through a series of tailored, in-depth ablations. Using this pre-training strategy, we train `MolBART`, a BART-like model with an order of magnitude more compute than previous self-supervised molecular representations. In-depth evaluations show that `MolBART` consistently outperforms other self-supervised representations across classification, regression, and generation tasks setting a new state-of-the-art on 11 tasks. We then quantitatively show that when applied to the molecular domain, the BART objective learns representations that implicitly encode our downstream tasks of interest. For example, by selecting seven neurons from a frozen `MolBART`, we can obtain a model having performance within two percentage points of the full fine-tuned model on task Clintox. Lastly, we show that standard attribution interpretability methods, when applied to `MolBART`, highlight certain substructures that chemists use to explain specific properties of molecules.

1 Introduction

Recent advancements in large-scale representation learning have significantly improved downstream performance on virtually every modality, from images Aghajanyan et al. (2022); Bao et al. (2021); He et al. (2022) to text Liu et al. (2019); Lewis et al. (2020) to speech Conneau et al. (2020); Radford et al.. Meanwhile, the domain of molecular representations has only seen a fraction of the scale of these other domains. In this paper, we show that pre-training with masked language modeling objective on a large collection of molecules allows us to learn generalizable representations that perform well on a wide range of computational chemistry benchmarks.

We train `MolBART`, a general purpose, pre-trained generative masked language model for molecules trained with an order of magnitude more compute than previous pre-trained molecular representations over all available molecules (over 1.7 billion molecules).

To validate the efficacy of our representations, we do in-depth evaluations across a large amount of classification, regression, and generation tasks achieving a new state of the art on ten tasks. We then explore the representations from a domain-expert perspective by comparing a series of attribution interpretability methods with structural alerts, a rule-based system developed by chemists to explain specific molecular properties. In particular, for a model trained to predict the results of a commonly used toxicity test, the Integrated Gradients method usually highlights atoms known to be part of the substructures responsible for toxicity.

Furthermore, our pre-trained model, in a completely unsupervised fashion, learned representations for specific tasks, i.e., certain neurons correlated highly with downstream tasks. For specific tasks, we see that a linear combination of frozen ten neurons is enough to reach 90% of the full fine-tuning performance.

To summarize our contributions are the following:

- A robust pre-training strategy tailored towards molecular representations discovered through a series of in-depth ablations for generative masked language models.
- We present `MolBART` pre-trained model for molecules trained with an order of magnitude more compute than previous pre-trained molecular representations.

- We disentangle the differences between different data splitting during evaluation used by previous works and provide concise and consistent evaluations on a large set of tasks across previous models and `MolBART`.

- We present a quantitative analysis of molecular properties learned during the pre-training process by analyzing individual neurons in the representations that correlate highly with downstream tasks.

- A qualitative analysis using attribution interpretability methods finds that the model internally learns patterns known to in-domain experts to correlate with certain chemical properties.

2 RELATED WORK

Chemical Property Prediction: Machine learning has been applied for the prediction of chemical properties since the 1960s (Dearden, 2016). Initially, the molecules were represented by fingerprints like Morgan fingerprints (Morgan, 1965), circular fingerprints (Glen et al., 2006) or extended-connectivity circular fingerprints (ECFP) (Rogers & Hahn, 2010) which were given as an input to machine learning models. Deep learning has been applied to fingerprints since at least 2014 (Dahl et al., 2014). Duvenaud et al. (2015) was the first to apply convolutional networks directly on the graphs by generalizing the computation of circular fingerprints. In contrast, Kearnes et al. (2016) introduced graph convolutions on molecular graphs.

Wu et al. (2018) introduced the MoleculeNet benchmark suite, which enabled direct comparisons among machine learning algorithms for several classification and regression tasks. It became the standard evaluation system for most of the subsequent work. Lowe (2012) extracted a large set of chemical reactions from US patents, becoming a standard benchmark for evaluating the generation abilities of machine learning methods.

Self-Supervised Learning for Molecular Representations: Jaeger et al. (2018) introduced Mol2vec, the adaptation of word2vec for molecules. Wang et al. (2019) was the first to apply BERT-like models on SMILES strings but unfortunately did not evaluate on MoleculeNet benchmarks. Chithrananda et al. (2020) further scaled up BERT like models for molecules showing consistent improvement with scale. MolCLR Wang et al. (2022d) and iMolCLR Wang et al. (2022c) use graph neural networks to perform contrastive learning on 10 million unlabeled molecules. Recently, Wang et al. (2022a) improved molecular representations by adding constraints on the sums of the representations of reactants.

The most similar paper to ours is ChemFormer Irwin et al. (2022), which also trains a BART-like model but, unlike our paper, does so with a fraction of the compute and scale. Additionally, our paper explores the further tweaks needed to the generative masked language model objective to make it optimal in the domain of molecular representations. We also present a unified fine-tuning recipe that removes the need for complex hyper-parameter tuning while consistently outperforming previous state-of-the-art models.

3 PRETRAINING

Dataset: We use the Zinc20 dataset from Irwin et al. (2020) for all of our pre-training ablations and experiments. We deduplicate the data based on the hashes of canonicalized SMILES using RDKit (Landrum et al., 2013), leaving us with a total of slightly north of 1.7 billion samples from which we reserve 10000 samples as our validation set.

Implementation: In all of our experiments, we parameterize our masked language model with the standard pre-layer norm Transformer architecture, precisely the BART-Large model proposed in Lewis et al. (2019). We have a maximum sequence length per sample of 128 tokens for all models. We train using the FairSeq framework Ott et al. (2019) with PyTorch Paszke et al. (2019) as the underlying framework. For our larger models, we use the fully sharded data-parallel implementation available in FairScale (Baines et al., 2021). All experiments use automatic mixed-precision available in FairSeq. We use the Aim ecosystem (Arakelyan et al., 2020) for experiment tracking.

Choice Of Objective: When choosing an objective, there are a certain number of pre-conditions to consider depending on the end goal of the model. Wang et al. (2022b) argues that for zero-shot or k-shot prompting causal language models with uni-directional attention is optimal, while bidirectionality (in both context and attention masks) is the primary driver of success in the fine-tuning setting.

Within the fine-tuning setting, there are different objectives conditioned on model type. For encoder-based models, the masked language modeling objective as initially proposed in Devlin et al. (2018) and further refined in Liu et al. (2019). The downside of encoder models is the inability to do generative fine-tuning, which led to the introduction of the denoising model for the encoder-decoder models Lewis et al. (2019). Decoder causal models are problematic because they are not bidirectional, although recently proposed objectives such as causal masking in Aghajanyan et al. (2022) are bidirectional in context but not in attention.

For `MolBART` we select both the denoising objective and architecture from Lewis et al. (2019) while previous works have focused on encoder-only (Wang et al., 2019; Chithrananda et al., 2020).

3.1 Ablation

Given the novelty of the molecular representation domain for pre-training, we aim to do an in-depth ablation analysis on how to successfully pre-train within this domain.

We measure the quality of learned representations by fine-tuning with a fixed set of hyper-parameters on three datasets, HIV from Wu et al. (2018), BBBP from Martins et al. (2012), and ClinTox from Wu et al. (2018), reporting the average. For the exact fixed hyper-parameters please refer to § E. We provide an overarching summary of our ablation in Figure 1.

Choice of Tokenization: Previous work, such as Wang et al. (2019) uses pre-defined tokenization rules on SMILES usually separating individual logical elements such as atoms (i.e., C, N, H) and chemical bond symbols like (i.e., #, -, =). These fundamental units for molecules are analogous to using characters in natural language. Recent advancements in natural language pre-training have ubiquitously used sub-word tokenization as a compromise between the full character coverage of character level tokenization and the need for large vocabularies for word-based tokenization. We train a uni-gram tokenizer on a random sample of 10 Million samples from our training set with a 1021 vocabulary size using SentencePiece (Kudo & Richardson, 2018).

Figure 1: A summary of decision made when pre-training `MolBART`, as well as their absolute gain on our ablation benchmark.

To ablate the benefit of learned tokenization vs. hand-crafted tokenization, we train two equivalent models (§3), using the `MolBART`-Base architecture. We fix the amount of compute for each run to the amount needed for one epoch, using learned tokenization with `MolBART`-Base on 100M samples. Since our hand-crafted tokenization produces more tokens, given a fixed compute budget, the model trained with this tokenization sees roughly 92M samples. Each model was trained on 128 A100 GPUs. As a baseline, we also fine-tune a randomly initialized model of the same architecture and size as `MolBART`-Base. We present our results in Table 1.

We significantly improve pre-training representations by using a learned tokenization method, and given that learned tokenization is more efficient (due to smaller context lengths) and pre-trains better, we select using a learned tokenizer for the rest of our experiments.
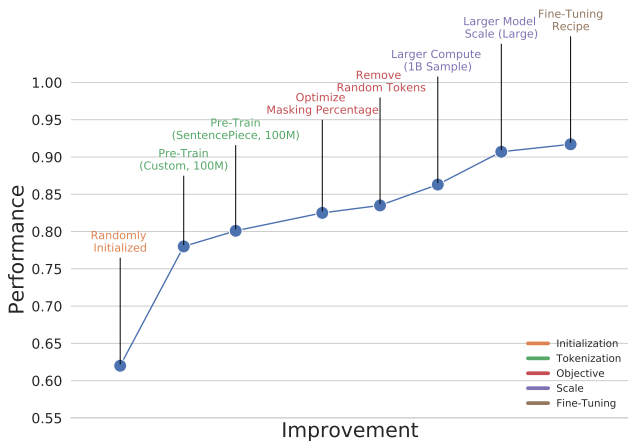
	Custom Tokenization (Random)	Custom Tokenization	Learned Tokenization
Performance	0.628	0.779	**0.801**

Table 1: Performance on ablation benchmark dataset over a randomly initialized network, a network pre-trained with hand-crafted tokenization rules, and a network pre-trained with SentencePiece tokenization.

Masking Strategy in Objective: We selected the denoising objective using encoder-decoder models from (Lewis et al., 2019) to learn fine-tunable representations for classification, regression, and generative tasks. Fundamentally the denoising objective has the following hyper-parameters `random_mask` (percent of tokens flipped to `<mask>`), λ (Poisson distribution for length of masked span), `randomize_tokens` (whether or not to randomize specific tokens, if true, 50% of `random_mask` is used for this). In conjunction with a `mask_token_budget` (the percent of tokens per sample allowed to be changed or masked), these three parameters cover the surface area of our ablation.

Given the computational cost of individual pre-training runs, even on the scale of `MolBART`-Base, we are unable to do a proper grid search; therefore, we aimed first to find an optimal `mask_token_budget` (Table 2a), then for that given budget, find optimal `random_mask` and λ (Table 2b), and finally given all the previous parameters we ablate whether or not to include to randomize specific tokens (Table 2c).

mask_token _budget	AUC
0.10	0.753
0.15	0.812
0.20	**0.821**
0.25	0.808
0.30	0.801
0.35	0.760
0.40	0.701

(a)

random _mask	λ	AUC
0.05	1.5	0.814
0.05	2.5	0.818
0.05	3.5	0.813
0.10	1.5	0.820
0.10	2.5	0.821
0.10	3.5	**0.821**

(b)

randomize _tokens	AUC
✓	0.821
✗	**0.835**

(c)

Table 2: All the ablations ran using the `MolBART`-Base model on 100 million samples from ZINC20. In (b) we ablate the impact of the random mask percentage and the poisson λ parameter which controls the length of the individual masks. For this we fix `mask_token_budget` to 0.20, the best forming configuration from (a). In (c) we ablate the impact of randomizing tokens in the input, using the best configurations found in (a) and (b)

In general, we see that the standard hyper-parameter of 0.3 for the `mask_token_budget` as proposed in Lewis et al. (2019) is suboptimal by a significant margin, and the use of randomizing tokens is suboptimal as well. The type of masking beyond that seems to not be of importance.

Importance of Pre-Training Scale As the last step in our ablation, we significantly increased the amount of compute to do a single pass over the full Zinc20 dataset, which required 20 hours on 1024 A100 GPUs. Training on the full dataset improves performance by five absolute points compared to the same setting trained on 100 million samples.

4 FINE-TUNING

4.1 DATASETS

To evaluate our pre-trained models, we fine-tune on multiple classification and regression datasets, mostly from the MoleculeNet benchmark suite (Wu et al., 2018).

MoleculeNet is a collection of 17 commonly used datasets for evaluating chemical representations. MoleculeNet defines splitting methods and evaluation metrics for all datasets. We have run experi-

	SIDER	ClinTox	Tox21	ToxCast	HIV	BACE	BBBP
RF MolCLR	0.684	0.713	0.769	-	0.781	0.867	0.714
SVM MolCLR	0.682	0.669	0.818	-	0.792	0.862	0.729
MoleculeNet GC	0.638	0.807	0.829	0.716	0.763	0.783	0.690
MGCN	0.552	0.634	0.707	-	0.738	*0.734	*0.85
D-MPNN	-	-	-	-	0.816	*0.875	*0.913
Attentive FP	0.637	0.94	**0.858**	0.805	*0.832	*0.85	*0.92
ChemBERTa 10M	-	-	-	-	0.622	-	0.643
MoLFormer-XL	0.69	0.948	0.847	-	0.822	*0.882	*0.937
GIN	0.627	0.726	0.781	0.657	0.799	0.845	0.687
GROVER large	0.658	0.944	0.831	0.737	-	*0.894	*0.94
MolCLR	0.68	0.932	0.798	-	0.806	0.89	0.736
iMolCLR	**0.699**	0.954	0.799	-	0.808	0.885	0.764
`MolBART`	**0.705**	**0.997**	**0.851**	**0.825**	0.745	0.855	0.74

Table 3: Results on the classification datasets from MoleculeNet. All numbers are AUC-ROC, higher is better. We bold the state-of-the-art numbers as well as all numbers within a 0.1 range. * indicates a non-standard split.

ments on 10 of them (excluding quantum mechanical datasets and the huge datasets PCBA and MUV). We use the AUC-ROC score to evaluate the classification models. Three of the selected datasets are regression tasks evaluated with RMSE. Four out of seven classification tasks have multiple target labels for each sample. We treat each target variable as a separate task and report the mean of the AUC scores. The remaining three are single-label classification tasks, and scaffolds of the molecules define their training-test splits. This implies some distribution shift in the test set, which introduces additional difficulties in model selection.

We additionally perform fine-tuning on two datasets developed for toxicology. The first one is called the Ames dataset Hansen et al. (2009), which holds the results of a bacterial reverse mutation test, commonly known as the Ames test. And the second one holds the results of another genotoxicity test, namely the micronucleus (MN) assay (Fan et al., 2018). MN assay is a cytogenetic test to assess the genotoxicity of chemicals and physical factors on a chromosomal level. We use cross-validation for the MN assay dataset to select the best hyperparameters and apply the best model to an external dataset of 65 compounds to obtain the final score. This makes our results comparable to the baselines introduced by (Hansen et al., 2009).

4.1.1 RECIPE

We fine-tune our pre-trained model with Fairseq (Ott et al., 2019), a sequence modeling toolkit based on PyTorch. We use the recommended hyperparameters for the GLUE benchmark with a few modifications (Wang et al., 2018). In particular, we set maximum input and output lengths to 128, changed dropout values, and set the gradient clipping norm to 0.1. We skip inputs longer than 128 and set the initial fp16 scaling to 128. We use batch size 16.

We train for ten epochs and use linear warmup, with the peak occurring at 16 percent of the training. After the peak. For each dataset, we do a fixed grid search over dropout and learning rate parameters with the values $[0.1, 0.2, 0.3]$ and $[5e-6, 1e-5, 3e-5]$, respectively. Finally, we apply Stochastic Weight Averaging (SWA) on three sets of four checkpoints: around the checkpoint having the best validation loss, the best validation accuracy, and the last checkpoint (Izmailov et al., 2018). We select the best one according to the performance on the validation set.

For multi-task problems, we train individual models for each task. We perform the grid search only on the first four tasks, pick the best hyperparameters and apply them to all tasks. We also tried to train a single multi-task classification model for each dataset to cover all tasks simultaneously, but the results were significantly worse.

4.2 BASELINES

MoleculeNet paper introduced its own set of baselines using multiple machine learning methods. MoleculeNet did not apply most baselines to all datasets. Among the ones applied to all datasets, the convolutional graph baseline performed the best, and we have included it as our baseline. The other graph-based supervised baselines include MGCN (Lu et al., 2019), D-MPNN (Yang et al., 2019), Attentive FP (Xiong et al., 2019) and the 3D InfoMax method (Stärk et al., 2022).

	ESOL	FreeSolv	Lipophilicity	Avg
RF MolCLR	1.070	2.030	0.880	1.320
SVM MolCLR	1.500	3.140	0.820	0.606
MoleculeNet GC	0.97	1.40	0.655	1.008
MGCN	1.270	3.350	1.110	1.910
D-MPNN	0.980	2.180	0.650	1.270
Attentive FP	0.503	0.736	0.578	0.605
Chemformer	0.633	1.230	0.598	1.231
MoLFormer-XL	0.279	0.231	0.529	0.346
GROVER large	0.831	1.544	0.560	0.978
MolCLR	1.110	2.200	0.650	1.320
iMolCLR	1.130	2.090	0.640	0.606
`MolBART`	**0.095**	**0.114**	**0.292**	**0.056**

(a) Results on the regression tasks. All numbers are RMSE, lower is better.

	AUC
FP_Pubchem_SVM	0.948
FP_MACCS_RF	0.947
Descriptor_SVM	**0.952**
Descriptor_RF	0.933
`MolBART`	0.914
`MolBART` (Frozen) + ECFP + SVM	**0.948**

(b) Micronucleus assay dataset

	AUC
FP_Pubchem_SVM	0.948
FP_MACCS_RF	0.947
Descriptor_SVM	**0.952**
Descriptor_RF	0.933
`MolBART`	0.914
`MolBART` (Frozen) + ECFP + SVM	**0.948**

(c) Ames dataset (cross-validation)

We also compare with baselines that use self-supervised pretraining. We skip SMILES-BERT, as SMILES-BERT did not test it on MoleculeNet. ChemBERTa (Chithrananda et al., 2020), Chemformer (Irwin et al., 2022) and MolFormer-XL (Ross et al., 2021) are based on regular text-based transformers. Graph Isomorphism Networks (GIN) with Context Prediction from Hu et al. (2020), GROVER-Large from Rong et al. (2020), MolCLR (Wang et al., 2022d) and iMolCLR (Wang et al., 2022c) are self-supervised methods based on graph neural networks. For Micronucleus Assay and Ames datasets we compare with the baselines introduced in the original papers.

4.2.1 RESULTS

Our model produced state-of-the-art results on all three regression tasks of the MoleculeNet suite, as seen in Table 2a. On classifications tasks the results are mixed (Table 3). On Clintox and Toxcast datasets, our model beats all previous models. Toxcast is a huge dataset with 617 tasks, so many papers do not include results. On SIDER and Tox21, the performance of our models is surprisingly poor, even compared to basic machine learning methods.

The other three classification tasks (HIV, BACE, and BBBP) require specific scaffold splits and are sensitive to the choice of the split. Unfortunately, many papers do not specify whether they used the split recommended by MoleculeNet, which makes comparisons difficult. We explored the split's impact on the BBBP dataset's results.

Molecule scaffolds induce 1024 clusters of 1940 molecules in the BBBP dataset. The split is organized so that all non-singleton clusters belong to the training set, while the validation and test sets contain only molecules with unique scaffolds. Most of the errors of our model are concentrated on roughly 30% of the singleton clusters in the test set. This implies that even a scaffold split with a different random seed can significantly affect the results. In Table 3, we have marked the results of these three classification datasets with an asterisk if we believe the authors used a different split.

The results for the two toxicology datasets that direct fine-tuning did not beat the baselines based on fingerprints and descriptors. To test whether the fingerprints still contain helpful information not covered by `MolBART`, we fine-tuned a classifier head that takes 2048-dimensional ECFP fingerprints (Rogers & Hahn, 2010) as an input in addition to the [CLS] vector. We also froze `MolBART`, averaged the last layers' representations per molecule, concatenated them with ECFP fingerprints, and ran the default SVM implementation from Kramer (2016). This helped close the gap with the baselines, as seen in Table 2c. Fundamentally this implies that fingerprints carry some information that is not trivially recoverable from our pre-trained representation. We do a deeper analysis of our pre-training dataset in § 5.2 and hypothesize that the Zinc20 dataset is not representative of the downstream tasks computational chemistry is interested in.

4.3 GENERATIVE TASKS

We follow a straightforward recipe for generative fine-tuning for all generative tasks, mainly derived from Aghajanyan et al. (2020a). Similarly to our results on SIDER, we found that training in full

32-bit precision (fp32) significantly impacted downstream results. We use the R3F method from Aghajanyan et al. (2020a), running a sweep over the noise types and λ regularization term for each task (See § E). Consistent with our fine-tuning recipe for classification and regression tasks, we utilize SWA and do not do any augmentation on the data. For additional stability, we initialize our fine-tuning optimizer (Adam) with the pre-training moving averages of the gradient and squared gradient from our pre-training. We also use a polynomial decay strategy where the peak learning rate is at $0.06 * \text{max\_updates}$ where the max number of updates is the number of updates needed to do ten epochs for the respective dataset.

During the evaluation, we implement two different strategies for sequence generation. The first runs a straightforward beam search with a beam size of ten. The second approach samples 128 samples from the natural distribution of tokens from the model (temperature of 1.0) and then re-ranks the list of 128 samples based on the perplexity of the entire predicted SMILE.

4.3.1 RETROSYNTHESIS

Retrosynthesis is a chemical synthesis technique involving the deconstruction of a target molecule into its starting materials to assess the best synthetic route. Retrosynthesis is a cornerstone of modern organic product synthesis, traditionally the domain of human experts' knowledge. Due to the algorithmic advances and the availability of large chemical reaction collections, computational approaches have been suggested to solve the problems in retrosynthesis (Davey, 2018; Watson et al., 2019). We apply our generative fine-tuning strategy to various subsets of the USPTO dataset Lowe (2012) in Table 3.

4.3.2 CHEMICAL REACTION PREDICTION

The dual problem of retrosynthesis is chemical reaction prediction which is the automatic construction of the target molecule given the set of reactant molecules. This can also be treated as a sequence-to-sequence task, as first shown by Schwaller et al. (2018a), and can be performed by fine-tuning `MolBART` on a relevant dataset. We follow the setup presented by Schwaller et al. (2019a) and apply our generative fine-tuning recipe to several subsets of the USPTO dataset. As seen in Table 4, `MolBART` outperforms all baselines across all subsets.

	Top-1	Top-5	Top-10
RetroComposer (Yan et al., 2021)	53.3	80.9	85.0
ATx100 (Tetko et al., 2020)	53.5	**81.0**	**85.7**
GraphRetro (Somnath et al., 2020)	53.7	72.2	75.5
ChemFormer (Irwin et al., 2022)	54.3	62.3	63.0
Dual-TF (Sun et al., 2020)	**55.3**	73.0	75.0
`MolBART` Beam-10	**55.2**	68.5	73.2
`MolBART` Sample-128 + PPL ReRank	**55.6**	74.2	80.9

Figure 3: We present our experimental results on the USPTO-50k retrosynthesis task across varying sampling strategies. We bold the best performing numbers within a 0.5 range.

USPTO Split	MIT (S)	MIT (M)	LEF (S)	LEF (M)	STEREO (S)	STEREO (M)
S2S (Schwaller et al., 2018b)	80.3				65.4	
WLDN (Jin et al., 2017)	79.6	74	84.0			
ELECTRO (Bradshaw et al., 2018)			87.0			
GTPN (Do et al., 2019)	82.4		87.4			
WLDN5 (Coley et al., 2019)	85.6		88.3			
Transformer Schwaller et al. (2019b)	90.4	88.6	**92.0**	90.3	78.1	76.2
`MolBART` Beam-10	**91.8**	**89.1**	**92.1**	**92.8**	**82.5**	**82.1**
`MolBART` Sample + ReRank	91.1	**89.1**	91.8	90.2	80.4	81.5

Table 4: We present our experimental results on the USPTO chemical prediction task across varying sampling strategies and data subsets. (S) signifies the "split" and (M) signifies the "mixed" sampling strategy from (mol_transformer). We bold the best performing numbers within a 0.5 range. The baselines scores are taken from Schwaller et al. (2019a).

5 INTERPRETABILITY

5.1 UNCOVERING THE KEY FEATURES

Radford et al. (2017) showed that a large language model trained to generate text had a particular neuron correlated with the sentence's sentiment, colloquially dubbed the sentiment neuron. Further

work showed that self-supervised representations had a very low intrinsic dimension for downstream tasks, arguing that self-supervised representations were capable of learning complex tasks implicitly (Aghajanyan et al., 2020b). For each dataset, we extracted the representations of each molecule by taking the average of the last layer representations of all tokens given by the pre-trained model and trained an L1-regularized logistic regression model to predict the label. We indirectly control the number of selected features by changing the regularization strength. The results are shown on Fig. 4.

Clintox tasks are among the easiest ones. As seen in Figure 4, a single unsupervised neuron can predict the label of its first subtask with a 0.77 ROC-AUC score, while with seven neurons, one can achieve 0.987 score. It is interesting to note that for the regularization parameter $C \in \{2^{-6}, 2^{-7}, 2^{-8}\}$, both subtasks of Clintox select the same set of features.

This might indicate that there are features learned in a completely unsupervised fashion that is highly correlated with the labels of these datasets, at least on the corresponding validation sets of the datasets.

5.2 Datasets in the Representation Space

To explore the relative positions of various datasets in the `MolBART` space, we compute the 1024-dimensional representations of all molecules by taking the average of token representations, and compute Frechet distance (Preuer et al., 2018) between the datasets. As the ZINC dataset is vast, we uniformly sample 0.05% of the molecules. As seen in Fig. 5, several datasets like BBBP and SIDER are further from ZINC, which might explain a relatively poor performance of `MolBART` compared to the models like iMolCLR, which do not use ZINC for pretraining.
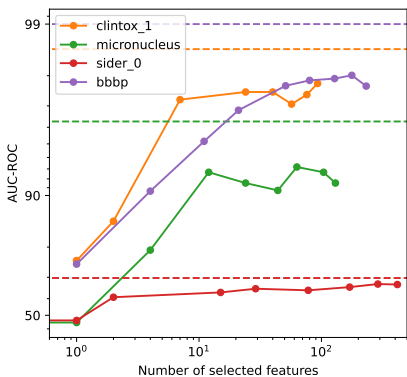

Figure 4: The AUC-ROC score plotted against the number of self-supervised features selected from `MolBART`. Dashed lines indicate the performance achieved on the same dataset with full fine-tuning of the model.
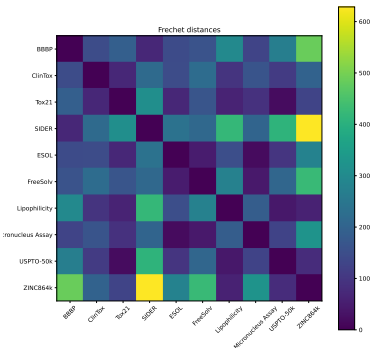

Figure 5: Frechet distances between datasets in the representation space of `MolBART`.

5.3 Interpreting Fine-Tuned Models

Methodology: In this section, we explore the contribution of each token of the given SMILES to the predictions of the fine-tuned model. We use the Integrated Gradients method (Sundararajan et al., 2017) implemented in Captum library (Kokhlikyan et al., 2020) to get the contributions of the tokens. Note that due to the tokenization we used in pretraining, there is no one-to-one mapping between tokens and the atoms. Nevertheless, we attempt to visualize the contributions of each atom on the molecular graph. If a token contains multiple symbols, we assume the contributions are additive and split the contribution of the token equally for all symbols in the token. Thus, we get the contribution of each atom and highlight the atom in the graph accordingly. Note that the SMILES string contains many non-atom symbols like brackets. We do not visualize the contributions of the brackets. We also visualize the contribution of a double-bond symbol $=$ by highlighting the bond in the graph.

Ames Dataset: In the Ames dataset, we explore the contributions of the atoms of the molecules from the test split. We compared the highlighted atoms in the graph with the known molecular
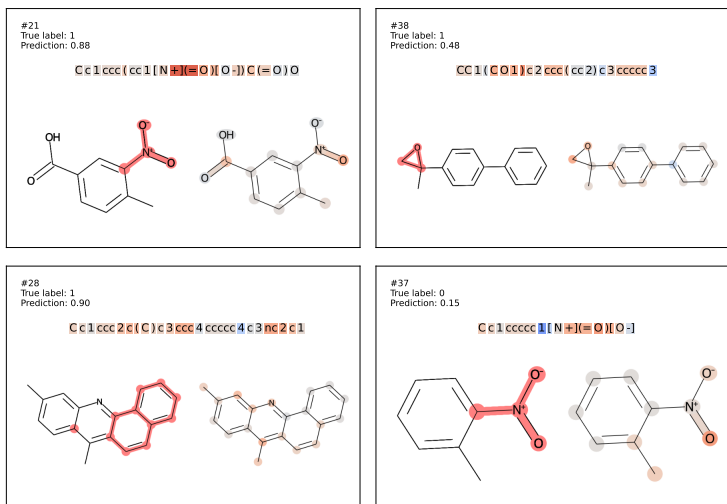
Figure 6: The contributions of each token and each atom to the prediction of our model fine-tuned on the Ames dataset, according to the Integrated Gradient algorithm.

substructures related to the mutagenic properties of molecules, so-called structural alerts (Plošnik et al., 2016). Figure 6 shows a couple of examples. On each of the plots, the left image highlights the structural pattern while the right one shows the contributions of each atom according to the Integrated Gradients method. Many more examples and a more thorough discussion can be found in Appendix C. The compound #21 in Fig. 6 is a nitroaromatic compound representing organic molecules that consist of at least one nitro group (-NO2) attached to an aromatic ring. The structural alert in this molecule is the nitro group (highlighted on the left part of the figure), whose components have been highlighted as a contribution in most cases of this class of compounds. The compound #38 on the same Figure is an epoxy group-bearing molecule. The epoxy group consists of an oxygen atom joined by single bonds to two adjacent carbon atoms, which form the three-membered epoxide ring (Plošnik et al., 2016) and is a well-known structural alert. It is worth noting that in the case of nitroaromatic compounds, the components of the structural alerts have been detected, along with the high probability (> 0.8) of correct prediction.

To summarize, the Integrated Gradients method, in most cases, recognizes contributions of non-complex substructures. In the case of more complex substructures, it provides explanations currently not understood through the lens of structural alerts. Additionally, the recognition of known substructures having the contribution to specific labels does not interfere with correct opposite label prediction, evidencing that the model learning ability is much deeper than estimating simple correlations between molecular substructures and their effects.

ESOL dataset: We did a similar analysis for the models trained on ESOL dataset. Integrated Gradients method successfully highlights a few substructures that are known to make molecules more soluble. The examples and details are presented in Appendix B. We hope the interpretability methods developed for deep learning, our visualization toolkit, and the strong predictive models made possible by the large pre-trained models will help chemists uncover new correlations between substructures and target variables or even causal links.

6 CONCLUSION

In this paper, we introduce `MolBART`, a large pre-trained model for molecular representations which sets state-of-the-art results across many classification, regression, and generation tasks. We showed the power of self-supervised training through a series of quantitative and qualitative analyses of representations showing that interpretability methods applied to these representations mirror that of chemists. There is still room for investigation of the impact of data on pre-training, i.e., ZINC contains more than a billion molecules, but its distribution might be irrelevant for many downstream tasks. Further collaboration between machine learning and chemistry experts might allow discovering of new molecular substructures linked to chemical phenomena.

REFERENCES

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*, 2020a.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020b.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Gor Arakelyan, Gevorg Soghomonyan, and The Aim team. Aim. 6 2020. doi: 10.5281/zenodo. 6536395. URL https://github.com/aimhubio/aim.

Mandeep Baines, Shruti Bhosale, Vittorio Caggiano, Naman Goyal, Siddharth Goyal, Myle Ott, Benjamin Lefaudeux, Vitaliy Liptchinsky, Mike Rabbat, Sam Sheiffer, Anjali Sridhar, and Min Xu. Fairscale: A general purpose modular pytorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Samuel Boobier, David RJ Hose, A John Blacker, and Bao N Nguyen. Machine learning with physico-chemical relationships: solubility prediction in organic solvents and water. *Nature communications*, 11(1):1–10, 2020.

John Bradshaw, Matt J Kusner, Brooks Paige, Marwin HS Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. *arXiv preprint arXiv:1805.10970*, 2018.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

Stephen G Davey. Retrosynthesis: Computer says yes. *Nature Reviews Chemistry*, 2(1):1–1, 2018.

John Dearden. The history and development of quantitative structure-activity relationships (qsars). *IJQSPR*, 1(1):1–44, 2016. doi: 10.4018/978-1-5225-0549-5.ch003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Kien Do, Truyen Tran, and Svetha Venkatesh. Graph transformation policy network for chemical reaction prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 750–760, 2019.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

Defang Fan, Hongbin Yang, Fuxing Li, Lixia Sun, Peiwen Di, Weihua Li, Yun Tang, and Guixia Liu. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicology research*, 7(2):211–220, 2018.

Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9(3):199, 2006.

Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9):2077–2081, 2009. doi: 10.1021/ci900161g. URL `http://dx.doi.org/10.1021/ci900161g`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJlWWJSFDH`.

John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073, 2020.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022, 2022.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608, 2016.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pp. 45–53. Springer, 2016.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

Thomas L Lemke. *Review of organic functional groups: introduction to medicinal organic chemistry*. Lippincott Williams & Wilkins, 2003.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1052–1060, 2019.

Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52 (6):1686–1697, 2012.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Alja Plošnik, Marjan Vračko, and Marija Sollner Dolenc. Mutagenic and carcinogenic structural alerts and their mechanisms of action. *Arhiv za higijenu rada i toksikologiju*, 67(3):169–182, 2016.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Do large scale molecular language representations capture important structural information? *arXiv preprint arXiv:2106.09553*, 2021.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018a.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018b.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019a.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019b.

Vignesh Ram Somnath, Charlotte Bunne, Connor W Coley, Andreas Krause, and Regina Barzilay. Learning graph models for template-free retrosynthesis. *arXiv preprint arXiv:2006.07038*, 2020.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.

Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Energy-based view of retrosynthesis. *arXiv preprint arXiv:2007.13437*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1): 1–11, 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://www.aclweb.org/anthology/W18-5446.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin Burke. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=6sh3pIzKS-.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 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*, pp. 429–436, 2019.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective work best for zero-shot generalization? *arXiv preprint arXiv:2204.05832*, 2022b.

Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2022c. doi: 10.1021/acs.jcim.2c00495.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022d.

Ian A Watson, Jibo Wang, and Christos A Nicolaou. A retrosynthetic analysis algorithm implementation. *Journal of cheminformatics*, 11(1):1–12, 2019.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Chaochao Yan, Peilin Zhao, Chan Lu, Yang Yu, and Junzhou Huang. Retrocomposer: Discovering novel reactions by composing templates for retrosynthesis prediction. *arXiv preprint arXiv:2112.11225*, 2021.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388, 2019.

A DATASET DESCRIPTIONS

Dataset	Task type	# of tasks	# of compounds	Split type	Metric
ESOL	Regression	1	1128	Random	RMSE
FreeSolv	Regression	1	642	Random	RMSE
Lipophilicity	Regression	1	4200	Random	RMSE
HIV	Classification	1	41127	Scaffold	ROC-AUC
BACE	Classification	1	1513	Scaffold	ROC-AUC
BBBP	Classification	1	2039	Scaffold	ROC-AUC
Tox21	Classification	12	7831	Random	ROC-AUC
ToxCast	Classification	617	8575	Random	ROC-AUC
SIDER	Classification	27	1427	Random	ROC-AUC
ClinTox	Classification	2	1478	Random	ROC-AUC
Ames	Classification	1	6512	Random	ROC-AUC
Micronucleus Assay	Classification	1	641	Random	ROC-AUC

Table 5: Dataset statistics for our fine-tuning evaluations.

B INTERPRETABILITY OF AQUEOUS SOLUBILITY PREDICTION

In this section we analyze the attribution scores Integrated Gradients gave for one of the regression tasks, ESOL, which aims to measure the solubility of the molecules in water.

We first noted a strange pattern: the first token of each molecule always received a positive score, regardless of the molecule's label or structure. We measured the mean attribution score across the tokens with respect to their positions and confirmed the bias (Fig. 7, top-left). We decided to "normalize" the attribution scores by subtracting those mean attributions from the scores given by the Integrated Gradients method.

Since water is a polar solvent, organic molecules having polar groups, such as hydroxyl group (-OH) and carbonyl group (C=O), will contribute to the solubility of the molecule (Lemke, 2003). The Integrated Gradients method highlighted almost all instances of hydroxyl groups, as seen in the bottom row of Fig. 7, and most instances of carbonyl groups, especially if the molecule is more soluble (top-right image of Fig. 7). It is known that long chains of hydrocarbon compounds are known to decrease the solubility of the molecules. This is captured by the negative attributions of "C" tokens in such molecules, as seen in the bottom row of Fig. 7.

More examples and analysis is presented in Appendix D.

C AMES INTERPRETABILITY: DETAILED ANALYSIS

In Figures 8 to 12 we show the structural alerts of more than a hundred molecules from the test set of Ames dataset. Chemicals under the numbers 10, 20, 21, 37, 40, 61, 71, 90, 93, 110, 117, and 127 are nitroaromatic compounds representing organic molecules that consist of at least one nitro group (-NO2) attached to an aromatic ring. These compounds and their residuals are well known for their carcinogenic and mutagenic potential. The nitro group in these molecules is characterized as structural alerts, i.e., molecular substructures related to the chemical's carcinogenic and mutagenic properties.

In most cases, Integrated Gradients recognized two components of the structural alert: the positively charged nitrogen atom and a double bonded oxygen atom. In some cases, only a positive charge was highlighted along with the oxygen atom. It is essential to mention that a positively charged nitrogen atom plays a crucial role in the toxic potential of this structural group since this positive charge determines the electrophilic nature of the molecule. An electrophile is a chemical species that form bonds with nucleophiles by accepting an electron pair. Whereas DNA is abundantly equipped with nucleophilic sites, reaction with electrophiles results in diverse chemical-DNA interactions.
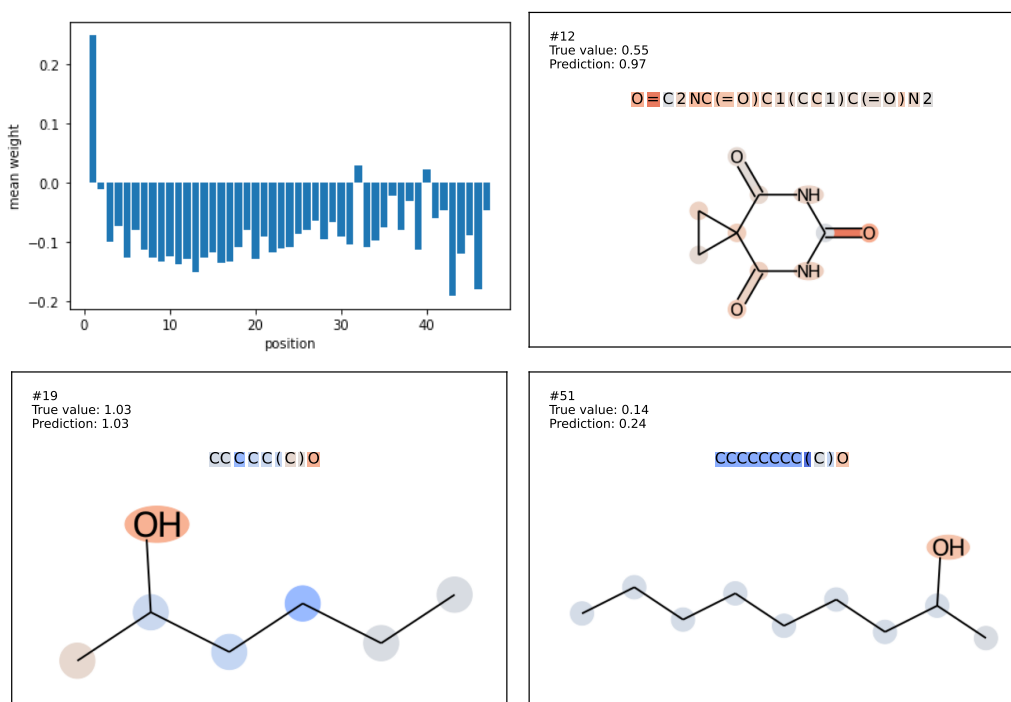
Figure 7: Top-left: the average attribution score of a token depending on its position in the tokenized SMILES. There is a significant positive bias on the first token. Others: the contributions of each token and each atom to the prediction of our model fine-tuned on the ESOL dataset, according to the Integrated Gradient algorithm.

However, in all cases, the most highlighted atom was a double-bonded oxygen atom. Two explanations may arise from this observation. First is that Integrated Gradients "detects" it as a nitroso group (N=O), e.g., as in the compounds under the numbers 42, 59, and 69, where they represent a separate structural alert group. Second, the activation of most nitroaromatic compounds' mutagenic properties is linked to the nitroreduction. As a result of nitroreduction, nitroso and superoxide species are produced.In both cases, double-bonded oxygen atoms play a vital role in forming toxic intermediates. So, in this case, it can be assumed that the algorithm recognizes target sites, such as positively charged nitrogen or oxygen atoms, responsible for producing DNA-damaging compounds.

The same pattern is observed in the case of compounds under the numbers 18, 38, 57, 60, 124, and 126. These compounds are characterized by the bearing of at least one epoxy group consisting of an oxygen atom joined by single bonds to two adjacent carbon atoms, which form the three-membered epoxide ring.The epoxy group is a well-known structural alert. Compounds bearing epoxy groups are alkylating agents in which carbon atoms represent electrophilic sites that react with nucleophilic DNA to form covalent bonds via nucleophilic substitution reaction, thus acting as direct genotoxins and carcinogens. Integrated Gradients recognized this structural alert in all cases, primarily highlighting the oxygen atom joined to one of the two carbon atoms. Most often, the electrophilic carbon atom responsible for chemical-DNA interaction was the most highlighted area, assuming that Integrated Gradients highlighted the most active site of the molecule from the genotoxicity point of view. Interestingly, the highlighting of the oxygen atom is always accompanied by one or two equally highlighted carbon atoms and never appears separately, only oxygen or carbon atoms. Presumably, the method highlights the bond of these atoms, which is the second most important characteristic of epoxy groups, since, during the nucleophilic substitution reaction, the three-membered ring is opened via one of the cleavages of the O-C bond, forming the favorable for the molecule unstrained acyclic intermediate.

It is worth noting that in the case of nitroaromatic compounds, the components of the structural alerts have been detected, along with the high probability (> 0.8) of correct prediction. In contrast, in the
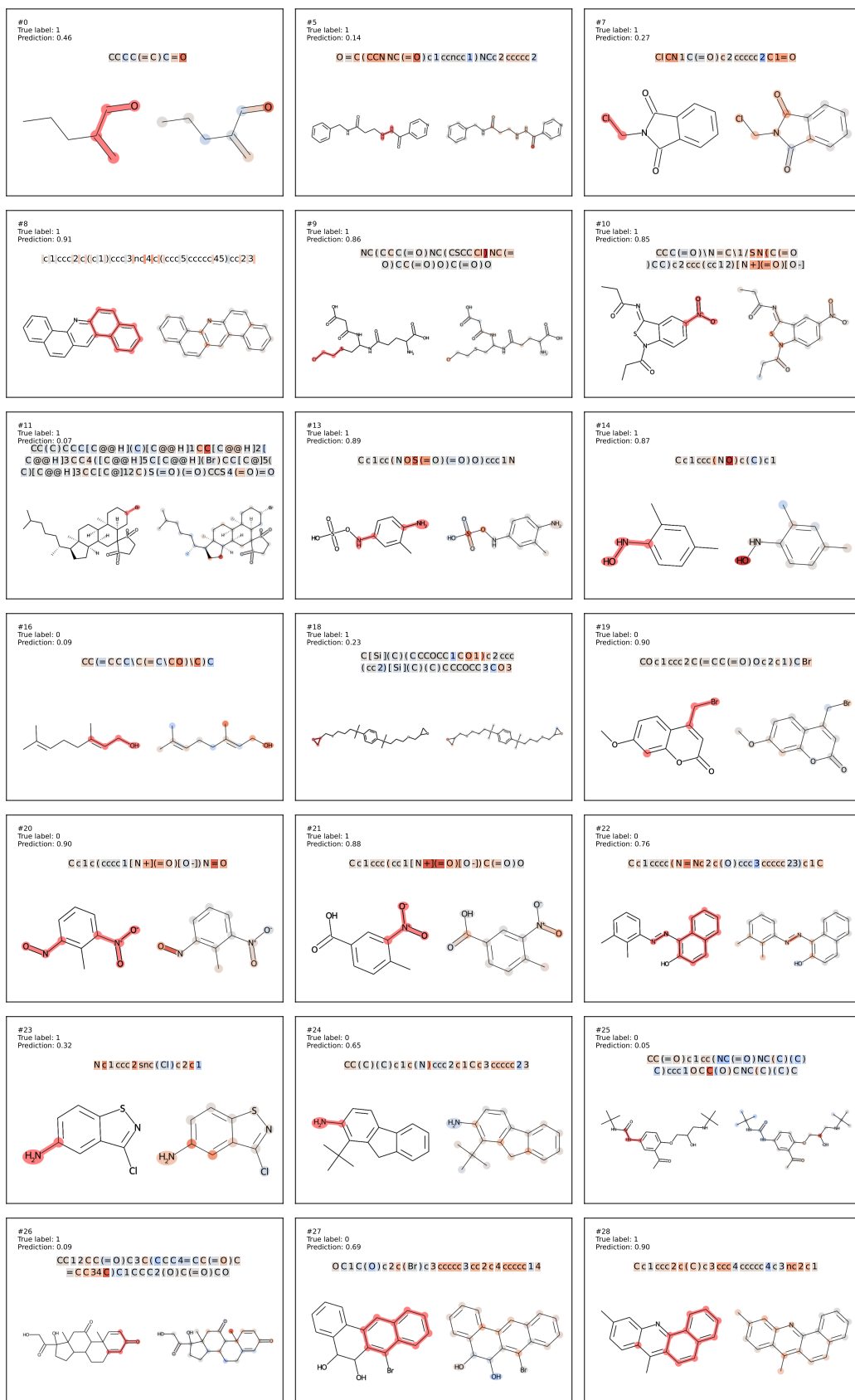
Figure 8: Explaining predictions of the fine-tuned model on Ames dataset. See Section 5.3. Part 1
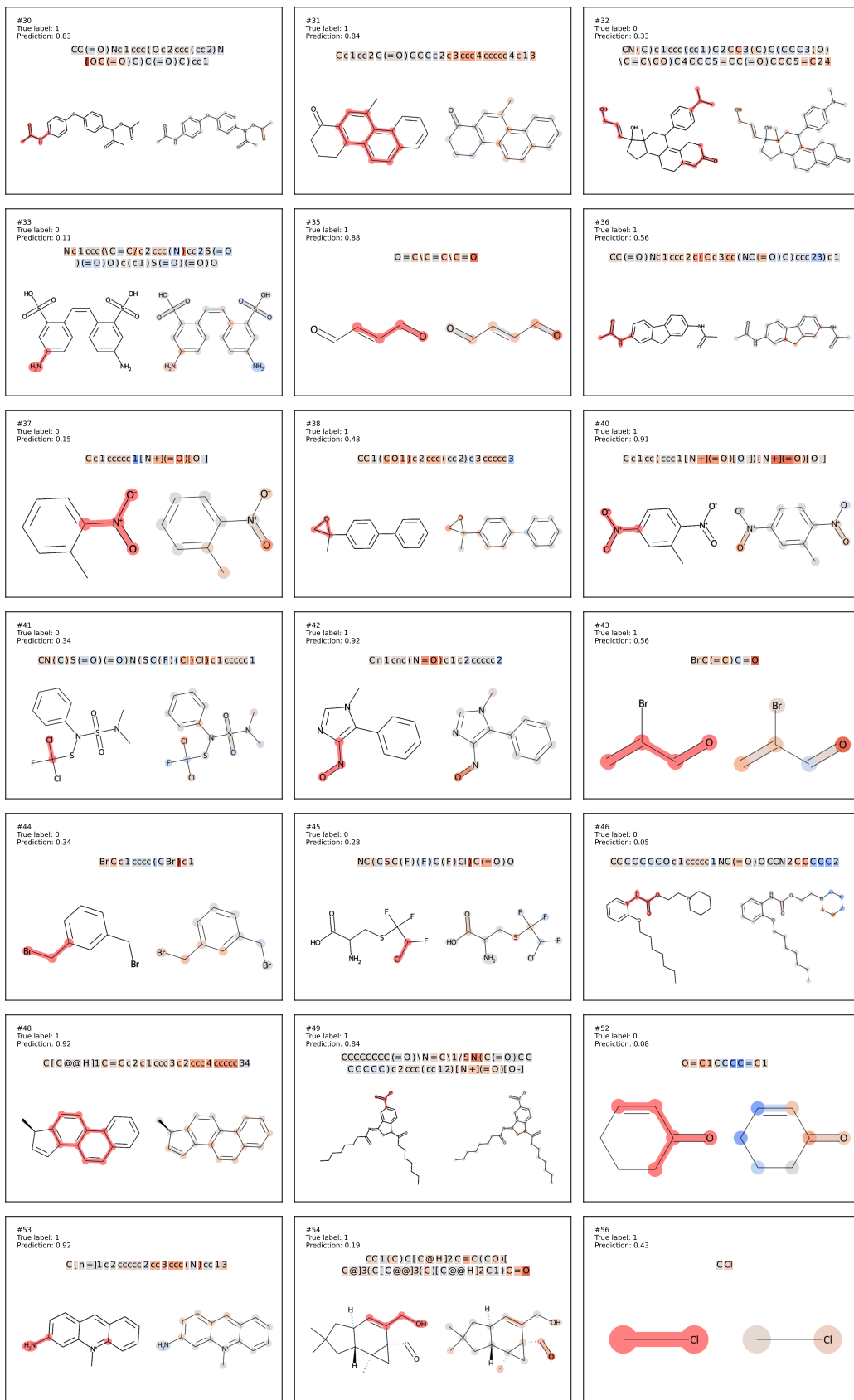
Figure 9: Explaining predictions of the fine-tuned model on Ames dataset. See Section 5.3. Part 2
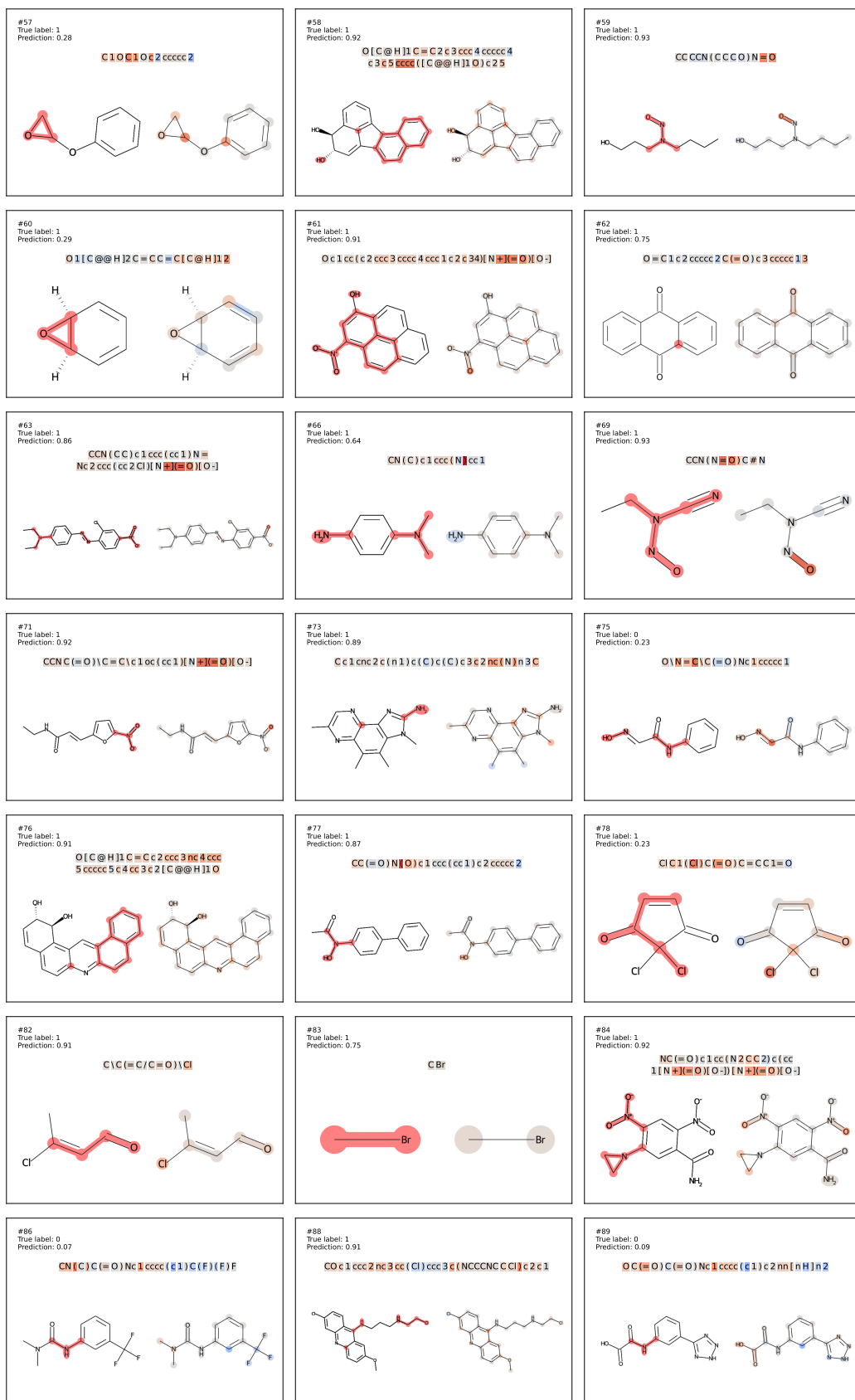
Figure 10: Explaining predictions of the fine-tuned model on Ames dataset. See Section 5.3. Part 3
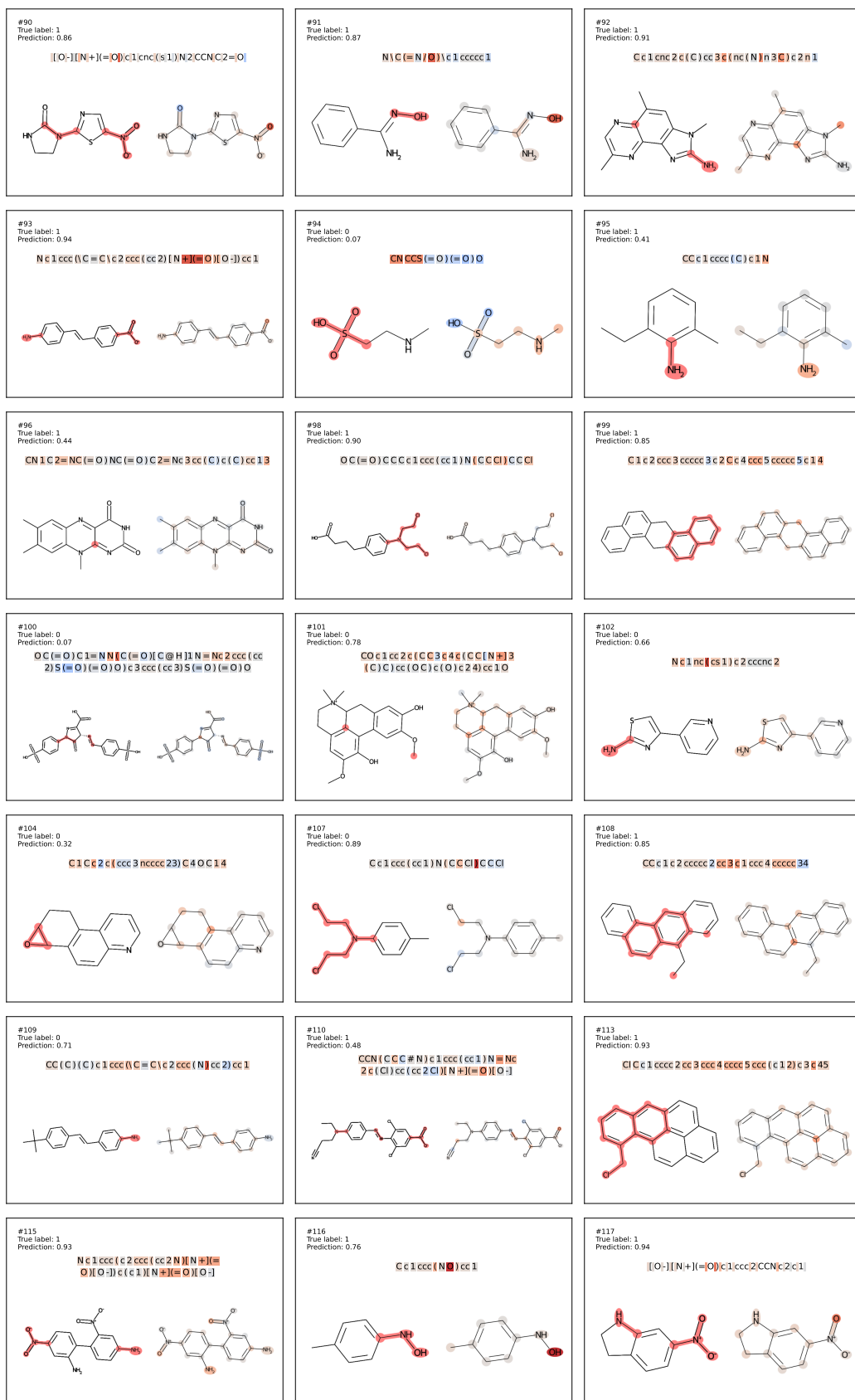
Figure 11: Explaining predictions of the fine-tuned model on Ames dataset. See Section 5.3. Part 4
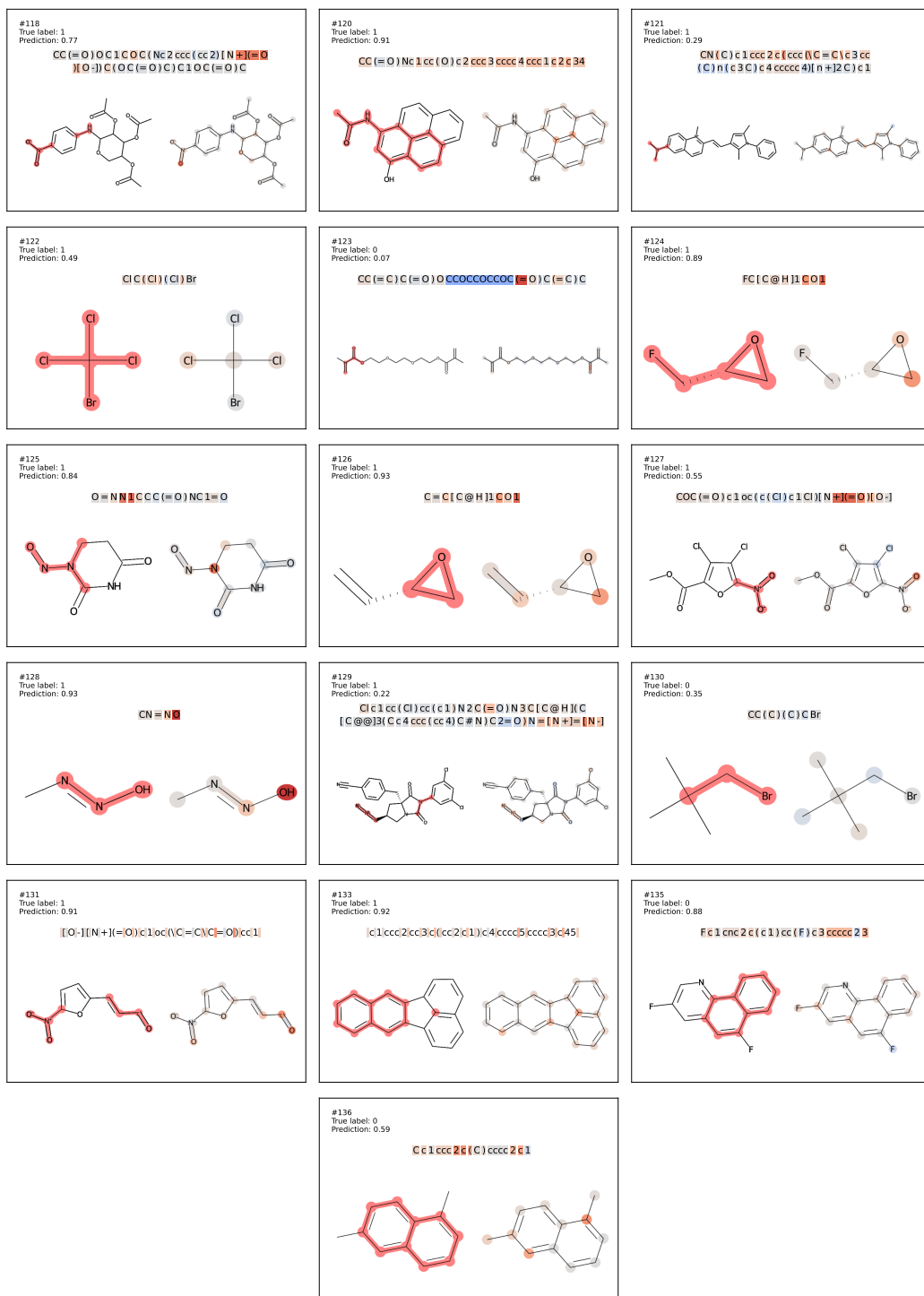
Figure 12: Explaining predictions of the fine-tuned model on Ames dataset. See Section 5.3. Part 5

case of epoxy compounds, the structural alerts have been detected correctly, but the probability level of correct predictions was around 0.22-0.48, except for compounds 124 and 126 (probability levels 0.891 and 0.932).

An example where the attributions given by Integrated Gradients is not correlated with the structural alerts is the case of the naphthalene group in polycyclic aromatic hydrocarbons (compounds under the numbers 8, 28, 31, 48, 58, 62, 61, 76, 99, 113, and 108). The naphthalene group is a bicyclic aromatic hydrocarbon representing a structural alert in all mentioned compounds. Although the attributions of Integrated Gradients were correct with the high probability level in all cases, the structural alert has not been recognized. Here, the activated neurons have been dispersed distributed throughout polycyclic aromatic hydrocarbons, rarely matching the part of the structural alert.

D ESOL INTERPRETABILITY: DETAILED ANALYSIS

In Figures 13 to 15 we show the attributions on 60 molecules from the test set of ESOL dataset. The images with indices 10, 16, 19, 27, 30, 37, 50, 51 and 55 contain hydroxyl groups, which are properly highlighted in red by the Integrated Gradient method. Same with carbonyl groups, as seen on molecules with indices 2, 5, 7, 12,25, 26, 46 and 47.

It is usually assumed that the noise in aqueous solubility datasets is roughly 0.7 Boobier et al. (2020). It is noteworthy that in all instances, the errors of the predicted values are less than that and, hence, can be assumed to be accurate.

In one case, out of all molecules bearing the hydroxyl group, in compound #48, the functional group was not highlighted and even marked as a negative contributor. Even so, the prediction value of solubility was correct.

One more interesting example where the hydroxyl group has been recognized, but the prediction was incorrect (LogS ± 0.76) is compound #34. In this compound, in addition to the hydroxyl group, four alkyl chloride (C-Cl) groups exist and contribute to the molecule's solubility (have a lower level of polarity). The lower proportion of alkyl chloride-bearing compounds in the training set may lead to underestimating these functional group contributions to the molecule's solubility.

In the case of carbonyl groups, there are more examples where the Integrated Gradients method didn't highlight the functional groups, for instance, compounds under the numbers 13, 14, 17, 35, 36, 40, 41, 42 and 43. Note that logS values of all these compounds were in the range of 0 to -2, which is the range of less soluble compounds, compared to the solubility of aforementioned compounds with recognized carbonyl groups (in the range of 0 and higher, highly soluble). Additionally, all compounds with unrecognized carbonyl groups are complex, polycyclic aromatic hydrocarbons, which, as mentioned in the Ames dataset, was more problematic for the analysis by the Integrated Gradients method.

Other interesting examples are in the aliphatic compounds, which are hydrocarbon compounds containing carbon and hydrogen joined together in straight chains with carbonyl groups (Compounds # 7, 25, 26, 28) or hydroxyl groups (Compounds # 19, 48, 51). A general rule for these compounds is that the longer the carbon chain, the lower the solubility in polar solvents such as water. Here, we see that the increasing amount of CH2 groups in a carbon chain enhances the hydrophobic effect, which decreases the solubility value. Hence, compounds #28 and #51 have the lowest solubility compared to the abovementioned examples. It is remarkable that the model somehow followed the rule, and predicted values also exhibit the same pattern. As to the Integrated Gradients method, the negative contribution of CH2 groups is constantly highlighted in all compounds, whereas peripheric CH2 groups have been marked as small contributions to solubility.

There are other examples where the predictions were accurate, but the underlying chemistry is more complicated, and the highlighted substructures sometimes are unexplainable.

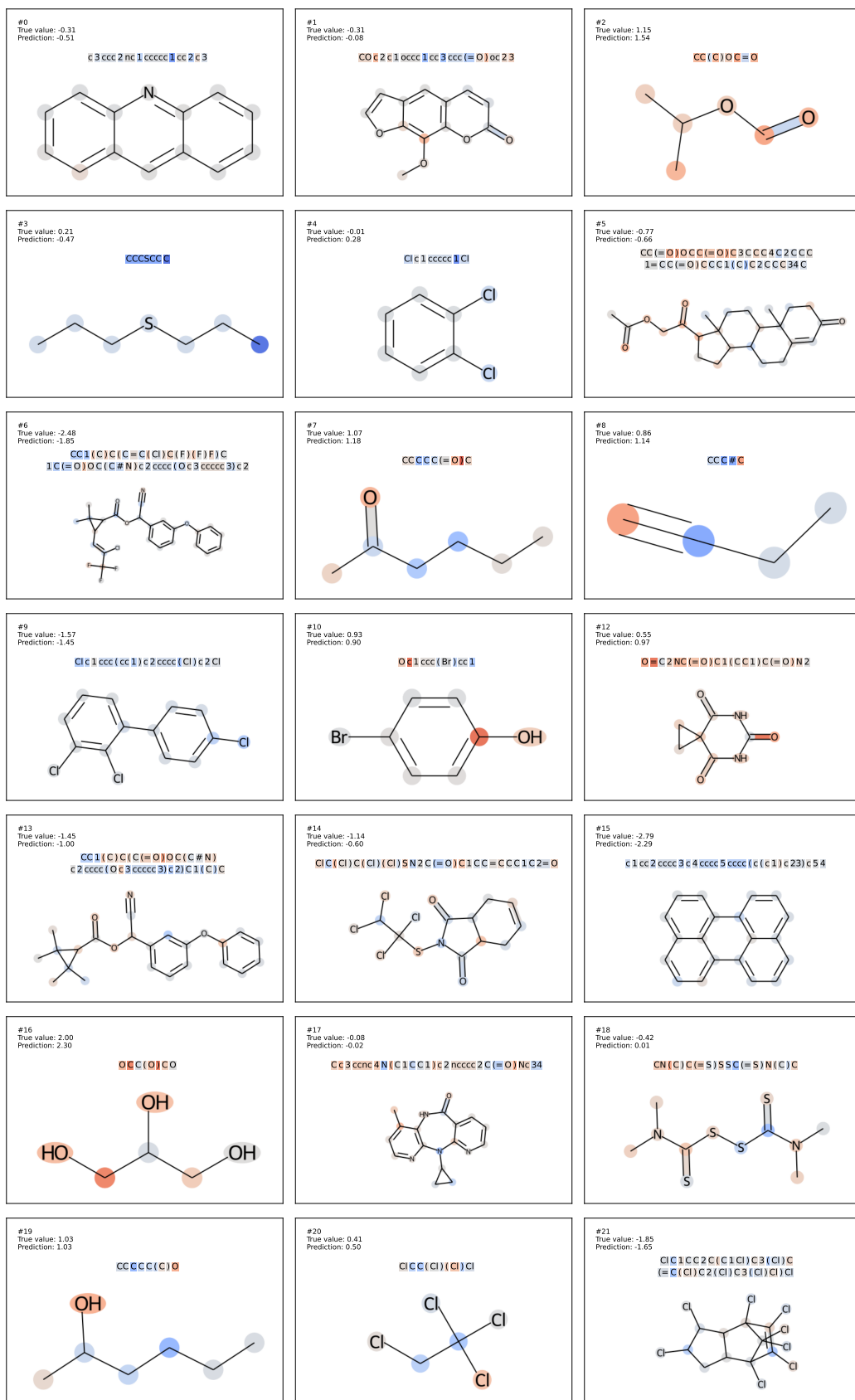## E FINE-TUNING HYPER-PARAMETERS

Figure 13: Explaining predictions of the fine-tuned model on ESOL dataset. See Section B. Part 1
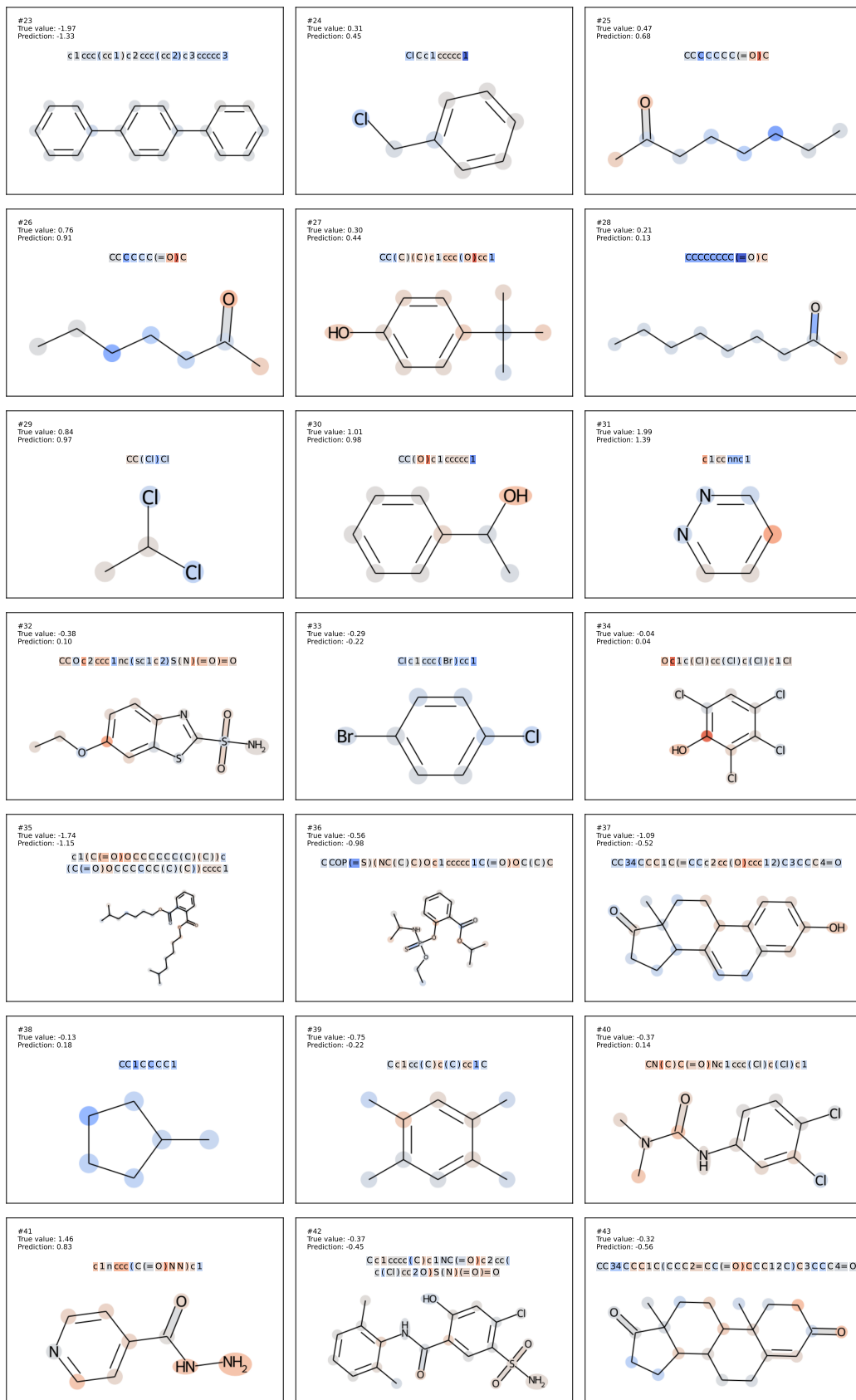
Figure 14: Explaining predictions of the fine-tuned model on ESOL dataset. See Section B. Part 2
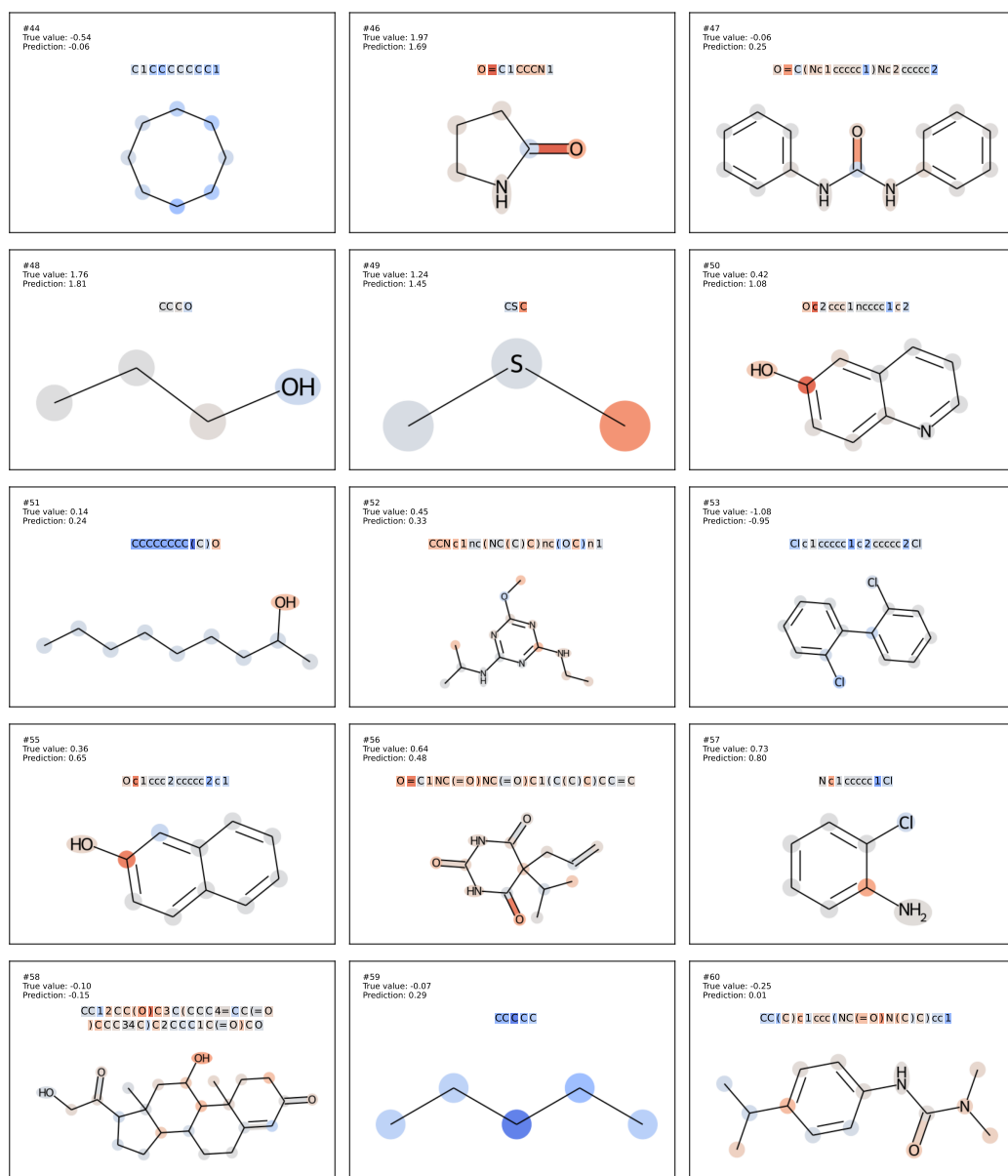
Figure 15: Explaining predictions of the fine-tuned model on ESOL dataset. See Section B. Part 3

Hyper parameter	Value
Optimizer	Adam
Adam-betas	(0.9, 0.98)
Adam-eps	1e-8
LR Scheduler	polynomial decay
Learning Rate	3e-05

Hyper parameter	Value
λ	[0.001, 0.01, 0.1]
Noise Types	$[\mathcal{U}, \mathcal{N}]$
σ	$1e-5$
Dropout	0.1
Weight Decay	0.01
Clip Norm	0.1

Table 6: Hyper parameters for the use of R3F on all generative fine-tuning experiments.