Probing Chemical Language Models: Effects of Pre-training and Fine-tuning

Anna Karnysheva[1,2], **Dietrich Klakow**[1,2,3]*, **Ji-Ung Lee**[1]*
[1]RTG Neuroexplicit Models [2]Spoken Language Systems, [3]PharmaScienceHub (PSH)
Saarland University
akarnysheva@lsv.uni-saarland.de,
dietrich.klakow@lsv.uni-saarland.de,
ji-ung.lee@uni-saarland.de

Abstract

Chemical language models (CLMs) are trained with linearized representations such as SMILES, yet it remains unclear which *chemically meaningful* substructures they encode. To foster a better understanding of CLMs, we conduct a systematic study and probe for 78 molecular substructures across eight pre-trained and six randomly initialized models. We furthermore study how fine-tuning on chemical downstream tasks affects the learned representations of molecular substructures. Our results show that pre-training generally improves molecular structure awareness of CLMs, particularly in the upper layers. Moreover, randomly initialized models already encode ring structures well in the first layer. Our analysis on two chemical downstream tasks further reveals that, interestingly, fine-tuning affects task-relevant molecular substructures more than others, indicating that the changes in the representations follow chemical theory.[1]

1 Introduction

Drug discovery is an inherently expensive process that includes labor- and time-intensive steps such as designing molecules that are both effective against a target disease and can be safely administered to humans (Jia et al., 2020). One important task in this process is molecular property prediction (MPP, Shen and Nicolaou 2019), i.e., to reliably predict properties such as the lipophilicity, solubility, permeability, bioactivity, or toxicity of a molecule.[2]

Many deep neural network (DNN) architectures—including general-purpose large language models (LLMs, Zhao et al. 2023), GNNs (Scarselli et al., 2009), graph transformers (Yun et al., 2019), and sequence-based chemical language models (CLMs, Wang et al. 2019)—have been explored for MPP tasks, however, they still frequently fall behind feature-based models (Dias et al., 2023; Xia et al., 2023; Sadeghi et al., 2024). Moreover, they often exhibit poor out-of-distribution generalization (Tossou et al., 2024) and are not evaluated on regression tasks which make up a substantial portion of MPP tasks (Liu et al., 2026). While a few works have tried to establish a better understanding of the shortcomings of DNNs by probing their representation for individual molecular substructures, they are often limited to a small set of molecular substructures and graph-based models which are often trained for individual MPP tasks (Akhondzadeh et al., 2023; Wang* et al., 2023; Volkov et al., 2022).

In this work, we focus on CLMs trained on linearized molecular representations (i.e., SMILES, Weininger 1988) which have been frequently used (Singh et al., 2026), but not well studied, especially regarding whether they learn to capture molecular substructures during pre-training (**RQ1**) and how fine-tuning on chemical downstream tasks affects these representations (**RQ2**). Our systematic empirical investigation based on a new probing dataset comprising 78 molecular substructures evaluated across eight pre-trained (PT) and six randomly initialized (RI) models reveals that:

- Pre-training improves molecular structure awareness towards the upper layers. Also, all molecular substructures exhibit a change larger than $\pm 1\%$ in at least one model.
- RI models already encode ring structures well, but not other substructures.
- Some molecular substructures are unlearned in all models during pre-training.

Studying how fine-tuning on lipophilicity and solubility prediction affects the representations of molecular substructures reveals that:

*Co senior authors.

[1]Code and data are available at: https://github.com/basagashka/substructure-probing-clms

[2]We provide an introduction into chemistry in Appendix A.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 60585–60627
October 24-29, 2026 ©2026 Association for Computational Linguistics

- Pre-training increases the robustness of molecular substructures during fine-tuning.
- Fine-tuning affects the representations less compared to pre-training with changes occuring more frequently in the upper layers.
- Molecular substructures that are theoretically more relevant for lipophilicity and solubility prediction are more affected by fine-tuning.

Finally, we explore how probing can be used to identify molecular substructures on which models may not have been sufficiently trained, and whether further pre-training on molecules containing these substructures can improve their representations.

2 Related Work

2.1 Molecular Representation Learning

Molecules can be represented in various ways in order to be processed by models or hand-crafted algorithms. The choice of representation directly affects what architectures are suited; e.g., representing molecules as graphs enables the use of different kinds of GNNs. To train language models, we use linearized molecule representations such as SMILES (Weininger, 1988). For chemistry, works have trained encoder-decoder, encoder-only, and decoder-only models.

Encoder-decoder models Works have utilized encoder-decoder models in tasks that mirror their sequence-to-sequence nature (Schwaller et al., 2019; Irwin et al., 2022; Lu and Zhang, 2022). Example tasks are chemical reaction prediction (CRP), i.e., predicting the outputs for a given set of inputs (Fooshee et al., 2018) and molecular optimization (MO), where an input molecule is altered to achieve desired properties (He et al., 2021).

Encoder-only models Encoder-only models (here, referred to as CLMs) have primarily been trained for MPP. To improve task performance, works have utilized different linear molecular representations (Krenn et al., 2020; Yüksel et al., 2023; Leon et al., 2024), domain-specific auxiliary training objectives (Fabian et al., 2020; Ahmad et al., 2022; Wu et al., 2022; Li and Jiang, 2021; Park et al., 2024b,a), different tokenization schemes (Chithrananda et al., 2020; Ahmad et al., 2022; Leon et al., 2024), positional encodings (Ross et al., 2022; Liu et al., 2023a) and attention mechanisms (Ross et al., 2022).

Decoder-only models While early works utilize decoder-only models to generate new molecules (Xue et al., 2022; Bagal et al., 2022; Wang et al., 2023), later studies consider them for other tasks such as MPP, CRP, and MO (He et al., 2022; Mazuz et al., 2023). Others have even devised novel text-centric tasks such as molecule captioning (Edwards et al., 2021) or augmented tasks with textual instructions (Edwards et al., 2022; Wang et al., 2023; Liu et al., 2023b; Christofidellis et al., 2023; Fang et al., 2024; Lin et al., 2026).

Limitations Despite all efforts to utilize LLMs in chemistry (Wang et al., 2023; Ye et al., 2025; Xian et al., 2025), recent works found that both chemical and general-purpose LLMs struggle to understand molecular structure (Jang et al., 2025; Ganeeva et al., 2024) or are outperformed by simple baselines (Guo et al., 2023).[3] For instance, Xian et al. (2025) show that GNNs outperform general-purpose LLMs on classification as well as regression tasks such as lipophilicity, which constitute a large portion of MPP tasks (Liu et al., 2026).

2.2 Probing

Probing is widely used in NLP to investigate the extent to which LMs capture linguistic knowledge, such as syntactic (Jawahar et al., 2019; Tenney et al., 2019b; Liu et al., 2019a; Hou and Sachan, 2021) or semantic information (Tenney et al., 2019b). Typically, probing involves training a classifier for a specific probing task (e.g., part-of-speech tagging) using hidden representations extracted from a pre-trained LM (Belinkov, 2022).

Probing for linguistic knowledge Various works localize linguistic knowledge in pre-trained LMs (particularly BERT), attributing syntax and semantics to different layers (Peters et al., 2018; Tenney et al., 2019a; Jawahar et al., 2019; Hewitt and Manning, 2019). Others investigate the effect of fine-tuning, finding that changes are more centered around upper layers (Mosbach et al., 2020; Zhou and Srikumar, 2022) and are task-dependent (Merchant et al., 2020). In general, works have found that many of these changes are less pronounced than during pre-training and vary from task to task.

Probing for chemical knowledge Few works explore the encoding of molecular substructures in models. Prior works focus on graph-based models, finding that GTs generally encode ten molec-

[3] Our experiments in Appendix J support these findings.

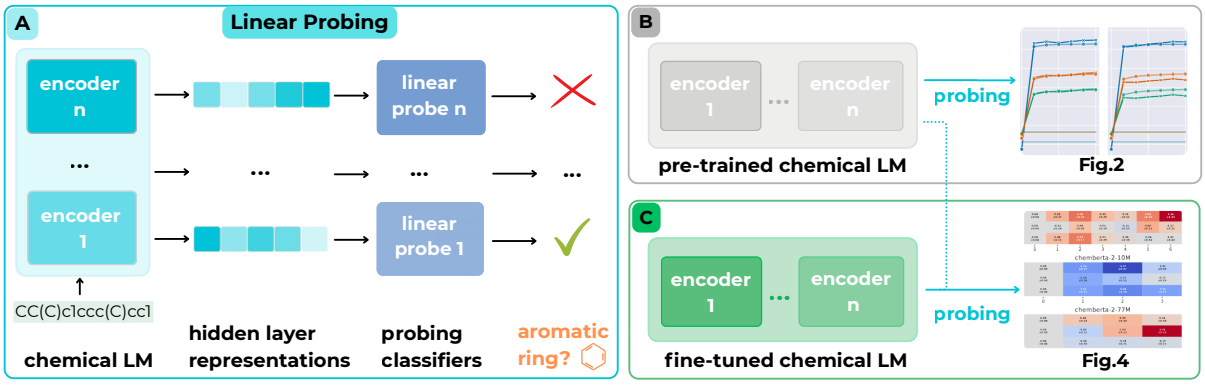


Figure 1: Overview of the experimental setup. A illustrates the general linear probing setup described in §4 for a single probing task, here the presence of an aromatic ring. B and C correspond to the two research questions investigated using this setup. RQ1 (§5) investigates the effect of pre-training on molecular substructure encoding, as shown in B. In contrast, RQ2 (§6) examines changes in molecular substructure encoding following fine-tuning by comparing the probing performance of pre-trained and fine-tuned models, as shown in C.

ular substructures better than message-passing GNNs; and that the molecular substructures are already encoded well by random initializations (Akhondzadeh et al., 2023). Others study differences between pre-training and fine-tuning of GNNs on MPP tasks and report a positive correlation between probing and MPP performance (Wang* et al., 2023). Only Payne et al. (2020) and Fender et al. (2025) investigate CLMs either for visualization or only for two molecular substructures. Finally, some works have shown that even image-based models capture chemical and biological knowledge in their representations (Alampara et al., 2025; Naghdloo et al., 2025).

In summary, it is not well understood whether CLMs learn molecular substructures well and if this follows any chemical theory. With this work, we make a first attempt to address this gap by conducting systematic probing experiments with CLMs across 78 molecular substructures.

3 Probing Dataset Creation

Our goal is to curate a probing dataset that captures a wide range of molecular substructures and at the same time allows us to conduct meaningful analysis. Each probing task is formulated as a binary classification problem predicting the presence or absence of a molecular substructure (e.g., functional group, ring, etc.) in a molecule. The probing dataset is derived from PCQM4Mv2, a publicly available dataset designed for predicting the HOMO-LUMO energy gap (Hu et al., 2021).

Preprocessing We first discard all molecules with invalid SMILES strings or those that lead to processing errors (cf. Appendix C.1). The remaining molecules—represented as SMILES strings—are then canonicalized and annotated with binary labels. Preprocessing results in an initial set of 101 unique molecular substructures. All processing except for binarization is performed using RDKit (Landrum et al., 2024).

Data sampling and cleaning The PCQM4Mv2 dataset comprises ≈ 3.7 million molecules—too many to conduct extensive probing experiments. Hence, we create three subsets by uniformly randomly sampling 100k and 20k instances from the preprocessed train and validation splits of PCQM4Mv2, respectively. Note that sampling from the original splits prevents any leakage between train and test sets. Finally, we discard probing tasks for molecular substructures that appear fewer than 200 times in either the train or test set.

Our final probing dataset comprises a diverse set of 78 molecular substructures—from functional groups such as amides or phenols to different types of ring structures (cf. Appendix C.2).

4 Experimental Setup

We investigate our research questions across eight pre-trained CLMs, which we first probe for the presence of molecular substructures, comparing them against their randomly initialized counterparts. We then study the effects of fine-tuning on two well-studied MPP tasks (lipophilicity and solubility prediction). This allows us to compare the

changes of a model's molecular substructure representation against existing chemical knowledge. An overview of the experimental setup is provided in Figure 1.

Pre-trained CLMs We focus on models pre-trained on small molecules, particularly encoder-only CLMs trained with the masked language modeling (MLM) objective. These models can consider both left and right context, in contrast to decoder-only models trained on causal language modeling.
Chemberta Chithrananda et al. (2020) release multiple six-layer models based on RoBERTa (Liu et al., 2019b). We use both publicly available models, `chemberta-base` and `chemberta`.
Chemberta-2 In subsequent work, Ahmad et al. (2022) release models pre-trained on different numbers of molecules (`chemberta-2-5M`, `chemberta-2-10M`, and `chemberta-2-77M`).
Chemberta-3 Most recently, Singh et al. (2026) released their training framework along with a 12-layer model trained on 100M molecules (`chemberta-3`).
Molformer Ross et al. (2022) train a model with 12 layers using linear attention and rotary positional embeddings. We use the publicly available model trained on 100M molecules (`molformer`).
Roberta-zinc-480m Heyer (2023) release a 14-layer RoBERTa-based model trained on 480M molecules (`roberta-zinc-480m`).

All models are trained on molecules sampled from either ZINC15 (Sterling and Irwin, 2015) or ZINC20 (Irwin et al., 2020). `molformer` is additionally trained on PubChem (Kim et al., 2018). A more detailed description of all models is provided in Appendix B.

Probing setup For each probing task, we train a linear classifier on the CLS token representation of each encoder layer (cf. Appendix D.1). We evaluate probing performance using the macro-averaged F1 score to account for class imbalances in the test splits. We compare each pre-trained model (PT) against its randomly initialized (RI) counterpart (note, `chemberta-2` has one shared RI model) and a majority class prediction baseline (maj). We further downsample instances in the training set to account for class imbalances. Preliminary experiments show that this substantially improves probing performance. For the remainder of the paper, all reported results refer to the performance on the downsampled probing tasks. All dataset statistics are provided in Appendix C.3.

5 The Effect of Pre-Training (RQ1)

We study the effect of pre-training (RQ1) using our probing dataset (§3) and analyzing the results with increasing levels of granularity. We first compare the probing performance of pre-trained (PT) and randomly initialized (RI) models (§5.1), then with respect to different ring structures (§5.2), and, finally, for individual molecular substructures (§5.3).

5.1 Probing Results

Figure 2 shows the macro-averaged F1 score of all eight PT and six RI models averaged across all 78 probing tasks and three datasetes each. In addition, we show the average performance of the majority class prediction baseline (maj). On average, models **learn to better encode molecular substructures during pre-training**: most pre-trained models (x) exhibit substantial improvements in performance relative to their randomly initialized counterparts (●) and the majority classifier (–). We also observe that for most PT models (excluding `chemberta` and `chemberta-base`), the probing performance is **higher in the upper layers** (0.5–1.0). In particular, we see that `chemberta-2-10M`, `molformer` and `chemberta-2-5M` benefit the most from pre-training. In contrast, we observe negligible improvement for `chemberta` or even small drops for `chemberta-base`. Most notably, we find that in the most recent model (`chemberta-3`), probing performance deteriorates substantially in the lower layers. We further investigate this phenomenon in Appendix I by further pre-training the model on different datasets.

5.2 Ring Structures

We further analyze the representations of rings and other molecular substructures, finding that the representations of both RI (●) and PT (x) models perform exceptionally well at identifying ring structures compared to all other groups (● and x). Moreover, RI models **encode ring structures well already at the first encoder layer**, suggesting that these surface-level patterns are easy for the models to extract directly from the input, even without pre-training. We also find that the benefit of pre-training diminishes for ring structures compared to that of other molecular substructures (i.e., the gap between x and ● is much smaller compared to the gap between x and ●).

Different types of rings A closer analysis of different types of ring structures (i.e., aliphatic, aro-

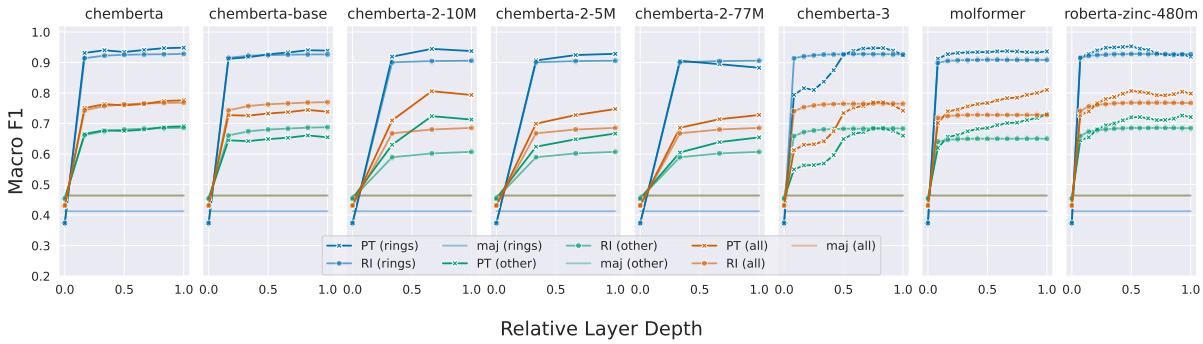


Figure 2: Average probing performance (macro-averaged F1 score ↑) of pre-trained (PT), randomly initialized (RI), and the majority class prediction (maj) models on molecular substructures. We report average performance on 12 ring types (avg rings), all other 66 substructures (avg other) and all 78 substructures (avg all).

matic, and saturated rings) reveals that particularly aromatic and aliphatic are already well encoded in random initializations, with pre-training slightly reducing probing performance in the upper layers (cf. Figure 6) for all models except for `chemberta-3` (see Appendix I). The high performance on aromatic rings might stem from a distinct surface-level pattern. When SMILES strings are canonicalized, atoms in aromatic rings (see Figure 5 for an example) are represented with lowercase letters. This contrast to other molecular substructures (which consist of upper-cased atoms) results in a strong signal for the model.

5.3 Individual Molecular Substructures

Finally, we investigate if there are molecular substructures that undergo changes consistently across all models. For visualization, we focus on the model with the most pronounced changes (`molformer`) and provide the rest in Appendix E. We observe three patterns shown in Figure 3.

PT > RI First, we find that **pre-training generally leads to a better encoding of most molecular substructures in upper layers**, as reflected by the higher density of the red shade in the middle and upper layers. In particular, all models exhibit substantial improvement on *carboxylic acids* (`COO`, `COO2`, `Al_COO`, `Ar_COO`), *aromatic hydroxy groups* (`Ar_OH`), *phenol groups* (`phenol`, `phenol_nonorthobound`), *amides*, and *ketones* (`ketone`, `ketone_Topliss`). Furthermore, all models improve on *carbonyls* (`C_O_noCOO`), *aldehyde* and *imide*, although to a lesser degree.

PT < RI Second, **some molecular substructures consistently exhibit lower performance after pre-training**. In particular, probing performance decreases for aromatic nitrogens (`Ar_N`) and certain *heterocycles* such as `thiazole`, `thiophene` and `furan` across all models. Likewise, some ring substructures such as `AromaticHeterocycles`, `AromaticCarbocycles` and `benzene` show slight performance degradation in upper layers. The decrease in performance on `AromaticHeterocycles` may potentially reflect the substantial drops in performance of `thiazole`, `thiophene` and `furan`—all aromatic heterocycles. Finally, most models (except for `chemberta-2-5M/10M`) show evidence of unlearning `halogens` in upper layers while preserving a better encoding in the lower layers.

PT ≈ RI Third, we observe that only a handful molecular substructures undergo small amounts of change ($\pm 1\%$), resulting in an almost uniform distribution of information across layers. However, this behavior is **not consistent across models**.

5.4 Discussion

Overall, our results suggest that **pre-training improves the molecular structure awareness** of CLMs, considerably changing the encoding of many molecular substructures. We further observe that **RI models already encode aromatic and aliphatic rings very well** performing on-par with or better than PT models. Interestingly, unlearning of molecular substructures largely varies between models, however, a **few molecular substructures are consistently unlearned during pre-training**. We conjecture this might stem from a disparity in the pre-training data and conduct further pre-training experiments for five molecular substructures and across different models (§7). Finally, we find that **all molecular substructures exhibit a change larger than $\pm 1\%$ in at least one model**.

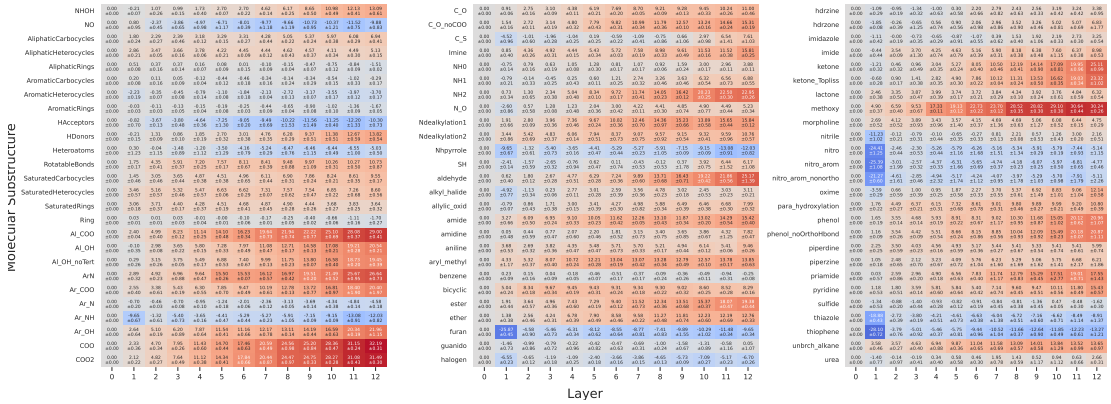


Figure 3: Relative difference in probing performance (% macro-averaged F1) between RI and PT `molformer` on 78 molecular substructures (cf. Table 3). Each cell denotes the relative difference for a specific probing task (y-axis) and layer (x-axis). Red indicates an increase in performance after pretraining, while blue denotes a decrease.

6 The Effect of Fine-Tuning (RQ2)

Our probing experiments have shown how pretraining reconfigures the information encoded in representations and that RI models already encode ring structures well. Next, we investigate changes of RI and PT models during fine-tuning on two well-studied tasks in chemistry, which allows us to contextualize our findings within chemical theory.

6.1 Chemistry Background

We focus on predicting the lipophilicity and aqueous solubility (in short, solubility) of molecules. Here, we provide brief task descriptions and introduce important molecular substructures that affect the lipophilicity and solubility of a molecule; and refer to Appendix A for more details.

Lipophilicity Lipophilicity refers to the ability of a chemical compound to dissolve in fat-like solvents (lipids, fats, oils; Morak-Młodawska et al. 2023). It is an important physicochemical property of molecules which correlates with the (oral) absorption, (tissue) distribution, metabolism, excretion, and toxicity (ADMET) properties of drugs (Mannhold et al., 2009), essential in determining how a candidate drug will interact with the human body (Waring, 2009). The goal of lipophilicity prediction is to estimate the octanol/water distribution coefficient (logD) of a specific molecule.

Aqueous solubility Aqueous solubility refers to the ability of a molecule to dissolve in water. For drug development, predicting the solubility of a molecule is equally important as predicting the lipophilicity as it also affects their biovailability and ADMET profiles (Llompart et al., 2024; Klopman et al., 1992). The goal of solubility prediction is to estimate the log solubility (logS) of a specific molecule in water. While solubility is closely related to lipophilicity, it is also dependent on other factors such as the melting point of a molecule (Hill and Young, 2010).

Important molecular substructures Chemical literature distinguishes between two groups of molecular substructures that are known to affect lipophilicity and solubility (Harrold and Zavod, 2023). First, *hydrophilic* substructures such as carboxylic acids substantially decrease a molecule's lipophilicity while increasing its solubility. Second, *lipophilic* substructures such as aromatic rings increase a molecule's lipophilicity while decreasing its solubility. We follow this classification of molecular substructures in our analysis and put all other molecular substructures that do not substantially affect lipophilicity into a third group (*other*). We provide a list of all molecular substructures along with their group in Appendix F.

6.2 Experimental Setup

For fine-tuning, we replace the classification head of the CLM with either a linear regression layer or a two-layer MLP and minimize the mean squared error loss. Following Wu et al. (2018), we use the root mean squared error (RMSE) as our evaluation metric for both tasks. Since all models were pre-trained on canonicalized SMILES strings, we canonicalize the input SMILES accordingly.

Dataset Both datasets are sampled from the MoleculeNet benchmark (Wu et al., 2018) and consist of 4,200 (lipophilicity) and 1,127 (solubility) molecules. We use the train–validation–test splits

(80/10/10) provided by Ross et al. (2022). Detailed dataset statistics and analysis for both tasks are provided in Appendix F.

Hyperparameters We perform hyperparameter tuning separately for both tasks, considering different batch sizes and learning rates. All pre-trained and randomly initialized models are trained for up to 10–20 epochs. We deploy early stopping with a patience of 2 and use AdamW as our optimizer. We report all hyperparameters in Appendix F.

Baselines As baselines, we evaluate multiple traditionally used models, namely, linear regression models (LR), support vector machines (SVM), and gradient boosted trees (XGB). For each model, we evaluate four algorithms to extract molecule representation vectors, also known as fingerprints, provided by RDKit (Landrum et al., 2024). Finally, we evaluate two large language models (LLMs): `Llama-3.2-3B-Instruct` (Grattafiori et al., 2024) and `gpt-oss-20B` (OpenAI, 2025) with additional chemical knowledge that is important for the respective downstream task. We provide detailed hyperparameters and experimental results for all baselines in appendices J.1 and J.2.

Probing dataset adjustment In order to conduct meaningful analyses, we accommodate changes to the probing dataset introduced in §3 that consider dataset-specific properties of the respective downstream task. More specifically, we discard any molecular substructure which appears fewer than ten times in either the training or test split of the task-specific dataset; effectively removing outliers from our analysis. This results in probing 60 and 39 molecular substructures for lipophilicity and solubility prediction, respectively.

6.3 Downstream Task Results

Table 1 shows the results of all randomly initialized (RI) and pre-trained models (PT) as well as the best performing model using fingerprints (SVM) and LLM (`gpt-oss-20b`) for both downstream tasks. We further include the results of the graph-based models (^{mol}) that were reported by Ross et al. (2022) who use the same data splits. Overall, we observe that PT models consistently outperform RI ones on both tasks with `molformer` consistently performing best, highlighting the benefit of pre-training CLMs. We further find that the `chemberta-2` models, differing only in the pre-training datasets, exhibit differences of 0.073 RMSE on lipophilicity (0.046 on solubility), suggesting that pre-training data plays a major role for downstream task performance. Moreover, the `chemberta-2-77M` model is often outperformed by its smaller counterparts. This indicates that data quality may play a more important role than data quantity. Finally, consistent with prior findings, we observe that fingerprint-based models perform rather well (Dias et al., 2023; Xia et al., 2023); and that LLMs perform even worse than the mean predictor (Zhao et al., 2023).

Model	Lipo		ESOL	
	RI	PT	RI	PT
GC^{mol}	-	0.655	-	0.970
A-FP^{mol}	-	0.578	-	0.503
MPNN^{mol}	-	0.719	-	0.580
`mean predictor`	-	1.013	-	2.057
SVM$_{(c=16)}$ + ATFP	-	0.640	-	0.830
`gpt-oss-20b`	-	2.532	-	8.964
`molformer`	0.832	**0.565**	0.808	**0.587**
`roberta-zinc-480m`	0.788	0.580	0.878	0.746
`chemberta-base`	0.785	0.663	0.832	0.739
`chemberta`	0.779	0.675	0.822	0.693
`chemberta-2-5M`	0.850	0.664	0.872	0.682
`chemberta-2-10M`	"	0.591	"	0.724
`chemberta-2-77M`	"	0.632	"	0.728
`chemberta-3`	1.026	0.637	0.960	0.757

Table 1: Test performance (RMSE, ↓) for lipophilicity (Lipo) and solubility (ESOL) prediction (Wu et al., 2018). Besides all CLMs, we also include results of the best performing fingerprint-based model (SVM) and LLM (`gpt-oss-20b`). ^{mol} denotes results of graph-based models reported by Ross et al. (2022).

6.4 Probing Results

We conduct probing experiments similar to §5 but with the difference that we now compare the layer-wise representations of molecular substructures in a model before and after fine-tuning. This is done for both RI and PT models to understand potential differences in their behavior during fine-tuning. In our analysis, we first inspect lipophilic and hydrophilic molecular substructures and then inspect individual molecular substructures. Due to a lack of space, we focus our analysis in the main paper on lipophilicity prediction and provide the results and analysis for solubility prediction in Appendix G.

Group analysis Figure 4 shows heatmaps for eight pre-trained (left) and six randomly initialized (right) models split into hydrophilic (top), lipophilic (middle), and other (bottom) groups. Similar to §5, red indicates an increase in prob-

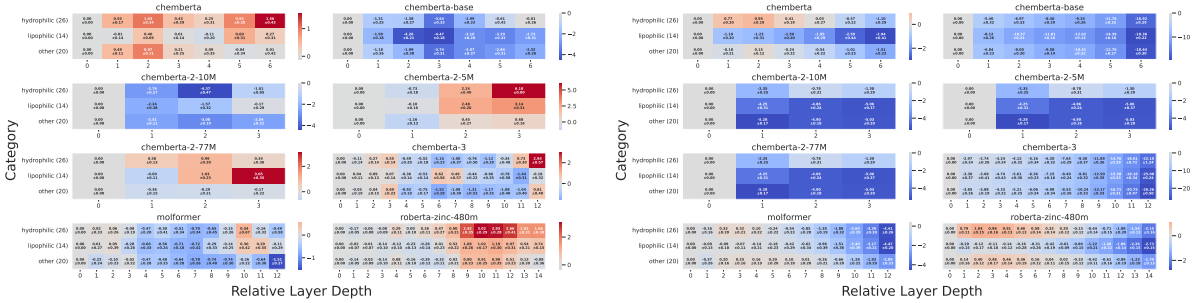


Figure 4: Average relative differences in probing performance (% macro-averaged F1) in probing performance of PT (left) and RI (right) models after fine-tuning on lipophilicity. We group into *hydrophilic* (top), *lipophilic* (middle), and *other* (bottom) groups (see Table 6), with numbers indicating the group size. The RI results for `chemberta-2` are based on a single model (hence, are the same for all three variants) as all models use the same architecture.

ing performance while blue indicates a decrease. We observe that the groups that are important for lipophilicity prediction (*hydrophilic* and *lipophilic*) undergo larger changes than the *other* group. This indicates that **molecular substructure learning follows chemical theory**. Interestingly, we find that **fine-tuning has a noticeably smaller effect than pre-training** on the molecular substructure representations; and that the changes are mostly **concentrated in upper layers** which corroborates prior observations in the NLP literature (Mosbach et al., 2020; Merchant et al., 2020; Fayyaz et al., 2021). In contrast, **RI models behave differently**, as fine-tuning appears to mostly negatively affect the encoding of substructures in upper layers.

Individual analysis A detailed analysis of individual molecular substructures reveals that the effect of fine-tuning varies across models. Furthermore, even among molecular substructures of the same group (lipophilic, hydrophilic, other), the magnitude may vary (we provide detailed heatmaps in Appendix G). Nevertheless, there are multiple substructures for which probing performance increases after fine-tuning on lipophilicity prediction. These are primarily *carboxylic acids* (`COO`, `COO2`, `Al_COO`, and `Ar_COO`[4]). This is **consistent with observations made by chemists suggesting that carboxylic acids contribute most negatively to the logD value** and are therefore highly indicative (Landry and Crawford, 2020). Interestingly, `chemberta-2` models consistently improve upon *halogens* after fine-tuning (we study this closer in §7). Again, we do not observe any consistent trends for RI models, except for a degradation of `AromaticHeterocycles` and `Ar_NH`.

[4]With the exception of `chemberta2-77M`.

6.5 Discussion

Our probing experiments on models before and after fine-tuning on lipophilicity (§6.4) and solubility (Appendix G) prediction reveal three major findings. First, **molecular substructures that are theoretically more relevant for a downstream task undergo larger changes during fine-tuning**. Reciprocally, molecular substructures that undergo major changes during fine-tuning for a specific downstream task might indicate a high importance. This might be especially interesting for tasks with a high variability in terms of important molecular substructures such as toxicity prediction. Second, **changes are less pronounced compared to pre-training and occur more frequently in the upper layers**. Considering the increasing model sizes and consequently, the increasing costs of probing, one way to reduce costs could be to restrict probing to the upper layers as they yield the largest changes. Third, **pre-training increases the robustness of molecular substructures in CLMs**, making them more likely to be retained during fine-tuning.

7 Practical Implications

The varying probing performance across different models (especially for the `chemberta-2` models with different pre-training data sizes, cf. §5) suggests that this might be attributed to a lack of molecules containing a specific molecular substructure during pre-training. To better understand the impact of pre-training data, we further pre-train the models. In particular, we curate pre-training datasets consisting of molecules with molecular substructures on which a model underperforms. Our results show that further pre-training on these molecules does indeed improve a model's internal representation of them (Appendix H). Moreover,

we find that models that share the same architecture converge towards the same upper bound shape in terms of probing performance (Figure 15).

Similarly, we conduct experiments for the `chemberta-3` model which exhibits a strange dome in the upper layers. Further pre-training the model on two different datasets indicates that, again, this dome becomes less pronounced as probing performance also increases in the lower layers (cf. Figure 17). Most notably, we find that the pre-training dataset can make a substantial difference on the resulting probing performance.

These experiments provide an exploratory demonstration of how probing may be used to identify signs of undertraining for specific molecular substructures. However, our limited-scale experiments also show that improvements in the representations of individual substructures do not necessarily translate into improved predictions on lipophilicity (cf. Tables 12, 14 and 15).

8 Conclusion

Our results provide a systematic analysis of the extent to which chemical language models (CLMs) trained on linearized molecular representations encode important molecular substructures. Across eight pre-trained and six randomly initialized models, we find that pre-training generally improves molecular structure awareness, with the most pronounced effects emerging in the upper layers. Although certain molecular substructures are unlearned during pre-training, only a small subset exhibits this behavior consistently across all models. Notably, we find that ring structures are linearly decodable from representations of randomly initialized models, suggesting that these are surface-level properties of the input representations and do not require pre-training. Our fine-tuning analysis reveals that, on average, groups of molecular substructures relevant to lipophilicity and solubility prediction undergo larger representational changes than others. Extending the analysis to other molecular property prediction tasks, such as toxicity, bioactivity, or permeability, represents an important direction for future work.

Finally, our further pre-training experiments showcase how probing may be used to identify potential signs of undertraining and mitigate the resulting representational gaps using a small and carefully curated dataset.

9 Limitations

True complexity of lipophilicity and solubility prediction Although Hansch et al. (1977) estimate the lipophilicity of a molecule using a linear combination of its substructures, we note that this approach is an oversimplification of the underlying process. The actual lipophilicity of a molecule also depends on the environment a certain substructure is in (i.e., other neighboring substructures) as well as its depth and position in the molecule and is subject to further research in chemistry.

While the logD value (measuring the lipophilicity) is part of the general solubility equation (corrected for ionization at pH 7.4), there are various other factors that influence the logS value (Hill and Young, 2010). Some of these challenges are part of ongoing chemistry research as highlighted by Llompart et al. (2024), who find that accurately predicting aqueous solubility requires the knowledge of many factors such as the solid-solvated phase transition, solid state, temperature, polymorphism, intermolecular interactions between solute-solvent etc.

Other downstream tasks Our experiments focus on lipophilicity and solubility prediction as the downstream tasks. While there exist other tasks such as predicting the bioactivity and toxicity of molecules, one limiting factors is the number of publicly available molecules that have been studied with the same experimental conditions. Moreover, many tasks are still subject to ongoing chemical research and are not understood well (yet). For instance, a non-trivial challenge in predicting the bioactivity are activity cliffs (Stumpfe et al., 2019), pairs of molecules with highly similar structures—i.e., close proximity in the molecular "landscape"—but different magnitudes in terms of bioactivity, resulting in a steep "cliff". A better chemical understanding of the underlying process would allow researchers to build models that capture fine-grained structural differences between molecules.

Effects of random- and downsampling We note that due to the random sampling of the three probing datasets and the downsampling for individual molecular substructures, the molecules across different splits may vary with each molecular substructure dataset corresponding to a distinct subset of the original 100k molecules, both in terms of size and composition.

Limitations of probing Probing remains a debated diagnostic method as it does not indicate whether a feature is used during prediction, but only how extractable the property is from the learned representations. Using linear probes (as in this work), we can therefore only assess the linear separability of the investigated property in these representations. Note, that there is no consensus on what probing classifier to use. While some works argue that utilizing linear classifiers prevents the possibility of memorization (Belinkov, 2022), others propose to instead consider memorization during evaluation (Pimentel et al., 2020).

More complex probing task Our probing tasks focus on detecting the presence of a molecular substructure, rather than counting occurrences or identifying its location. Consequently, some probing tasks may be easier to learn due to the underlying nature of a molecule. For instance, a molecule may contain multiple occurrences of a single molecular substructure, making it easier to detect its presence (as is the case for aromatic rings that often occur multiple times in many molecules). More challenging tasks could offer additional insights but are subject to future investigation; and moreover, also require respective chemistry research.

10 Impact Statement

The primary goal of this work is to provide insights on how chemical language models are affected by pre-training and fine-tuning on chemical data. While this work falls under the category of fundamental research without direct implications on downstream applications, the authors acknowledge that some of the findings (e.g., fine-tuning mostly aligns with chemical theory) may lead to research which could be abused to identify molecular substructures that are harmful to the human body. The authors emphasize that the experiments and tasks presented here have no direct connection to harmful applications (including tasks such as predicting the toxicity of molecules).

Acknowledgments

We thank Shubham Dokania, Max Rausch-Dupont and Afnan Sultan for their helpful discussions and feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *Preprint*, arXiv:2209.01712.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, and Aleksandar Bojchevski. 2023. Probing graph representations. In *International Conference on Artificial Intelligence and Statistics*, pages 11630–11649. PMLR.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, 5(10):952–961.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076. PMID: 34694798.

Hartmut Beck, Michael Härter, Bastian Haß, Carsten Schmeck, and Lars Baerfacker. 2022. Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the bayer chemical research laboratory. *Drug Discovery Today*, 27(6):1560–1574.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. 2019. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108. PMID: 30887799.

Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. In *Machine Learning for Molecules Workshop (NeurIPS 2020)*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

William Curatolo. 1998. Physical chemical properties of oral drug candidates in the discovery and

exploratory development settings. *Pharmaceutical Science & Technology Today*, 1(9):387–393.

Ana Laura Dias, Latimah Bustillo, and Tiago Rodrigues. 2023. Limitations of representation learning in small molecule property prediction. *Nature Communications*, 14:6394.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. In *Machine Learning for Molecules Workshop (NeurIPS 2020)*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, pages 1–31.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Inken Fender, Jannik Adrian Gut, and Thomas Lemmin. 2025. Beyond performance: how design choices shape chemical language models. *Journal of Cheminformatics*, 17(1):1–15.

David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. 2018. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3(3):442–452.

Veronika Ganeeva, Kuzma Khrabrov, Artur Kadurin, and Elena Tutubalina. 2025. Two steps from hell: Compositionality on chemical LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1042–1049, Suzhou, China. Association for Computational Linguistics.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. 2024. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. In *Advances in Neural Information Processing Systems*, volume 37, pages 134721–134746. Curran Associates, Inc.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Corwin Hansch, Sharon D Rockwell, Priscilla YC Jow, Albert Leo, and Edward E Steller. 1977. Substituent constants for correlation analysis. *Journal of medicinal chemistry*, 20(2):304–306.

Marc W. Harrold and Robin M. Zavod. 2023. *Basic concepts in medicinal chemistry*, 3rd edition. American Society of Health-System Pharmacists, Bethesda, MD.

Jiazhen He, Eva Nittinger, Christian Tyrchan, Werngard Czechtizky, Atanas Patronov, Esben Jannik Bjerrum, and Ola Engkvist. 2022. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of Cheminformatics*, 14(1):18.

Jiazhen He, Huifang You, Emil Sandström, Eva Nittinger, Esben Jannik Bjerrum, Christian Tyrchan, Werngard Czechtizky, and Ola Engkvist. 2021. Molecular optimization by capturing chemist's intuition using deep neural networks. *Journal of Cheminformatics*, 13(1):26.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Karl Heyer. 2023. Roberta-zinc-480m.

Alan P. Hill and Robert J. Young. 2010. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discovery Today*, 15(15):648–655.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*, pages 1–16.

Yifan Hou and Mrinmaya Sachan. 2021. Bird's eye: Probing for linguistic graph structures with a simple information-theoretic approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1844–1859, Online. Association for Computational Linguistics.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. 2021. OGB-LSC: A large-scale challenge for machine learning on graphs. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. 2020. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073. PMID: 33118813.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):1–13.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2025. Structural reasoning improves molecular understanding of LLM. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21016–21036, Vienna, Austria. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Chen-Yang Jia, Jing-Yi Li, Ge-Fei Hao, and Guang-Fu Yang. 2020. A drug-likeness toolbox facilitates admet study in drug discovery. *Drug Discovery Today*, 25(1):248–258.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. 2018. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109.

Gilles Klopman, Shaomeng Wang, and D. M. Balthasar. 1992. Estimation of aqueous solubility of organic molecules by the group contribution approach. application to the study of biodegradation. *Journal of Chemical Information and Computer Sciences*, 32(5):474–482. PMID: 1400663.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, and 11 others. 2024. Rdkit: Open-source cheminformatics.

Matthew L. Landry and James J. Crawford. 2020. Logd contributions of substituents commonly used in medicinal chemistry. *ACS Medicinal Chemistry Letters*, 11(1):72–76.

Miguelangel Leon, Yuriy Perezhohin, Fernando Peres, Aleš Popovič, and Mauro Castelli. 2024. Comparing smiles and selfies tokenization for enhanced chemical language modeling. *Scientific Reports*, 14.

Juncai Li and Xiaofei Jiang. 2021. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021(1):1–7.

Xuan Lin, Long Chen, and Yile Wang. 2026. Attrilensmol: Attribute guided reinforcement learning for molecular property prediction with large language models. *Preprint*, arXiv:2508.04748.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Shikun Liu, Deyu Zou, Nima Shoghi, Victor Fung, Kai Liu, and Pan Li. 2026. Roft-mol: Benchmarking robust fine-tuning with molecular graph foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 1–39.

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

Yunwu Liu, Ruisheng Zhang, Tongfeng Li, Jing Jiang, Jun Ma, and Ping Wang. 2023a. Molrope-bert: An enhanced molecular representation with rotary position embedding for molecular property prediction. *Journal of Molecular Graphics and Modelling*, 118:108344.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Pol Llompart, Christian Minoletti, Sapark Baybekov, Dragos Horvath, Gilles Marcou, and Alexandre Varnek. 2024. Will we ever be able to accurately predict solubility? *Scientific Data*, 11(1):303.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387. PMID: 35266390.

Gerald M Maggiora. 2006. On outliers and activity cliffs why qsar often disappoints. *Journal of chemical information and modeling*, 46(4):1535–1535.

Favour Danladi Makurvet. 2021. Biologics vs. small molecules: Drug costs and patient access. *Medicine in Drug Discovery*, 9:100075.

Raimund Mannhold, Gennadiy I. Poda, Claude Ostermann, and Igor V. Tetko. 2009. Calculation of molecular lipophilicity: State-of-the-art and comparison of logp methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, 98(3):861–893.

Eyal Mazuz, Guy Shtar, Bracha Shapira, and Lior Rokach. 2023. Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports*, 13(8799).

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Beata Morak-Młodawska, Małgorzata Jeleń, Emilia Martula, and Rafał Korlacki. 2023. Study of lipophilicity and adme properties of 1,9-diazaphenothiazines with anticancer action. *International Journal of Molecular Sciences*, 24(8).

H. L. Morgan. 1965. 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.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Amin Naghdloo, Dean Tessone, Rajiv M Nagaraju, Brian Zhang, Jeffrey Kang, Shouyi Li, Assad Oberai, James B Hicks, and Peter Kuhn. 2025. Representation learning enables robust single cell phenotyping in whole slide liquid biopsy imaging. *Scientific Reports*, 15(1):36589.

Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. 1987. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Jun-Hyung Park, Yeachan Kim, Mingyu Lee, Hyuntae Park, and SangKeun Lee. 2024a. MolTRES: Improving chemical language representation learning for molecular property prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14241–14254, Miami, Florida, USA. Association for Computational Linguistics.

Jun-Hyung Park, Hyuntae Park, Yeachan Kim, Woosang Lim, and SangKeun Lee. 2024b. Moleco: Molecular contrastive learning with chemical language models for molecular property prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–420, Miami, Florida, US. Association for Computational Linguistics.

Josh Payne, Mario Srouji, Dian Ang Yap, and Vineet Kosaraju. 2020. Bert learns (and teaches) chemistry. *Preprint*, arXiv:2007.16012.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4609–4622, Online. Association for Computational Linguistics.

Douglas E. V. Pires, Tom L. Blundell, and David B. Ascher. 2015. pkcsm: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of Medicinal Chemistry*, 58(9):4066–4072. PMID: 25860834.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Shaghayegh Sadeghi, Ali Forooghi, Jianguo Lu, and Alioune Ngom. 2024. Moleval: An evaluation toolkit for molecular embeddings via LLMs. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. 2019. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583. PMID: 31572784.

Jie Shen and Christos A Nicolaou. 2019. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, Derek Jones, Kevin McLoughlin, Hyojin Kim, Anoushka Bhutani, Shreyas Vinaya Sathyanarayana, Venkat Viswanathan, Jonathan E. Allen, and Bharath Ramsundar. 2026. Chemberta-3: an open source training framework for chemical foundation models. *Digital Discovery*, 5:662–685.

Teague Sterling and John J. Irwin. 2015. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Dagmar Stumpfe, Huabin Hu, and Jurgen Bajorath. 2019. Evolving concept of activity cliffs. *ACS omega*, 4(11):14360–14368.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. 2024. Real-world molecular out-of-distribution: Specification and investigation. *Journal of Chemical Information and Modeling*, 64(3):697–711. PMID: 38300258.

Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. 2022. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958.

Michael A. Walker. 2017. Improvement in aqueous solubility achieved via small molecular changes. *Bioorganic & Medicinal Chemistry Letters*, 27(23):5100–5108.

Hanchen Wang*, Jean Kaddour*, Shengchao Liu, Jian Tang, Joan Lasenby, and Qi Liu. 2023. Evaluating self-supervised learning for molecular graph embeddings. In *NeurIPS 2023, Datasets and Benchmarks Track*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. 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*, BCB '19, page 429–436, New York, NY, USA. Association for Computing Machinery.

Ye Wang, Honggang Zhao, Simone Sciabola, and Wenlu Wang. 2023. cmolgpt: A conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules*, 28(11):4430.

Michael J. Waring. 2009. Defining optimum lipophilicity and molecular weight ranges for drug candidates—molecular weight dependent lower logd limits based on permeability. *Bioorganic & Medicinal Chemistry Letters*, 19(10):2844–2851.

Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alex Weir. 2015. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475–486.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

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

Zhenxing Wu, Dejun Jiang, Jike Wang, Xujun Zhang, Hongyan Du, Lurong Pan, Chang-Yu Hsieh, Dongsheng Cao, and Tingjun Hou. 2022. Knowledge-based bert: a method to extract molecular features like computational chemists. *Briefings in Bioinformatics*, 23(3):1–13.

Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z. Li. 2023. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. MolRAG: Unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531, Vienna, Austria. Association for Computational Linguistics.

Dongyu Xue, Han Zhang, Xiaohan Chen, Dongling Xiao, Yukang Gong, Guohui Chuai, Yu Sun, Hao Tian, Hua Wu, Yukun Li, and Qi Liu. 2022. X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *Science Bulletin*, 67(9):899–902.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2025. Drugassist: a large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):1–12.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. In *Advances in Neural Information Processing Systems*, pages 1–11.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. 2023. Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4(2):025035.

Lawrence Zhao, Carl Edwards, and Heng Ji. 2023. What a scientific language model knows and doesn't know about chemistry. In *NeurIPS 2023 AI for Science Workshop*.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

A Chemical Background for Molecular Modeling

A.1 Drug Development

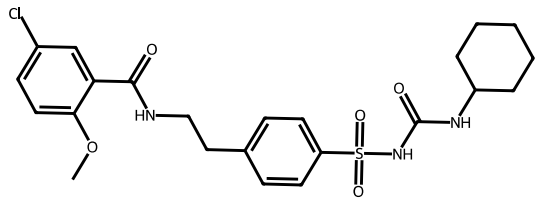
Drug development aims at identifying molecules that are both effective against a target disease and can be safely administered to humans (Jia et al., 2020). One crucial bottleneck in drug development is synthesizing potential molecules in the laboratory, which is a time- and cost-intensive process. Machine learning offers the potential to accelerate this process by improving efficiency and reducing costs: for example, by prioritizing candidate compounds with desirable properties, thereby reducing the need to synthesize nonviable molecules. Consequently, *molecular property prediction*—the estimation of physicochemical, biological and functional properties such as lipophilicity, toxicity, permeability, and reactivity—constitutes a core component of this process (Waring et al., 2015; Pires et al., 2015). In particular, *ADMET* properties, which characterize the *drug-likeness* of a compound, are essential for assessing how a drug candidate will interact with the human body. In this work, we focus on small molecules (those with molecular weight ≤ 1000 Da), which possess better ADMET profiles (Beck et al., 2022) and constitute most approved pharmaceuticals (Makurvet, 2021).

Molecular property prediction Typically, molecular property prediction encompasses a wide range of tasks such as predicting bioactivity, solubility, permeability, and toxicity which are often addressed by different models. It is important to note that many of these tasks are not classification, but regression tasks. Reliably predicting the magnitude of even a single property can be challenging, as the arrangement and composition of molecular substructures play an important role. In many cases, models must capture fine-grained structural differences between molecules that lead to considerably different behavior. One such example is activity cliffs (Stumpfe et al., 2019), which describe pairs of molecules with highly similar structures—i.e., close proximity in the molecular "landscape"—but different magnitudes in terms of bioactivity, resulting in a steep "cliff". Activity cliffs have been subject to chemistry research for over a decade, with evolving insights on what molecular substructures constitute them (Maggiora, 2006; Stumpfe et al., 2019; Xia et al., 2023).

A.2 Linearized Representations

In order to process molecules in models or handcrafted algorithms, works have devised various methods. One such method called SMILES proposes to represent molecules as linearized represen-

tation of molecular graphs (Weininger, 1988). The standard SMILES encoding is non-unique, resulting in a one-to-many mapping between a molecule and its possible SMILES representations. However, canonicalization of SMILES enforces a deterministic graph traversal order, resulting in a one-to-one mapping between molecular graphs and SMILES strings. SMILES encode some graph structural information explicitly. For example, double bonds are represented by =, triple bonds by #. Single bonds, on the other hand, are not explicitly specified. Thus, consecutive atoms are assumed to be connected by either a single or an aromatic bond. Furthermore, as shown in Figure 5, when SMILES are canonicalized, atoms in aromatic rings (i.e., rings with alternating single and double bonds) are represented with lowercase letters.



COc1ccc(Cl)cc1C(=O)NCCc2ccc(S(=O)(=O)NC(=O)NC3CCCCC3)cc2

Figure 5: An example molecule: Glibenclamide and its SMILES representation. The two aromatic rings are highlighted in green and orange. Compared to the non-aromatic ring (purple), all atoms in the aromatic rings are lowercased.

A.3 Lipophilicity Prediction

Lipophilicity refers to the ability of a chemical compound to dissolve in fat-like solvents (lipids, fats, oils; Morak-Młodawska et al. 2023). It is an important physicochemical property of molecules which correlates with the (oral) absorption, (tissue) distribution, metabolism, excretion, and toxcicity (ADMET) properties of drugs (Mannhold et al., 2009), essential in determining how a candidate drug will interact with the human body (Waring, 2009). Compared to other MPP tasks such as toxicity prediction, the impact of specific molecular substructures is better understood for lipophilicity prediction. For instance, Landry and Crawford (2020) study how specific molecular substructures affect the lipophilicity of a molecule.

The goal of lipophilicity prediction is to estimate the octanol/water distribution coefficient (i.e., logD at pH 7.4) of a specific molecule. Chemical literature distinguishes between two groups of molecular substructures that are known to affect lipophilicity (Harrold and Zavod, 2023). In particular, *hydrophilic* substructures such as carboxylic acids substantially decrease a molecule's logD value while *lipophilic* substructures such as aromatic rings increase the logD value. We follow this classification of molecular substructures in our analysis and put all other molecular substructures that do not substantially affect lipophilicity into the *other* group (Appendix F provides a list of all molecular substructures).

A.4 Aqueous solubility

Aqueous solubility refers to the ability of a molecule to dissolve in water. Similar to lipophilicity, aqueous solubility is a physicochemical property of molecules which affects their biovailability as well as ADME profiles (Llompart et al., 2024). Therefore, predicting a molecule's solubility is of high importance to the development of orally active drugs/compounds (Klopman et al., 1992). The goal of aqueous solubility prediction is to estimate the logS (log solubility of a molecule in water).

A.5 Relation between Lipophilicity and Solubility

Generally, a high lipophilicity is negatively correlated with aqueous solubility, decreasing oral absorption of a drug (Curatolo, 1998). Moreover, both tasks share the same categories of important molecular subgroups—i.e., *hydrophilic* substructures increase the hydrogen-bonding ability of a molecule, making it more likely to be soluble in water but decreasing lipophilicity while *lipophilic* substructures increase its hydrophobicity. Although both tasks are closely related, computing the logD and logS values requries the consideration of additional, molecule-specific factors such as the melting point (Hill and Young, 2010).

B Model Selection

In our experiments, we consider the following CLMs trained with the masked language modeling (MLM) objective on datasets of small molecules.

Chemberta Chithrananda et al. (2020) release multiple versions of a six-layer model based on RoBERTa (Liu et al., 2019b) trained on different amounts of data sampled from the ZINC15 database (Sterling and Irwin, 2015). We use the two publicly available models, namely, `seyonec/ChemBERTa-zinc-base-v1` (here, referred to as `chemberta-base`; trained on 100k molecules) and `seyonec/chemberta-zinc250k-v1` (`chemberta`; trained on 250k molecules). Both employ a BPE tokenizer with respective vocabulary sizes of 767 and 52k tokens. While `Chemberta` has 83,450,880 parameters, `chemberta-base` has only 44,103,936 parameters.

Chemberta-2 In subsequent work, Ahmad et al. (2022) release models pre-trained on a larger set of molecules sampled from ZINC15 (Sterling and Irwin, 2015). We investigate `DeepChem/chemberta-5M-MLM`, `DeepChem/chemberta-10M-MLM` and `DeepChem/chemberta-77M-MLM`, denoted as `chemberta-2-5M`, `chemberta-2-10M` and `chemberta-2-77M` resepectively, as the other models are pre-trained with domain-specific auxiliary objectives[5]. `chemberta-2-5M/10M/77M` models all share the same number of parameters – 3,427,440 parameters. Compared to the earlier version of Chemberta (Chithrananda et al., 2020), these models have only three encoder layers, a hidden layer size of 384, and a BPE tokenizer with a vocabulary size of 600 tokens. The main difference among the three variants is the training data size: 5M, 10M, and 77M molecules, respectively. We note that `chemberta2` models exhibit tokenization problems [6]. In particular, some halogens such as chlorine (Cl) and bromine (Br) are tokenized incorrectly.

Chemberta-3 Most recently, Singh et al. (2026) released their training framework along with a 12-layer model `DeepChem/ChemBERTa-100M-MLM` (referred to as `chemberta-3`) trained on 100M molecules sampled from ZINC20 (Irwin et al., 2020). In contrast to the `chemberta-2` models, `chemberta-3` is again based on RoBERTa using a self-trained tokenizer with a vocabulary size of 7,924 tokens. Finally, compared to the `chemberta-2` models, both the size of the hidden vectors and the size of the intermediate FFNNs are larger (3072 and 768 dimensions, respectively). `chemberta-3` has 92,126,976 parameters.

Molformer Ross et al. (2022) train a 12 layer RoBERTa-based model using linear attention, rotary positional embeddings and a regex-based tokenizer with a vocabulary size of 2362 tokens. We use the publicly available `ibm-research/MoLFormer-XL-both-10pct` model[7] trained on 100M molecules sampled equally from ZINC15 (Sterling and Irwin, 2015) and PubChem (Kim et al., 2018) with 44,375,040 parameters.

roberta-zinc-480m (Heyer, 2023) is a RoBERTa-based 14 layer, ∼102M parameter model trained on 480M molecules sampled from the ZINC20 database (Irwin et al., 2020). The model is available on Huggingface under `entropy/roberta_zinc_480m`.

For the experiments, we select models that are comparable and only differ with respect to their architecture and training data sizes. We thus omit models trained with domain-specific auxiliary objectives such as MolBERT (Fabian et al., 2020), SELFormer (Yüksel et al., 2023), Mol-BERT (Li and Jiang, 2021), K-BERT (Wu et al., 2022) and those that are not publicly available (e.g., SMILES-BERT, Wang et al. 2019).

C Probing Dataset

We derive our probing dataset from PCQM4Mv2, a publicly available dataset designed for predicting the HOMO-LUMO energy gap. It is part of the open graph benchmark (Hu et al., 2021) and is published under an open source license (CC-BY 4.0).

[5]These are the `chemberta-5M-MTR`, `chemberta-10M-MTR` and `chemberta-77M-MTR` where MTR refers to multitask regression of 200 molecular properties.

[6]see Issue 1 and Issue 2 and Issue 3

[7]We note that the best model (`Molformer-XL`) whose results are reported by (Ross et al., 2022) is unavailable

C.1 Dataset Preprocessing

In Section 3, we described the preprocessing steps used to obtain the final set of probing datasets. Specifically, all molecules were preprocessed with RDKIT (Landrum et al., 2024)[8], and molecules were discarded if they exhibited any of the following issues: 1) invalid SMILES strings: (MolFromSmiles cannot be created), 2) incorrect conformational information (GetConformer().Is3D() returned False), 3) chemistry problems (rdkit.Chem.DetectChemistryProblems is not empty), 4) other processing errors (e.g., due to unrecoverable rotational or double bond information). Finally, the remaining molecules were annotated with binary labels with 1 denoting the presence of a substructure. To extract functional group information, we used the rdkit.Chem.Fragments module from RDKit. Ring structure labels and bond information were annotated using rdkit.Chem.Lipinski. We then binarized the labels.

Runtime estimates The sheer number of instances in the PCQM4Mv2 dataset requires the subsampling of molecules for the probing dataset. For reference, running all 78 probing tasks for a single 14-layer model with 100k instances requires ∼5.4 hours. Scaling this to the whole dataset would require almost 200 hours for only one (out of four) experimental configuration.

C.2 Probing Tasks

Our final probing dataset comprises 78 probing tasks, as summarized in Table 2. The *Molecular Substructure* column lists the abbreviations used for each probing task, while *Description* provides a brief explanation of the corresponding task.

C.3 Dataset Statistics

Table 3 summarizes the sizes of the downsampled training sets for each probing task. Table 4 reports the class distribution in the corresponding test sets. Note that many molecular substructures are highly infrequent, resulting in significant class imbalance. To address this, we evaluate probing performance using the macro-averaged F1 score, which weighs each class equally. For training the probing classifier, we also downsample instances in the training set (i.e., the probing classifier is always trained on a balanced dataset).

[8]version 2024.3.6

D Probing Setup and Compute Infrastructure

D.1 Probing Setup

For each probing task and each encoder layer of a CLM, we train a logistic regression classifier using scikit learn[9] with default parameters and set *max_iter*=2000. The input features are the hidden representations from the corresponding layer. We apply padding to the longest sequence in each batch. Probing experiments were conducted on CPU-only nodes on the same cluster as described in Appendix D.2. As an indicative reference, a single probing run for one probing task on a 14-layer model (e.g., `roberta-zinc-480m`) typically completes in approximately 4.14 minutes on an Intel Xeon E5-2620 CPU. Runtime varies slightly with the number of layers and probing tasks.

D.2 Computing Environment

All experiments were performed on a high performance computing cluster with varying CPU and GPU architectures. Fine-tuning experiments were conducted on a single GPU; either NVIDIA Titan X (12GB), NVidia V100 (32GB), or A100 (40GB). Probing experiments were executed on 2 CPUs equipped with AMD EPYC 7662 (64 cores, 512 GB RAM) and did not require a GPU.

D.3 Compute Time

Overall, $\sim$ 677.9 CPU-only hours were spent for probing. For fine-tuning, the experiments required $\sim$ 57.9 GPU hours for lipophilicity prediction and 77.3 GPU hours for solubility prediction. The downstream task experiments for the baselines required 692.5 GPU hours for `gpt-oss-20b` and 7.2 GPU hours for `Llama-3.2-3B-Instruct`. All baseline experiments using the feature-based models using fingerprints required less than 1 CPU-only hour in total. Finally, the further pre-training experiments required $\sim$ 32.5 GPU hours. In total, this amounts to approximately 678.9 CPU-only hours and 867.4 GPU hours.

E Effect of Pre-training (RQ1)

We provide further evidence and analysis for our findings regarding the effect of pre-training on ring structures (§5.2) and individual molecular substructures (§5.3). Moreover, we perform analysis excluding the strangely behaving `chemberta-3`

[9]sklearn.linear_model.LogisticRegression

Molecular Substructure	Description
NHOH	NHs or OHs
NO	nitrogens and oxygens
AliphaticCarbocycles	aliphatic (containing at least one non-aromatic bond) carbocycles
AliphaticHeterocycles	aliphatic (containing at least one non-aromatic bond) heterocycles
AliphaticRings	aliphatic (containing at least one non-aromatic bond) rings
AromaticCarbocycles	aromatic carbocycles
AromaticHeterocycles	aromatic heterocycles
AromaticRings	aromatic rings
HAcceptors	hydrogen bond acceptors
HDonors	hydrogen bond donors
Heteroatoms	heteroatoms
RotatableBonds	rotatable bonds
SaturatedCarbocycles	saturated carbocycles
SaturatedHeterocycles	saturated heterocycles
SaturatedRings	saturated rings
Ring	rings
Al_COO	aliphatic carboxylic acids
Al_OH	aliphatic hydroxyl groups
Al_OH_noTert	aliphatic hydroxyl groups excluding tert-OH
ArN	N functional groups attached to aromatics
Ar_COO	aromatic carboxylic acid
Ar_N	aromatic nitrogens
Ar_NH	aromatic amines
Ar_OH	aromatic hydroxyl groups
COO	carboxylic acids
COO2	carboxylic acids
C_O	carbonyl O
C_O_noCOO	carbonyl O excluding COOH
C_S	thiocarbonyl
Imine	imines
NH0	tertiary amines
NH1	secondary amines
NH2	primary amines
N_O	hydroxylamine groups
Ndealkylation1	XCCNR groups
Ndealkylation2	tert-alicyclic amines (no heteroatoms, not quinine-like bridged N)
Nhpyrrole	H-pyrrole nitrogens
SH	thiol groups
aldehyde	aldehydes
alkyl_halide	alkyl halides

Molecular Substructure	Description
allylic_oxid	allylic oxidation sites excluding steroid dienone
amide	amides
amidine	amidine groups
aniline	anilines
aryl_methyl	aryl methyl sites for hydroxylation
benzene	benzene rings
bicyclic	bicyclic
diazo	diazo groups
ester	esters
ether	ether oxygens (including phenoxy)
furan	furan
guanido	guanidine groups
halogen	halogens
hdrzine	hydrazine groups
hdrzone	hydrazone groups
imidazole	imidazole
imide	imide groups
ketone	ketones
ketone_Topliss	ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha
lactone	cyclic esters (lactones)
methoxy	methoxy groups -OCH3
morpholine	morpholine
nitrile	nitriles
nitro	nitro groups
nitro_arom	nitro benzene ring substituents
nitro_arom_nonortho	non-ortho nitro benzene ring substituents
oxime	oxime groups
para_hydroxylation	para-hydroxylation sites
phenol	phenols
phenol_noOrthoHbond	phenolic OH excluding ortho intramolecular Hbond substituents
piperdine	piperdine
piperzine	piperzine
priamide	primary amides
pyridine	pyridine
sulfide	thioether
thiazole	thiazole
thiophene	thiophene
unbrch_alkane	unbranched alkanes of at least 4 members (excludes halogenated alkanes)
urea	urea groups

Table 2: **Probing tasks:** Molecular substructure abbreviations and descriptions

model which we study more extensively in Appendix I. Excluding `chemberta-3` this model, we observe multiple additional patterns which we did not discuss in the main paper (as the analysis in the main paper includes `chemberta-3`).

E.1 Ring Structures

For ease of visualization, we omitted the shading that shows the upper and lower quartiles in Figure 2. Figure 7 presents the average probing performance along with its variability for rings, other and all. Notably, rings have the lowest variability across both PT and RI models, while probing performance of other shows greated variability. We provide the 12 ring types used to compute rings in Table 5.

We further analyze the representations of rings and other molecular substructures, finding that the representations of both RI (●) and PT (x) models perform exceptionally well at identifying ring structures compared to all other groups (● and x). Moreover, both RI and PT models **encode ring structures well already at the first encoder layer**, suggesting that these surface-level patterns are easy for the models to extract directly from the input, even without pre-training. We also find that the benefit of pre-training diminishes for ring structures compared to that of other molecular substructures (i.e., the gap between x and ● is much smaller compared to the gap between x and ●).

Different rings For different ring types, we find that—in contrast to aliphatic and aromatic rings—pre-training does improve the probing performance for saturated rings for all models except for `chemberta-3`.

E.2 Individual Molecular Substructures

Figure 8 illustrates the layer-wise differences in probing performance after pre-training for individual molecular substructures. Results for `molformer` are shown in Figure 8a, while `roberta-zinc-480m` is presented in Figure 8b. The performance of `chemberta` and `chemberta-base` is shown in Figure 8c and Figure 8d, respectively. The three `chemberta-2` models trained on 5M, 10M, and 77M molecules are shown in Figure 8e, Figure 8f, and Figure 8g, respectively. Finally, the latest `chemberta-3` model's performance is shown in Figure 8h. We additionally note the following for analysis:

PT > RI In addition to the molecular substructures discussed in §5, we further identify consistent improvements across following substructures (when excluding `chemberta-3`). For *hydroxy groups* (`Al_OH`) and *aldehyde* as well as `C_O`, although to a lesser degree.

F Fine-Tuning Setup

In the following, we provide details about the two downstream tasks, namely, lipophilicity prediction

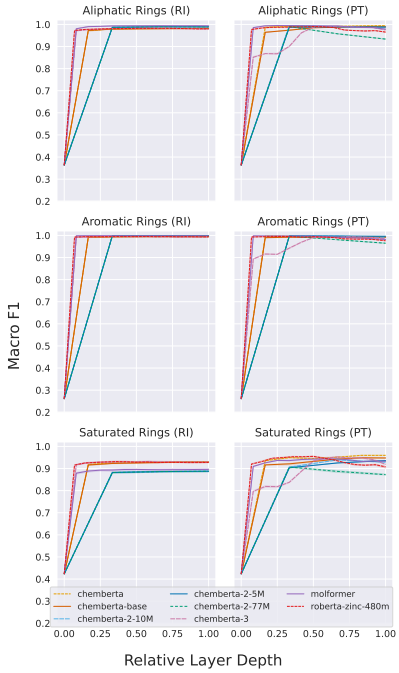


Figure 6: Probing performance on aliphatic (top), aromatic (mid), and saturated (bottom) rings for randomly initialized (left) and pre-trained (right) models. As can be seen, the differences between randomly initialized and pre-trained models are minimal for aliphatic and aromatic rings (except for `chemberta-3`, for which we provide an explanation in Appendix I). This shows that random initializations already encode these ring structures very well.

and solubility prediction. For each task, we provide a dataset description, report the hyperparameters considered for tuning, and elaborate the dataset pre-processing along with the final molecular substructures considered for probing. For both tasks, we use the splits provided by Ross et al. (2022).[10]

F.1 Lipophilicity

Lipophilicity dataset The lipophilicity dataset originates from MoleculeNet (Wu et al., 2018), an open source dataset (MIT license) and comprises 4,200 molecules and their corresponding logD values. Figure 9a shows the distribution of logD values in the training, validation and test splits. As can be seen, the distributions of the logD value follow a similar shape across the training, validation, and test data with peaks around 0–1. Table 8 shows that the validation and test sets have very similar distributions, with means close to -0.079 and -0.004, while the training set is centered around 0. The validation set values are more dispersed (σ= 1.044) compared to the similar variability of the training (σ= 1.0 and 1.01, respectively). The more negative logD values mean that the validation and test sets are centered slightly towards molecules with lower lipophilicity.

Hyperparameters We used the AdamW optimizer with the default settings of $\beta = 0.9$ and $\beta = 0.999$, $\epsilon = 1 \times 10^{-8}$ and weight decay of 0.01, together with a linear decay of the learning rate every 10 epochs with $\gamma = 0.1$. We performed grid search across different batch sizes and learning rates (presented in Table 9). We trained the pre-trained for up to 10 epochs. Note, that we did not conduct a separate hyperparameter search for randomly initialized models and instead, used the optimal hyperparameter values obtained for the corresponding pre-trained models. However, we adjusted the maximum number of training epochs for randomly initialized models to 20 epochs. In addition, because the pre-trained `chemberta-2` models share the same architecture, we fine-tuned only a single randomly initialized `chemberta-2` model, marked with * in Table 10. The final hyperparameters are presented in Table 10.

Probing Setup To conduct meaningful analyses of the effect of fine-tuning on molecular substructures, we excluded any molecular substructures that appear fewer than ten times in either the training or test split of the lipophilicity dataset. Table 6 summarizes the final set of 60 molecular substructures, along with their relative frequencies in the training, validation and test splits of the lipophilicity dataset. Because our fine-tuning analyses categorize molecular substructures as either relevant (hydrophilic and lipophilic) or non-relevant (other) (Appendix A.3), Table 6 also indicates the group assignment (hydrophilic, lipophilic, other) for each molecular substructures.

F.2 Solubility

Solubility dataset We fine-tune all models on the ESOL solubility dataset from the MoleculeNet benchmark (Wu et al., 2018). The dataset comprises 1,127 molecules and their corresponding logS values (log solubility in mols per liter). Notably, the dataset is considerably smaller than the lipophilicity one (1,128 vs 4,200 molecules). We use the train-validation-test splits (80/10/10) provided by Ross et al. (2022). Figure 9b illustrates the distribution of logS values across the training, validation and test splits of the dataset.

[10] Available at `https://github.com/IBM/molformer`

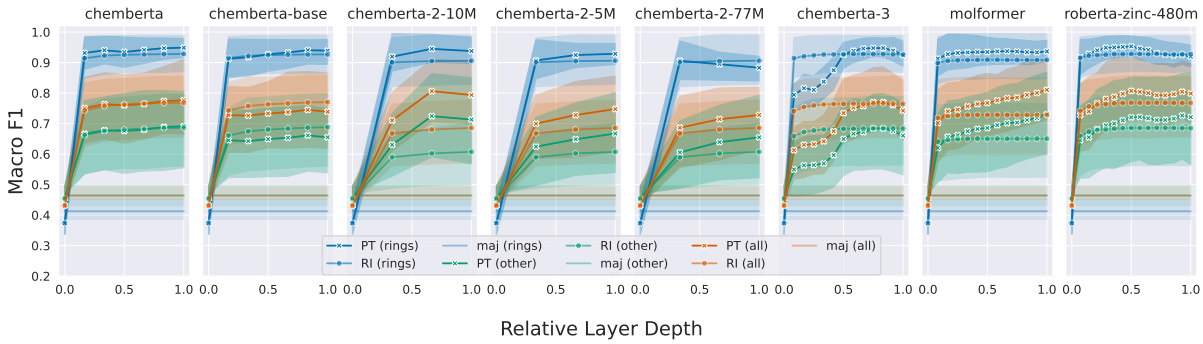


Figure 7: Average probing performance (macro-averaged F1 score ↑) of pre-trained (PT), randomly initialized (RI) models, and the majority class prediction (maj) on molecular substructures with **shading between the upper and lower quartiles**. We report average performance on all 12 classes of rings (avg rings), all other 66 substructures (avg other) and all 78 substructures (avg all).

As can be seen, the distribution of logS values differs substantially between different splits of the datasets; considerably more than in the lipophilicity prediction dataset.

Most notable is the tail difference around a logS value of -8, which occur more frequently in the validation set ($\sim 7.5\%$) compared to the training and the test sets ($\leq 2.5\%$). As shown in Figure 9, while the train and test means are centered around -3 (-3.05 and -2.92, respectively), the validation set has a slightly lower mean of -3.19. Furthermore, the validation set exhibits greater dispersion (standard deviation $\sigma = 2.30$) compared to the training ($\sigma = 2.08$) and test ($\sigma = 2.06$) sets. This might explain lower performance of almost all models on the ESOL dataset, as they were selected on a validation set with a different distribution compared to the training and test sets.

Hyperparameters We largely followed the fine-tuning setup described for lipophilicity in the previous section and evaluate the same hyperparameter ranges (Table 9). We only adjusted the number of training epochs to 20, as our initial experiments showed a slower convergence of models on this task. We furthermore used two types of prediction heads: a linear one and a two-layer MLP. Table 11 lists the final hyperparameter values for fine-tuning the models on the ESOL dataset.

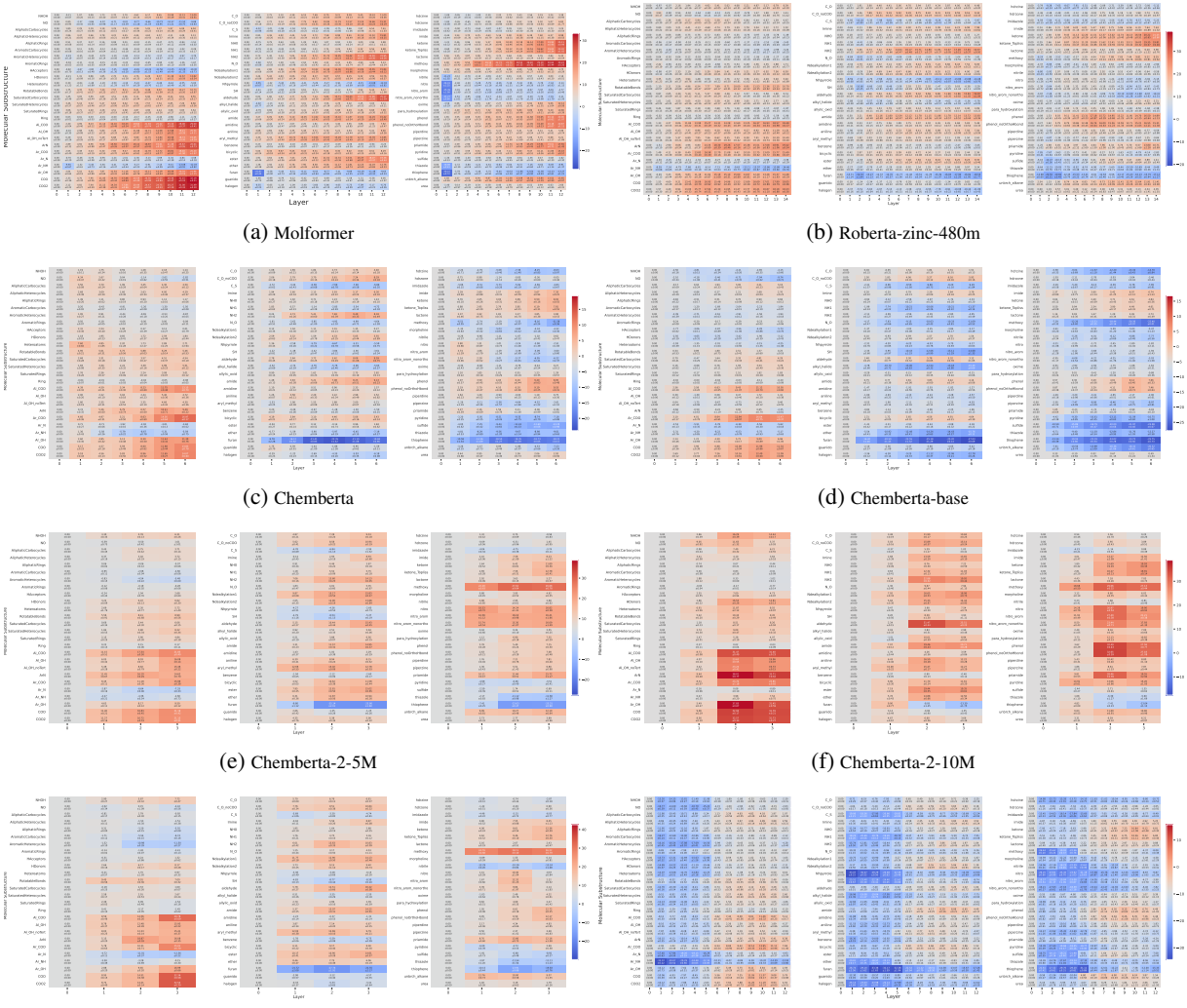
Probing Setup Our probing setup for solubility follows the general probing setup described in Section 4 and Appendix D.1. Similar to lipophilicity prediction (cf. Appendix F.1), we exclude probing tasks for substructures appearing fewer than ten times in either the training or test split of the ESOL dataset. As a result, our ESOL fine-tuning analyses are based on 39 molecular substructures listed in Table 7.

F.3 Lipophilicity and Solubility: Differences

After preprocessing, we analyze both datasets in terms of the distribution of molecular substructures as well as SMILES strings (i.e., molecules).

Overlap in molecular substructures We observe that the distributions of substructures between the lipophilicity and ESOL datasets are quite different, making a direct comparison on a common set of substructures difficult. For example, $\sim$ 90% molecules in the lipophilicity dataset contain `AromaticHeterocycles`, whereas `AromaticHeterocycles` are found in only 10%-15% of molecules in ESOL. Some substructures appearing frequently in the lipophilicity dataset (e.g. `pyridine`, `COO` and `methoxy`) are absent from the ESOL dataset altogether, and vice versa (`unbrch-alkane`, `imide` and `allylic-oxid`). We note that this also results in slight differences for groups of hydrophilic, lipophilic and other molecular substructures (cf. Table 6 and Table 7).

Overlap in SMILES We further inspected the lipophilicity and solubility datasets for overlapping SMILES strings, which would have enabled us to investigate how the two tasks influence the encoding of substructures present in a molecule. In particular, if the effects are reversed. However, we identified only 37 overlapping molecules in the training sets and a single one in the test sets across the two datasets. In light of this, we decide against a direct comparison between the effect of fine-tuning on lipophilicity and ESOL on substructure encoding. Instead, we conduct separate analyses for the two datasets, based on 60 and 39 substructures, respectively.



(a) Molformer (b) Roberta-zinc-480m

(c) Chemberta (d) Chemberta-base

(e) Chemberta-2-5M (f) Chemberta-2-10M

(g) Chemberta-2-77M (h) Chemberta-3

Figure 8: **Effect of pre-training on molecular substructure encoding in CLMs**. We plot the layer-wise difference in performance (% macro-averaged F1) on each probing task between each pre-trained model and its randomly initialized counterpart. Red denotes improvement in probing performance after pre-training. Blue designates degradation in probing performance after pre-training

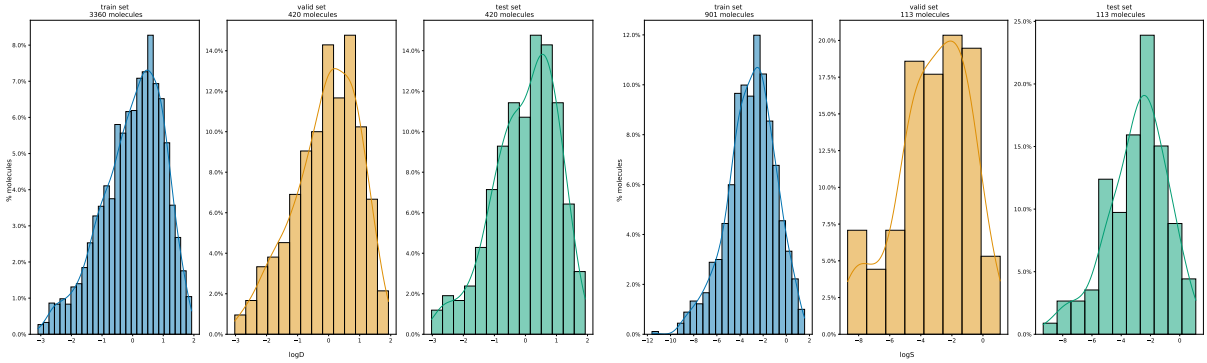
G Effect of Fine-Tuning (RQ2)

Similar to lipophilicity prediction (§6), we analyse the probing performance on solubility prediction for different groups and individual molecular substructures.

G.1 Solubility: Group Analysis

Figure 10 shows the changes (before and after fine-tuning on solubility) in terms of probing performance. We observe similar trends: fine-tuning leads to larger changes in task-relevant groups (lipophilic and hydrophilic), concentrated in the upper layers. In contrast to lipophilicity, fine-tuning on solubility has a more pronounced effect on hydrophilic substructures, as highlighted by the darker red shading. This corroborates our lipophilicity results that molecular substructure learning is consistent with chemical theory. This still holds when we observe `chemberta-2-10M`, which exhibits a consistent unlearning effect on both downstream tasks: For solubility prediction, the unlearning effect is stronger for the lipophilic groups while for lipophilicity prediction it is stronger for the hydrophilic groups.

We can furthermore observe that the effect of fine-tuning on the probing performance is noticeably smaller compared to pre-training (similar as we observed for lipophilicity prediction). Interestingly, we see several common patterns with respect to how fine-tuning on solubility affects RI and PT models. Namely, we find that all RI models



(a) logD (b) logS

Figure 9: The distribution of predicted values across the training, validation and test splits of the **lipophilicity** (9a) and **solubility (ESOL)** (9b) datasets.

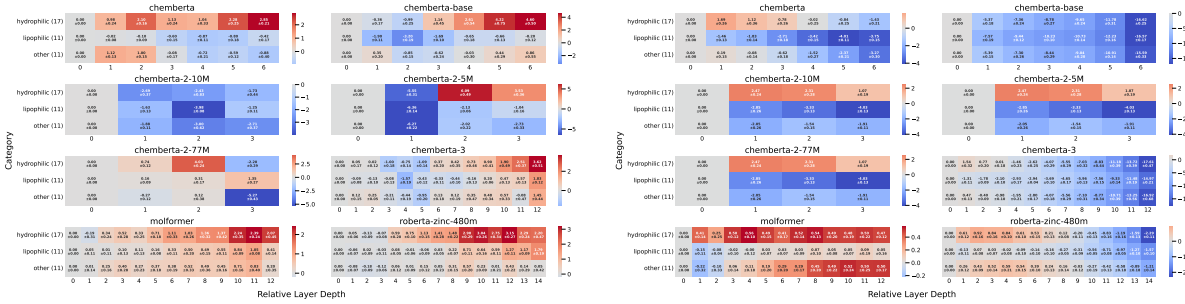


Figure 10: Average differences in probing performance of PT (left) and RI (right) models after fine-tuning on **solubility** (ESOL). We group into *hydrophilic* (top), *lipophilic* (middle), and *other* (bottom) groups (see Table 7). The number of substructures in the corresponding group is indicated next to it. The RI results for `chemberta-2` are based on a single model (hence, are the same for all three variants) as all models use the same architecture.

improve on hydrophilic substructures during fine-tuning. However, in contrast to PT models, these changes are more prominent in lower and middle layers.

G.2 Solubility: Individual Analysis

The heatmaps in Figure 12 present the layer-wise differences in probing performance after fine-tuning on solubility prediction for each substructure (Figure 11 presents the respective heatmaps for lipophilicity prediction).

First, our results show that fine-tuning on solubility improves encoding of `NO` (nitrogens and oxygens), `Heteroatoms` and `HAcceptors` across all PT models. In addition, all models except for `chemberta-2-10M` exhibit improved probing performance on *phenols* (`phenol`, `phenol_noOrthoHbond`), various *hydroxyl groups* (i.e., aliphatic hydroxyl groups (`Al_OH`), aromatic hydroxyl groups (`Ar_OH`), aliphatic hydroxyl groups excluding tert-OH (`Ar_OH_noTert`)), as well as *aromatic nitrogens* (`Ar_N`). These results are consistent with chemical theory. More specifically, Walker (2017) state that replacing a carbon atom with polar heteroatoms N or O is one of the most common approaches to increasing solubility. Furthermore, Harrold and Zavod (2023) note that the ability of polar hydroxyl groups (`OH`) to form hydrogen bonds leads to an increase in aqueous solubility. `Phenols` on the other hand have a moderate effect on solubility due to the presence of a lipophilic aromatic ring and a hydrophilic hydroxyl group. In summary, after fine-tuning on solubility, we observe improvements on hydrophilic substructures known to positively contribute to the solubility of a molecule. For decreased performance, we do not observe such systematic patterns.

While in general there are no consistent trends for RI models, fine-tuning on solubility prediction leads to an improvement of RI models on `NO`, `Heteroatoms`, `HAcceptors` in lower layers. While this might suggest that RI models might be also capable of capturing task-relevant substructures, we do not observe such a trend for lipophilicity

prediction.

H Practical Implications: Further Pre-Training

Our probing experiments have allowed us to identify three interesting patterns. First, we saw that pre-training leads to worse encodings of `thiophene`, `thiazole` and `furan` in all pre-trained models compared to the randomly initialized ones. Second, all `chemberta-2` models use the same architecture and exhibit consistent improvements in probing performance on halogens while we observe the reverse for all other models. Finally, we observed diverging patterns in substructure encoding for `chemberta-2-5M` and `chemberta-2-10M` which differ only in terms of pre-training data. We conjecture that these differences might stem from different compositions of the pre-training data and conduct experiments to see if further pre-training models on specific datasets can mitigate the low probing performance.

PT<RI In Section 5.3, we saw that pretraining degraded probing performance on substructures such as `thiophene`, `thiazole` and `furan` across all models. We hypothesize that the models were undertrained on these substructures. While we cannot conduct a frequency analysis of these substructures due to the unavailability of the pre-training data, further pre-training models on molecules that contain these molecular substructures should allow models to somewhat recover from the low performance. For this experiment, we select three models—`chemberta-2-5M` (3 layers), `chemberta` (6 layers) and `molformer` (12 layers)—each representing a small, mid-sized, and large model.

Chemberta-2 and Halogens Our analysis in Section 5.3 showed that `chemberta-2-5M` and `chemberta-2-10M` improved on halogens after pre-training, whereas performance degraded for all other models (which generally showed a very high probing performance). Moreover, in Section 6.4 we found that, unlike the other models, fine-tuning on lipophilicity led to further improvements on halogens for `chemberta-2-5M` and `chemberta-2-10M`. We thus hypothesize the following:

1. Further pre-training on halogen-containing molecules will improve `chemberta-2-5M` and `chemberta-2-10M`'s performance on halogens.
2. As further pre-training improves probing performance for halogens, we expect smaller probing performance gains on halogens after fine-tuning on lipophilicity.

Same architecture, different performance We conjecture that probing can furthermore be used to identify molecular substructures that a model has seen less during pre-training and that further pre-training can be used to mitigate this gap. We study this hypothesis on two models with the same architecture (`chemberta-2-5M` and `chemberta-2-10M`) which exhibit large differences in probing performance on `phenol` (Figure 16, left). We conjecture that this is a result of the difference in the pre-training data (and its phenol distribution).

H.1 Pre-training Dataset

For further pre-training, we subsample from the Guacamol dataset (Brown et al., 2019), a dataset that is designed for benchmarking de novo molecular design. It comes with pre-defined train, validation and test splits with 1,273,104, 79,568 and 238,706 molecules, respectively.

Pre-processing We first canonicalize all SMILES strings and discard 3,217 molecules which appear in the downstream lipophilicity and solubility datasets. We then annotate the canonicalized and cleaned training and validation splits of Guacamol with binary labels (where 1 denotes the presence of a substructure) using RDKIT. For each substructure of interest, we select SMILES strings containing this substructure from the training and validation sets. Having obtained the substructure-containing subsets, we randomly sample 50,000 and 3,000 molecules from them.

H.2 Experimental Setup

Pre-training We conducted pre-training for all models on the subset of data sampled from Guacamol for 20 epochs (15,625 steps) with a batch size of 64. We used the AdamW optimizer with the default settings of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$ and weight decay of 0.01. We use a scheduler with a linear learning rate decay of 0.1 and 100 warm-up steps.

Probing Setup Our probing setup is identical to that summarized in Appendix D.1.

Finetuning Setup For fine-tuning on lipophilicity prediction, we follow the setup described in

Appendix F.1 and use the best hyperparameters for pre-trained models listed in Table 10.

H.3 Results

PT<RI Figure 13 shows the impact of further pre-training on more data with specific substructures (`furan`, `thiazole`, `thiophene`) on the probing performance on the corresponding substructure for `chemberta-2-5M`, `chemberta` and `molformer`.

First, we observe an **improvement in probing performance across all three substructures and models**. While further pre-training benefits the `chemberta` model most (originally with a pre-training dataset of only 250k molecules), large improvements for `molformer` are generally observed in the upper layers. Compared to the other two models, `chemberta-2-5M` exhibits more variation in the magnitude of improvement.

Second, **further pre-training has a mixed effect on the downstream performance on lipophilicity prediction.** Whereas `molformer`'s downstream performance is negatively affected in all three cases, `chemberta-2-5M` mostly benefits from further pre-training. In contrast, `chemberta`'s performance improves when further pre-trained on furan and is negligibly affected by pre-training on `thiazole` and `thiophene`. This may change under extensive hyperparameter tuning.

Chemberta-2 and Halogens Figure 15 presents the results of further pre-training of models from the `chemberta-2` family on halogen data.

First, we observe slight performance improvements for `chemberta-2-5/10M` and substantial performance gains for `chemberta-2-77M` when probing for `halogens`. Interestingly, we find that the probing performance on halogens converges towards a joint upper bound which is also shared in the randomly initialized `chemberta-2` model. This is in stark contrast to all other models shown in Figure 14 which have a substantially higher performance in the upper layers. One reason for this might be the tokenization issues of the `chemberta-2` models which result in turning halogens such as Cl and Br into C and B. [11]

Second, improved probing performance on halogens after further pre-training makes their encoding more robust, resulting in smaller changes after fine-tuning, detailed in Table 13. We conjecture that these changes might be more prominent without the tokenization issues.

[11] see Issue 1 and Issue 2 and Issue 3

Finally, similar to the effect of further pre-training on `furan`, `thiazole` and `thiophene` on downstream performance, the results in Table 14 demonstrate that pre-training on more `halogen` data has a mixed effect on downstream performance. Whereas `chemberta-2-5M` improves on the downstream task and `chemberta-2-10M`'s performance is substantially negatively affected, `chemberta-2-77M` remains largely unchanged. One reason for this disparity despite all models sharing the same architecture might be the different hyperparameters which were not tuned individually for further pre-trained models. Tables 10 and 11 show different optimal hyperparameters for different `chemberta-2` models across both tasks. Hence, extensively tuning the hyperparameters for the further pre-trained models might mitigate drops in the performance.

Same architecture, different performance Figure 16 shows the effect of pre-training `chemberta-2-5M` and `chemberta-2-10M` on more data containing phenol substructures.

Although further pre-training on data containing phenol improved the respective probing performance, the performance gap between the two models persisted. We conjecture that the difference in the amount of original pre-training data (5 million vs 10 million) might be responsible for this and that longer pre-training might further reduce this gap. Again, Table 15 shows that further pre-training can both negatively and positively affect downstream performance.

H.4 Discussion

Our results suggest that substructure probing may serve as a diagnostic tool to identify whether a model was undertrained on specific substructures. Our additional pre-training experiments provide an exploratory demonstration that targeted further pre-training on infrequent substructures can improve the representations of these substructures and increase their representational robustness during fine-tuning. Based on our findings, systematically studying the pre-training dynamics of CLMs via probing could be a promising endeavor for future work.

I The Peculiar Case of Chemberta-3

`Chemberta-3` exhibits a peculiar "dome" in the average probing performance at lower layers (Figure 2). We hypothesize that this might stem from

the irregularities during the pre-training process. To investigate this as well as the impact of different training data, we further pre-train `chemberta-3` on subsets of the Guacamol (Brown et al., 2019) (CC BY-SA 3.0) and ZINC-100M (Singh et al., 2026) datasets.

I.1 Datasets

To obtain the training data, we subsample the Guacamol (Brown et al., 2019) and ZINC-100M (Singh et al., 2026) datasets. Guacamol comes with pre-defined train, validation and test splits with 1,273,104, 79,568 and 238,706 molecules, respectively. Singh et al. (2026) curated ZINC-100M by sampling 100M molecules from the ZINC20 database (Irwin et al., 2020).

I.2 Data Preprocessing and Sampling

Guacamol We first canonicalize all SMILES strings. We then discard any molecules which appear in the downstream lipophilicity and solubility datasets. In the case of Guacamol, we discard 3,217 molecules. We then randomly sample 100,000 and 10,000 molecules from the training and validation sets, respectively.

ZINC-100M Due to the large size of ZINC-100M (∼7.7GB), we first performed random sampling on the byte positions, obtaining 100,000 and 10,000 molecules for the training and validation sets, respectively. We then canonicalize the SMILES strings and check for SMILES strings which overlap with the those in the fine-tuning datasets.

I.3 Pre-Training Setup

We follow the same pre-training setup described in Appendix H.2. However, we train the model for 32 epochs (50k steps).

I.4 Effect of Further Pre-Training

Figure 17 shows the effect of further pre-training `chemberta-3` on subsets of Guacamol and ZINC-100M. Interestingly, the model further pre-trained on Guacamol exhibits substantially better probing performance compared to the model trained on ZINC across all molecular substructures. Further pre-training on ZINC benefits other (non-ring) substructures more than rings, with improved performance mostly in upper layers.

Importantly, the lower layers which are characterized by substantially lower performance remain mostly unaffected by further pre-training on ZINC while showing a slight improvement for Guacamol. To better understand these effects, we analyze the frequency distributions of molecular substructures in Figure 18 across both datasets, finding that they are mostly similar. Surprisingly, the actual overlap in terms of molecules (i.e., SMILES string) is almost zero with only four overlapping strings in the training split. This suggest that small differences in the training frequency of molecular substructures may already substantially affect the encodings during pre-training. Furthermore, there might be other effects beyond molecular substructures that need to be studied in future work.

Nonetheless, we conclude that further pre-training a model on a small, carefully curated dataset can already mitigate some of the negative probing performances.

J Lipophilicity and Solubility Prediction with Other Models

We provide additional experiments on the downstream tasks using models other than CLMs. More specifically, we investigate models using traditional chemical features, referred to as "molecular fingerprints" and more recent, decoder-only foundation models.

J.1 Models Trained on Fingerprints

Fingerprints are a common way of representing molecules in chemistry to be used as features in machine learning models. They are often generated using hand-crafted algorithms that output a binarized feature vector representing the presence of specific molecular substructures in a molecule. One of the most famous methods is Morgan's algorithm (Morgan, 1965), with various other methods that have been developed over time. In this work, we evaluate four different types of fingerprints:

MFP Morgan fingerprints as introduced by Morgan (1965).

RDFP The fingerprinting method provided in RDKit (Landrum et al., 2024).

ATFP Atom pair fingerprints introduced by (Carhart et al., 1985). In the RDKit implementation, an atom is represented by a tuple of the atomic number, number of pi electrons, and the degree of the atom, with the option of adding chirality information.

TTFP Topoligical torsion fingerprints are similar to atom pair fingerprints but use 4-atom se-

quences to capture more local graph structure. (Nilakantan et al., 1987).

We extract fingerprints using a radius of 2, which is equivalent to the same value used in MoleculeNet (Wu et al., 2018)—the benchmark from which we use the lipophilicity and solubility datasets—and evaluate four different vector sizes $dim \in \{300, 512, 1024, 2048\}$ For each of the extracted fingerprints, we then train a logistic regression model (LR), a support vector machine (SVM), and a gradient boosted tree (XGB). For the SVM, we additionally tune different values for $c=\{0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 4, 16, 64, 256, 1024\}$ using the validation set.

Results. Table 16 and Table 17 show the performance (RMSE) of different fingerprinting algorithms across different models for lipophilicity and solubility, respectively. Most notably, the SVM consistently performs best across all dimensions, fingerprints, and tasks, showing the best performance for atom pairs fingerprinting (ATFP) with a dimension of 2048 and a c of 16. Interestingly, we find that Morgan fingerprints (MFP) perform most stable across different dimensions and best on average. Finally, we find that for lipophilicity prediction the SVM and XGB both benefit from higher dimensions, while this is the opposite for the LR model. Similarly, the SVM and XGB both perform robustly across all dimensions and fingerprints, while LR performance varies a lot; especially for higher dimensions.

J.2 Foundation Models

Recently, an increasing number of works have investigated the potential of foundation models for chemistry tasks (Guo et al., 2023, 2024). While they find that some do perform decently on MPP tasks, they primarily focus on classification tasks. In this work, we provide complementary results on two regression tasks (i.e., lipophilicity and solubility prediction).

Experimental setup. We conduct experiments for two large language models, namely `Llama-3.2-3B-Instruct` (Grattafiori et al., 2024) and `gpt-oss-20B` (OpenAI, 2025). We prompt each model in a zero-shot setting, asking it to predict the logD (or logS) value of the given molecule (basic). We further evaluate three additional setups to accommodate for the model's lack of chemical knowledge. First, we provide explanations on which functional groups increase and decrease the logD (or logS) value (expl). Second, we provide a list of extracted molecular substructures present in the molecule (hint). Finally, we pass both piece of information to the model (both). For generation, we use nucleus sampling (Holtzman et al., 2020) with $p = 0.95$, a temperature of 0.8 and a maximum token budget of 4,096 tokens. We evaluate all three reasoning levels for the `gpt-oss-20B` model, i.e., low, medium, and high. For the high reasoning level, we find that the model tends to generate very lengthy responses that exceed the token budget. We thus further evaluate token budgets of 8,192 and 16,384. All experiments were conducted on a high performance computing cluster with 4 NVIDIA A100 (40GB). Each GPU ran for ≈112 hours (∼4.7 days in total).

Results. Table 18 and Table 19 provide the results of our experiments for lipophilicity prediction and solubility prediction, respectively. Overall, we can see that all LLMs perform worse compared to the models that use CLMs or fingerprints. We further find that a high reasoning level does not `gpt-oss-20b` automatically lead to an improve performance but instead, can produce responses that exceed the token budget (as it consistently happens for the high reasoning level). Interestingly, providing the models with either explanations or information about the present functional groups does improve their performance, however providing both leads to a worse performance for lipophilicity prediction. This follows the findings by Ganeeva et al. (2024, 2025) who find that LLMs might still be lacking in terms of compositionality (as they seem incapable of putting together the provided explanation and the functional groups). In contrast, we do not consistently observe this behavior for solubility prediction, as providing both even leads to the best result (for $\texttt{gpt-oss-20b}_{4,096}$, low). However, we also see that the performance of smaller models or lower and medium reasoning levels is substantially worse for solubility prediction compared to lipophilicity prediction. Figure 19 showcases an example response for the best performing LLM ($\texttt{gpt-oss-20b}_4, 096$, with hints and medium reasoning level). As can be seen, the model seemingly "reasons" about the task, but assigns the wrong sign to the predicted logD value, indicating that it does not have actual knowledge about the task.

Prompt templates. In the following, we provide all prompt templates that we used in our experiments. Note, that the basic prompt (we provide individual templates for lipophilicity and solubility prediction) is always present, and that the respective sub-prompts are appended accordingly. We always provide the molecule last with the prefix ("This is the molecule") shifted accordingly. The setting both combines all three templates (one of the basic templates, chemical explanations, and preprocessed functional groups).

Basic Prompt (lipophilicity)

System: You are now working as an excellent expert in chemistry and drug discovery.
User: Predict the lipophilicity the following molecule. The lipohilicity of a molecule is defined by the logD value, the logarithmic form of the distribution coefficient D. The molecule is represented using the canonical form of its SMILES string.

This is the molecule:
{canonicalized SMILES string}

Basic Prompt (solubility)

System: You are now working as an excellent expert in chemistry and drug discovery.
User: Predict the solubility of the following molecule. The solubility of a molecule is measured by the log solubility in mols per liter. The molecule is represented using the canonical form of its SMILES string.

This is the molecule:
{canonicalized SMILES string}

+ Chemical Explanations:

For prediction it is also important to consider the following functional groups that affect {lipophilicity, solubility}. These are functional groups that increase {lipophilicity, solubility}:
{List of groups that increase logD or logS value}
These are functional groups that decrease {lipophilicity, solubility}:
{List of groups that decrease logD or logS value}

This is the molecule:
{canonicalized SMILES string}

+ Preprocessed Functional Groups:

As a hint, you are also given the functional groups that are present in the molecule. Following functional groups are present in the molecule:
{all present groups}

This is the molecule:
{canonicalized SMILES string}

Probing task	Training Split Size			Probing task	Training Split Size		
	Seed 42	Seed 77	Seed 4		Seed 42	Seed 77	Seed 4
NHOH	54574	54936	54628	aldehyde	4128	4054	4228
NO	5308	5374	5454	alkyl_halide	12224	12168	12022
AliphaticCarbocycles	44174	43876	44168	allylic_oxid	19950	20470	20198
AliphaticHeterocycles	66372	66602	66996	amide	19672	19528	19746
AliphaticRings	99440	99306	99924	amidine	9086	9024	9054
AromaticRings	76818	77146	77070	aniline	27334	27340	27794
AromaticCarbocycles	75432	75536	76012	aryl_methyl	42070	41994	41350
AromaticHeterocycles	65176	64620	64144	benzene	75432	75530	76012
HAcceptors	4662	4778	4826	bicyclic	35830	35808	35426
HDonors	53418	53890	53548	ester	15866	15856	15890
Heteroatoms	2076	2060	2156	ether	63050	63536	63826
RotatableBonds	15078	15180	15118	furan	5358	5272	5252
SaturatedRings	70664	70306	71046	guanido	2322	2364	2346
SaturatedCarbocycles	34032	33544	33866	halogen	35422	35528	35384
SaturatedHeterocycles	42976	43060	43562	hdrzine	3384	3348	3280
Ring	23322	23484	23270	hdrzone	1754	1792	1944
Al_COO	14092	14152	13996	imidazole	7152	7336	7074
Al_OH	52366	52306	52222	imide	1254	1166	1206
Al_OH_noTert	48914	48874	48840	ketone_Topliss	13072	13170	13448
ArN	10872	10808	11102	lactone	1930	1940	2090
Ar_COO	3992	3956	3878	methoxy	22138	22602	22642
Ar_N	56262	55878	55370	morpholine	2316	2170	2256
Ar_NH	10330	10088	9820	nitrile	11304	11414	11520
Ar_OH	8548	8220	8362	nitro	4444	4384	4396
COO	18034	18050	17816	nitro_arom	2410	2418	2368
COO2	19182	19154	18944	nitro_arom_nonortho	1182	1200	1116
C_O	68424	68080	68792	oxime	1570	1476	1556
C_O_noCOO	52970	52798	53526	para_hydroxylation	12858	12690	12736
C_S	870	812	792	phenol	5432	5382	5422
Imine	28506	28680	28360	phenol_noOrthoHbond	5354	5284	5306
ketone	15014	15276	15558	piperdine	10696	10720	10628
NH0	75328	75390	76000	piperzine	3064	3168	3176
NH1	70862	70678	71058	pyridine	18140	18244	18118
NH2	39752	39668	40030	priamide	576	556	572
N_O	3462	3568	3484	sulfide	9854	9744	9946
Ndealkylation1	14322	14354	14098	thiazole	3874	3852	3782
Ndealkylation2	13014	13062	13032	thiophene	5610	5394	5370
Nhpyrrole	10330	10088	9820	unbrch_alkane	12118	12130	11878
SH	3550	3444	3436	urea	758	752	768

Table 3: **Probing training dataset statistics.** Each row shows the size of the undersampled training split for each probing task across three random samples. Some substructures occur infrequently in the original PCQM4Mv2 dataset, resulting in smaller training sets for certain tasks (e.g., *priamide*, *urea*).

Probing task	% of molecules w/substructure			Probing task	% of molecules w/substructure		
	Seed 42	Seed 77	Seed 4		Seed 42	Seed 77	Seed 4
NHOH	61.20	61.51	60.86	aldehyde	2.04	2.02	2.00
NO	95.38	95.48	95.42	alkyl_halide	6.27	6.12	6.12
AliphaticCarbocycles	17.38	17.34	17.29	allylic_oxid	12.45	12.53	12.56
AliphaticHeterocycles	29.50	29.54	30.05	amide	14.00	14.27	14.69
AliphaticRings	42.71	42.77	43.22	amidine	5.04	5.41	5.44
AromaticRings	64.66	64.69	64.67	aniline	13.87	13.58	13.42
AromaticCarbocycles	49.26	48.84	49.48	aryl_methyl	17.68	17.71	18.09
AromaticHeterocycles	27.84	28.32	27.61	benzene	49.20	48.78	49.43
HAcceptors	95.94	96.14	96.03	bicyclic	24.58	24.25	24.30
HDonors	62.17	62.59	61.91	ester	10.78	11.16	11.09
Heteroatoms	97.82	97.92	97.84	ether	30.86	31.51	31.12
RotatableBonds	88.03	88.13	87.86	furan	2.37	2.66	2.63
SaturatedRings	26.55	26.68	26.68	guanido	1.56	1.64	1.59
SaturatedCarbocycles	11.27	11.21	11.04	halogen	18.18	17.68	17.57
SaturatedHeterocycles	17.39	17.46	17.59	hdrzine	2.75	2.65	2.61
Ring	85.53	85.84	86.00	hdrzone	3.01	3.13	3.33
Al_COO	7.17	7.14	7.42	imidazole	3.17	3.29	3.16
Al_OH	21.76	21.98	21.84	imide	1.52	1.57	1.62
Al_OH_noTert	20.09	20.36	20.27	ketone_Topliss	7.95	7.81	7.54
ArN	4.68	4.50	4.27	lactone	1.38	1.44	1.39
Ar_COO	2.06	2.09	2.11	methoxy	10.91	11.30	11.19
Ar_N	23.54	23.74	22.96	morpholine	1.12	1.18	1.14
Ar_NH	5.59	5.83	5.48	nitrile	5.19	5.31	5.36
Ar_OH	5.36	5.11	4.88	nitro	4.65	4.46	4.46
COO	9.20	9.19	9.48	nitro_arom	2.81	2.69	2.75
COO2	9.24	9.24	9.54	nitro_arom_nonortho	1.77	1.71	1.71
C_O	42.93	43.13	43.42	oxime	1.71	1.76	1.74
C_O_noCOO	35.73	36.13	36.20	para_hydroxylation	10.64	10.49	10.81
C_S	1.06	1.01	1.01	phenol	3.87	3.75	3.58
Imine	12.85	12.77	13.32	phenol_noOrthoHbond	3.75	3.65	3.50
ketone	10.04	9.84	9.61	piperdine	3.48	3.38	3.58
NH0	57.19	57.01	57.49	piperzine	1.17	1.23	1.21
NH1	28.09	28.98	28.75	pyridine	7.32	7.14	6.98
NH2	12.85	12.63	12.24	priamide	1.42	1.49	1.47
N_O	1.97	1.95	2.06	sulfide	6.33	6.22	6.25
Ndealkylation1	3.51	3.48	3.54	thiazole	1.45	1.55	1.46
Ndealkylation2	3.62	3.61	3.87	thiophene	2.36	2.44	2.44
Nhpyrrole	5.59	5.83	5.48	unbrch_alkane	7.11	7.14	7.00
SH	3.23	3.26	3.12	urea	1.76	1.70	1.71

Table 4: **Class distributions in the probing test set.** Each row shows the percentage of molecules containing a molecular substructure probed for in each of the three randomly drawn test samples.

Molecular Substructure
AliphaticCarbocycles
AliphaticHeterocycles
AliphaticRings
AromaticRings
AromaticCarbocycles
AromaticHeterocycles
SaturatedCarbocycles
SaturatedHeterocycles
SaturatedRings
benzene
bicyclic
Ring

Table 5: List of the 12 ring types used for computing the average probing performance for rings shown in Figure 2.

	Molecular Substructure	% of molecules w/substructure		
		train	val	test
lipophilic	AliphaticCarbocycles	14.79	16.67	19.29
	AliphaticHeterocycles	48.87	44.29	48.33
	AliphaticRings	56.58	55.24	61.43
	AromaticCarbocycles	89.70	90.00	90.48
	AromaticHeterocycles	73.42	76.43	74.05
	AromaticRings	99.02	98.33	98.33
	SaturatedCarbocycles	12.59	14.05	17.14
	SaturatedHeterocycles	40.27	35.48	40.00
	SaturatedRings	47.74	45.48	52.86
	benzene	89.70	90.00	90.48
	halogen	43.51	40.00	40.95
	pyridine	29.64	28.57	29.29
	thiazole	6.31	6.43	7.14
	thiophene	6.04	4.76	6.90
hydrophilic	NHOH	86.37	88.57	85.00
	NO	99.97	100.00	100.00
	Heteroatoms	99.97	100.00	100.00
	HAcceptors	99.94	100.00	100.00
	HDonors	86.37	88.81	85.00
	Al_COO	7.71	5.48	6.90
	Al_OH	13.30	16.43	10.00
	Al_OH_noTert	11.19	14.05	8.81
	Ar_COO	3.36	5.71	4.29
	Ar_NH	16.85	19.05	19.52
	Ar_OH	8.66	11.19	7.86
	COO	11.07	11.19	10.95
	COO2	11.07	11.19	10.95
	NH0	87.68	83.81	84.76
	NH1	65.48	68.57	66.90
	NH2	19.55	20.71	20.71
	amide	51.25	49.29	52.38
	aniline	53.01	58.33	53.33
	ether	45.92	44.05	42.38
	morpholine	6.22	7.38	6.90
	para_hydroxylation	16.93	18.81	21.19
	phenol	6.70	8.81	6.67
	phenol_noOrthoHbond	6.58	8.81	6.19
	piperdine	18.21	15.48	17.62
	piperzine	9.64	8.81	10.00
	priamide	4.76	4.29	6.90
other	RotatableBonds	97.65	96.19	96.90
	Ring	99.94	99.76	99.52
	ArN	9.97	12.38	8.10
	Ar_N	69.43	71.90	69.05
	C_O	63.15	61.43	63.57
	C_O_noCOO	56.43	55.71	57.38
	Imine	1.93	1.67	2.62
	Ndealkylation1	9.43	7.14	11.67
	Ndealkylation2	16.37	11.19	14.29
	Nhpyrrole	16.85	19.05	19.52
	alkyl_halide	8.39	6.19	7.14
	aryl_methyl	25.77	27.14	25.00
	bicyclic	53.57	54.29	53.33
	imidazole	9.26	8.10	7.86
	ketone	4.85	6.90	7.38
	ketone_Topliss	4.08	5.71	6.43
	methoxy	16.88	15.48	17.62
	nitrile	7.83	7.62	6.67
	sulfide	4.97	6.90	4.76
	urea	4.35	4.05	3.81

Table 6: **Molecular substructure frequencies in the lipophilicity dataset**: Percentage of molecules in the training, validation and test splits of the lipophilicity dataset containing each of the 60 specific molecular substructures.

	Molecular Substructure	% of molecules w/substructure		
		train	val	test
lipophilic	AliphaticCarbocycles	12.99	13.27	15.04
	AliphaticHeterocycles	14.65	9.73	16.81
	AliphaticRings	25.08	22.12	26.55
	AromaticCarbocycles	49.72	48.67	53.98
	AromaticHeterocycles	15.32	10.62	19.47
	AromaticRings	58.60	55.75	61.95
	SaturatedCarbocycles	9.66	8.85	12.39
	SaturatedHeterocycles	9.32	7.96	12.39
	SaturatedRings	17.20	16.81	21.24
	benzene	49.72	48.67	53.98
	halogen	29.41	30.97	27.43
hydrophilic	NHOH	43.17	34.51	53.98
	NO	71.37	67.26	83.19
	Heteroatoms	86.13	84.07	89.38
	HAcceptors	72.25	68.14	83.19
	HDonors	43.62	34.51	53.98
	Al_OH	13.98	15.93	11.50
	Al_OH_noTert	11.54	12.39	9.73
	Ar_OH	7.88	4.42	13.27
	NH0	26.75	27.43	33.63
	NH1	18.87	14.16	30.97
	amide	18.87	15.93	28.32
	aniline	15.76	10.62	23.89
	ester	8.88	7.08	8.85
	ether	21.31	14.16	22.12
	para_hydroxylation	10.43	7.96	15.04
	phenol	6.99	4.42	11.50
	phenol_noOrthoHbond	6.88	4.42	11.50
other	RotatableBonds	68.04	67.26	68.14
	Ring	71.81	69.03	75.22
	Ar_N	14.21	8.85	15.93
	C_O	34.63	31.86	43.36
	C_O_noCOO	34.63	31.86	43.36
	allylic_oxid	11.43	12.39	12.39
	aryl_methyl	12.32	14.16	15.04
	bicyclic	22.09	18.58	28.32
	imide	4.99	7.08	11.50
	unbrch_alkane	10.65	12.39	9.73
	urea	6.44	6.19	12.39

Table 7: **Molecular substructure frequencies in the ESOL dataset**: Percentage of molecules in the training, validation and test splits of the ESOL dataset containing each of the 39 specific molecular substructure.

	logD				logS		
	train	**valid**	**test**		**train**	**valid**	**test**
mean	0.000000	-0.078882	-0.004254	**mean**	-3.047764	-3.186230	-2.919593
std	1.000149	1.043810	1.014424	**std**	2.076592	2.294987	2.061854
min	-3.091151	-3.024248	-3.024248	**min**	-11.600000	-8.710000	-9.332000
25%	-0.632465	-0.718184	-0.665917	**25%**	-4.300000	-4.570000	-3.955000
50%	0.145283	0.061654	0.157827	**50%**	-2.900000	-3.000000	-2.630000
75%	0.755773	0.707686	0.755773	**75%**	-1.614000	-1.456000	-1.600000
max	1.926575	1.926575	1.926575	**max**	1.580000	1.110000	1.100000

Table 8: Summary statistics for the distribution of predicted values – logD (lipophilicity) and logS (solubility)– in the training, validation and test splits of the lipophilicity and solubility datasets, respectively.

Task	**Regression Head**	**Batch size**	**Learning rate**	**# Epochs**	
		PT	RI		
Lipo	LR	$[8, 16, 32, 64]$	$[0.01, 1 \times 10^{-3}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-5}, 5 \times 10^{-5}]$	10	20
ESOL	LR, 2L-MLP	$[8, 16, 32, 64]$	$[0.01, 1 \times 10^{-3}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-5}, 5 \times 10^{-5}]$	20	20

Table 9: Hyperparameter ranges explored. For randomly initialized models, we did not perform extensive hyperparameter tuning, instead adopting the optimal batch size and learning rates from their pre-trained counterparts. However, we increased the number of epochs. For ESOL, we evaluate both, linear regression (LR) and 2-layer MLP (2L-MLP) heads.

Model	**Batch size**		**Learning rate**		**Best Epoch**	
	PT	RI	PT	RI	PT	RI
`chemberta-base`	16	16	1×10^{-4}	1×10^{-4}	8	18
`chemberta`	16	16	1×10^{-5}	1×10^{-5}	9	10
`chemberta-2-5M`	8	32*	5×10^{-5}	5×10^{-4}*	8	20*
`chemberta-2-10M`	32	32*	5×10^{-4}	5×10^{-4}*	10	20*
`chemberta-2-77M`	32	32*	5×10^{-4}	5×10^{-4}*	7	20*
`molformer`	16	16	1×10^{-4}	1×10^{-4}	10	20
`roberta-zinc-480m`	32	32	1×10^{-5}	1×10^{-5}	7	17
`chemberta-3`	64	64	1×10^{-4}	1×10^{-4}	10	10

Table 10: Final hyperparameter values for pre-trained and randomly initialized models fine-tuned for **lipophilicity**. Both pre-trained and ranodmly initalized models share the hyperparameters except for the best epoch. Note that `chemberta-2-5M`, `chemberta-2-10M`, `chemberta-2-77M` models share the same architecture. Thus, we fine-tune a single randomly initialized model for all of them (`chemberta-2`), selecting the average optimal hyperparameters across all three pre-trained models (*).

Model	Head type	Batch size		Learning rate		Best Epoch	
		PT	RI	PT	RI	PT	RI
chemberta-base	linear	16	16	1×10^{-4}	1×10^{-4}	19	18
chemberta	linear	8	8	5×10^{-5}	5×10^{-5}	12	17
chemberta-2-5M	linear	64	8*	0.001	5×10^{-4}*	19	19*
chemberta-2-10M	MLP	8	8*	5×10^{-4}	5×10^{-4}*	20	19*
chemberta-2-77M	linear	8	8*	5×10^{-4}	5×10^{-4}*	6	19*
molformer	linear	16	16	5×10^{-5}	5×10^{-5}	18	19
roberta-zinc-480m	linear	32	32	5×10^{-5}	5×10^{-5}	16	20
chemberta-3	linear	32	32	1×10^{-4}	1×10^{-4}	19	20

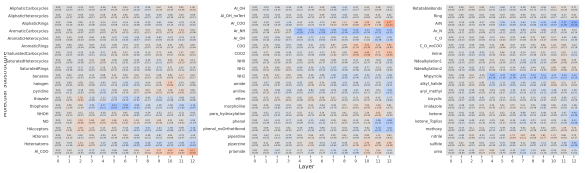
Table 11: Final hyperparameter values for pre-trained and randomly initialized models fine-tuned on **ESOL**. Both pre-trained and ranodmly initalized models share the hyperparameters except for the best epoch. Note that chemberta-2-5M, chemberta-2-10M, chemberta-2-77M models share the same architecture. Thus, we fine-tune a single randomly initialized model for all of them (chemberta-2), selecting the average optimal hyperparameters across all three pre-trained models (*).

Model	RMSE		
	furan	**thiazole**	**thiophene**
molformer	0.596 (+0.031)	0.597 (+0.032)	0.594 (+0.029)
chemberta	0.641 (-0.034)	0.669 (-0.006)	0.667 (-0.008)
chemberta-2-5M	0.661 (-0.003)	0.631 (-0.033)	0.646 (-0.018)

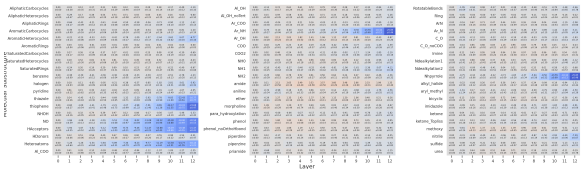
Table 12: Fine-tuning performance (RMSE, lower is better) of models further pre-trained on data containing **furan**, **thiazole**, **thiophene** on the lipophilicity dataset. green indicates improvement in RMSE (lower is better) while red denotes increased RMSE. Gray indicates negligible changes (<0.01) in RMSE.

Model	Layer	Δ macro F1 (PT)	Δ macro F1 (halogens)
chemberta-2-5M	0	0.000	0.000
	1	-0.209	-0.700
	2	0.015	0.499
	3	2.955	0.573
chemberta-2-10M	0	0.000	0.000
	1	-1.252	-2.341
	2	-1.395	-0.212
	3	2.407	-1.257
chemberta-2-77M	0	0.000	0.000
	1	-1.062	-2.606
	2	-0.419	-1.297
	3	4.381	-2.869

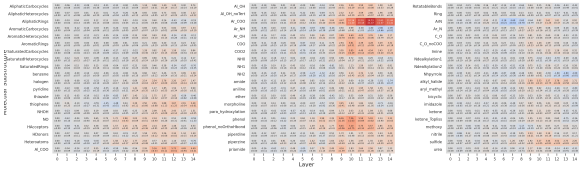
Table 13: **Further pretraining on halogens**: Layer-wise difference in probing performance on **halogens** for chemberta-2 models (left) and their counterparts further pre-trained on halogens (right) after fine-tuning.



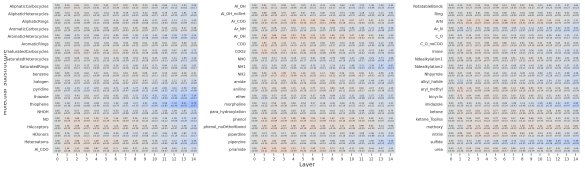
(a) Molformer (PT)



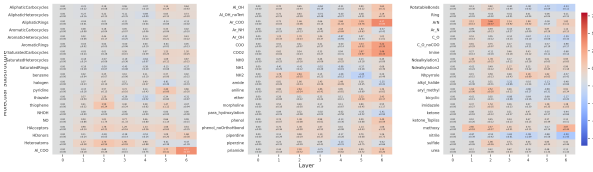
(b) Molformer(RI)



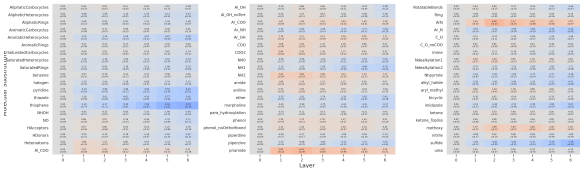
(c) Roberta-zinc-480m (PT)



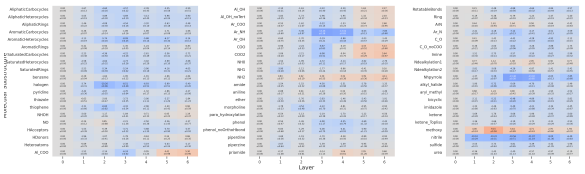
(d) Roberta-zinc-480m (RI)



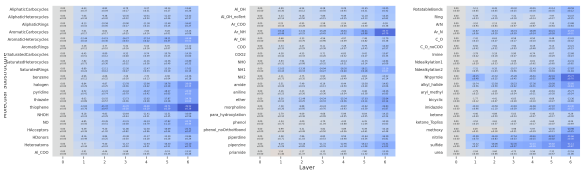
(e) Chemberta (PT)



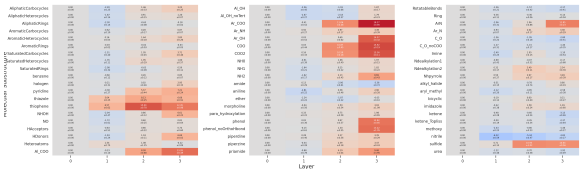
(f) Chemberta (RI)



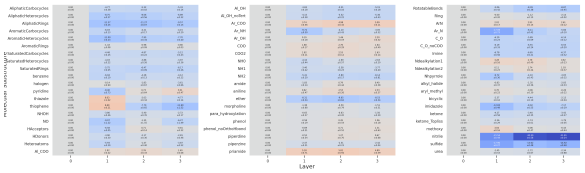
(g) Chemberta-base (PT)



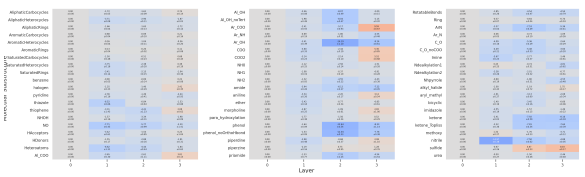
(h) Chemberta-base (RI)



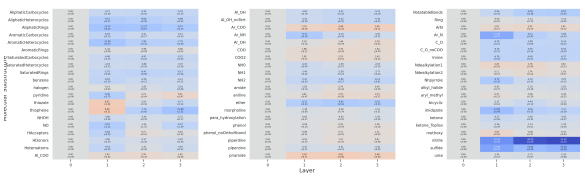
(i) Chemberta-2-5M (PT)



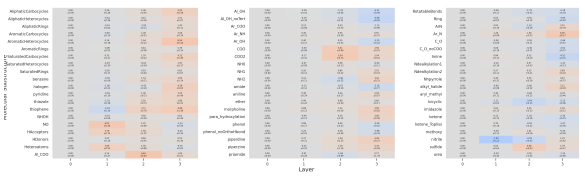
(j) Chemberta-2 (RI)



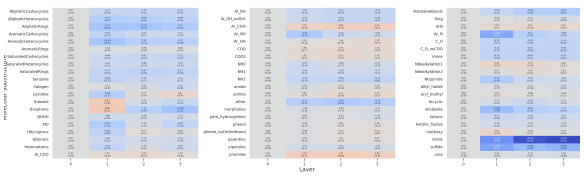
(k) Chemberta-2-10M (PT)



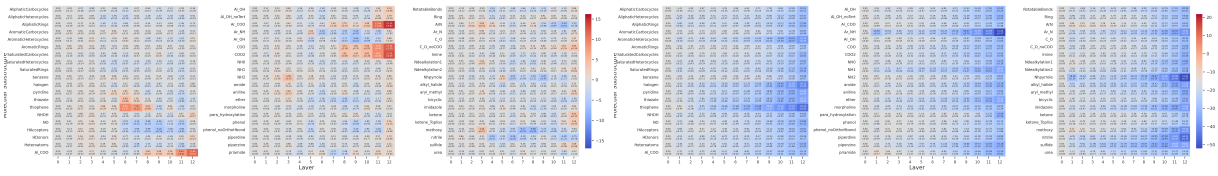
(l) Chemberta-2 (RI)



(m) Chemberta-2-77M (PT)



(n) Chemberta-2 (RI)



(o) Chemberta-3 (PT) (p) Chemberta-3 (RI)

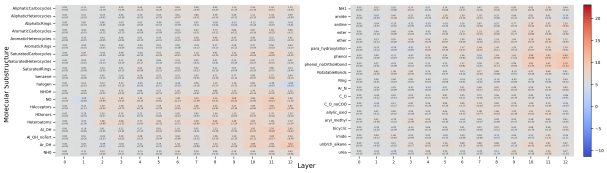
Figure 11: **Effect of fine-tuning on lipophilicity on molecular substructure encoding in CLMs**. We report the layer-wise difference in probing performance (macro-averaged F1, in %) after fine-tuning across 60 tasks (cf. Table 6). Figures 11j, 11l and 11n correspond to the same model architecture. PT denotes to the pre-trained models (left), while RI refers the same models but with randomly initialized weights (right). Improvements in probing performance after fine-tuning are shown in red, while degradation is indicated in blue.

Model	RMSE
`chemberta-2-5M`	0.646 (-0.018)
`chemberta-2-10M`	0.699 (+0.108)
`chemberta-2-77M`	0.635 (+0.003)

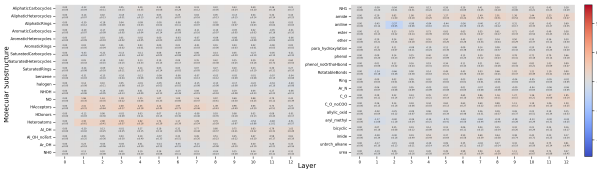
Table 14: Fine-tuning performance (RMSE, lower is better) of `chemberta-2-5M`, `chemberta-2-10M` and `chemberta-2-77M` further pre-trained on data containing **`halogens`**. green indicates improvement in RMSE (lower is better) while red denotes increased RMSE. Gray indicates negligible changes (<0.01) in RMSE.

Model	RMSE
`chemberta-2-5M`	0.622 (-0.042)
`chemberta-2-10M`	0.618 (+0.027)

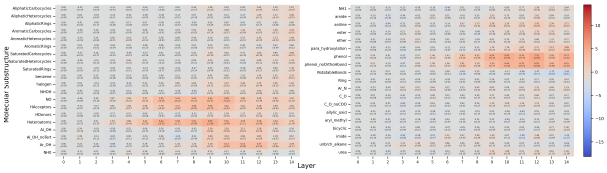
Table 15: Fine-tuning performance (RMSE, lower is better) of `chemberta-2-5M` and `chemberta-2-10M` further pre-trained on data containing **`phenol`**. green indicates improvement in RMSE (lower is better) while red denotes increased RMSE.



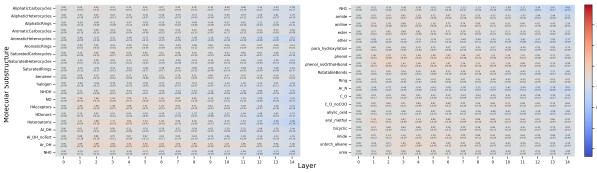
(a) Molformer (PT)



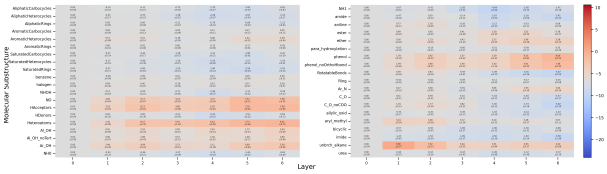
(b) Molformer(RI)



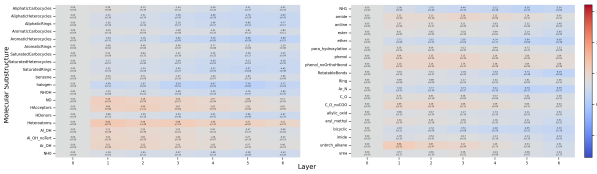
(c) Roberta-zinc-480m (PT)



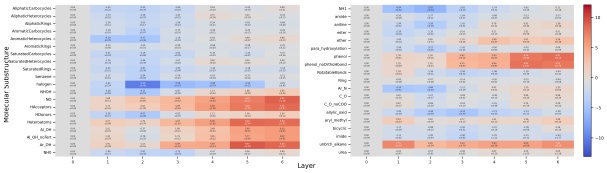
(d) Roberta-zinc-480m (RI)



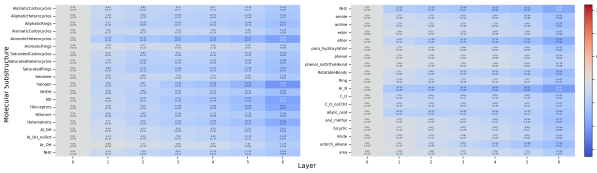
(e) Chemberta (PT)



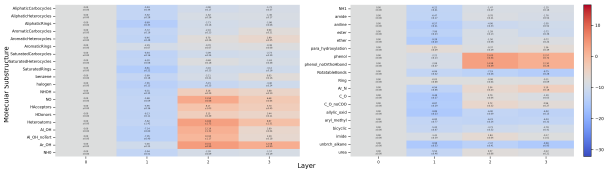
(f) Chemberta (RI)



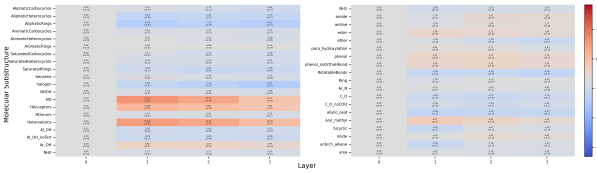
(g) Chemberta-base (PT)



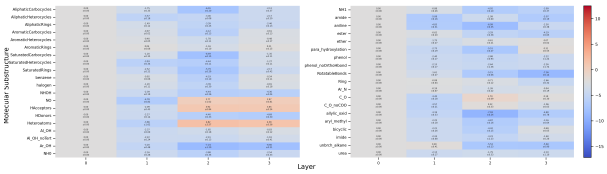
(h) Chemberta-base (RI)



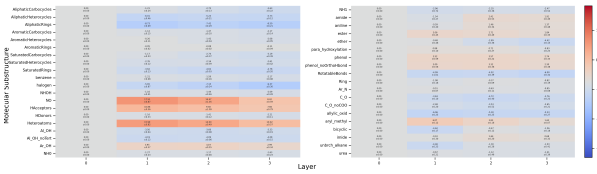
(i) Chemberta-2-5M (PT)



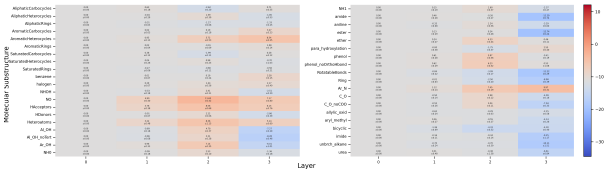
(j) Chemberta-2 (RI)



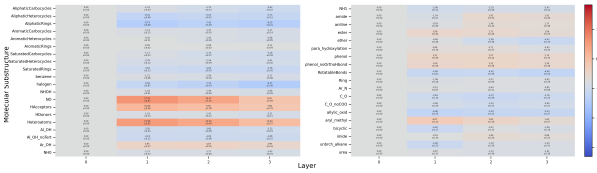
(k) Chemberta-2-10M (PT)



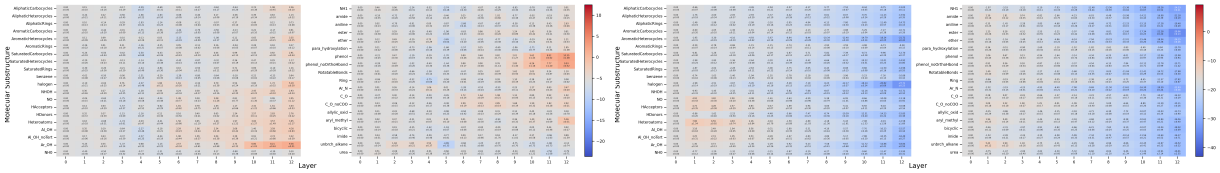
(l) Chemberta-2 (RI)



(m) Chemberta-2-77M (PT)



(n) Chemberta-2 (RI)



(o) Chemberta-3 (PT) (p) Chemberta-3 (RI)

Figure 12: **Effect of fine-tuning on solubility (ESOL) on molecular substructure encoding in CLMs**. We report the layer-rwise difference in probing performance (macro-averaged F1, in %) after fine-tuning across 39 tasks (cf. Table 7). Figures 11j, 11l and 11n correspond to the same model architecture. PT denotes to the pre-trained models (left), while RI refers the same models but with randomly initialized weights (right). Improvements in probing performance after fine-tuning are shown in red, while degradations are indicated in blue.

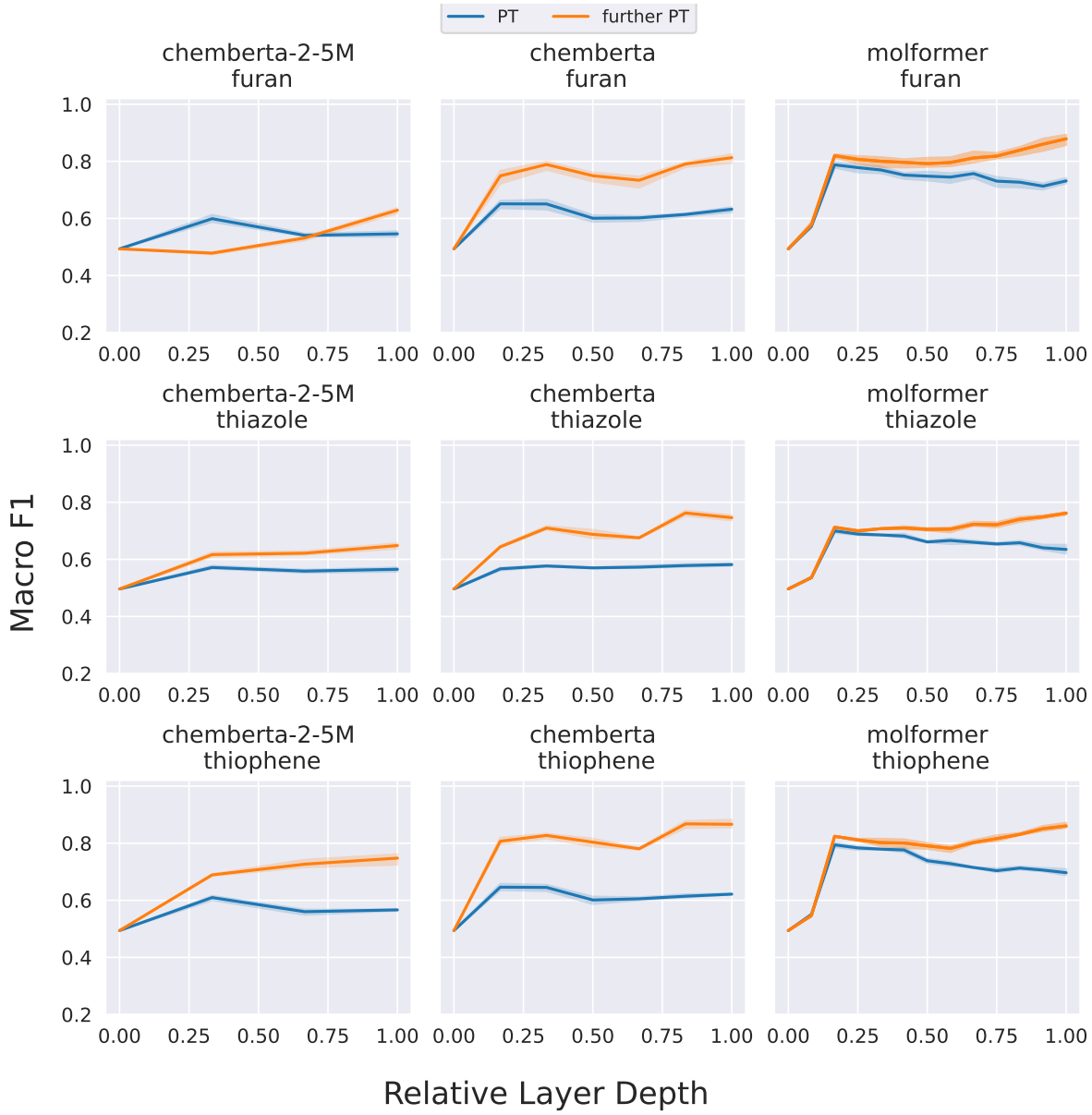


Figure 13: Probing performance of pre-trained models (PT) and those further pre-trained (further PT) on data containing `furan` (row 1), `thiazole` (row 2) and `thiophene` (row 3) on the respective substructure. Across all models, we observe an improvement in probing performance on the specific substructure after additional pre-training with data containing the corresponding substructure.

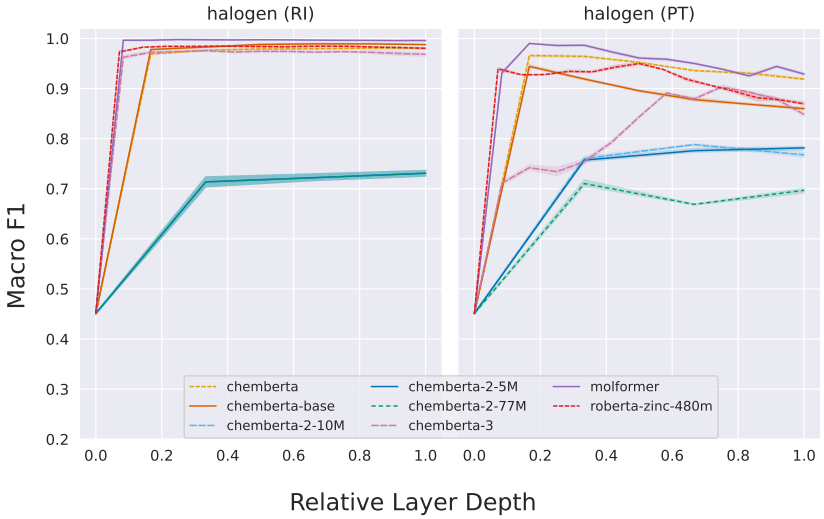


Figure 14: Layerwise probing performance (macro F1 score) on `halogens` for randomly initialized (RI) and pre-trained (PT) models. We observe that all models except for `chemberta-2-5M` and `chemberta-2-10M` unlearn halogens in the middle and upper layers.

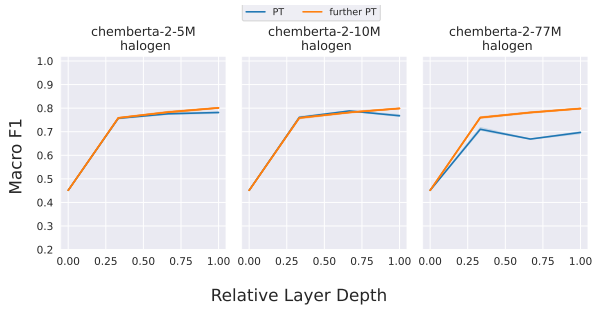


Figure 15: Probing performance of pre-trained models (PT) `chemberta-2-5M`, `chemberta-2-10M` and `chemberta-2-77M` and their counterparts further pre-trained (further PT) on data containing `halogens` on the `halogen` substructures.

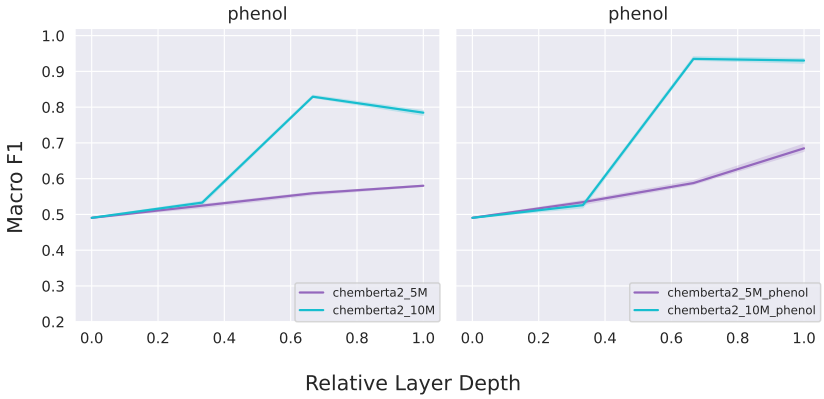


Figure 16: Probing performance of pre-trained models `chemberta-2-5M` and `chemberta-2-10M` (left) and the same models further pre-trained on data containing `phenols` (right) on `phenols`. We observe that while further pre-training on data containing phenol leads to better probing performance on the corresponding substructure, the performance gap between the two models does not close. We hypothesize that longer pre-training might narrow this gap.

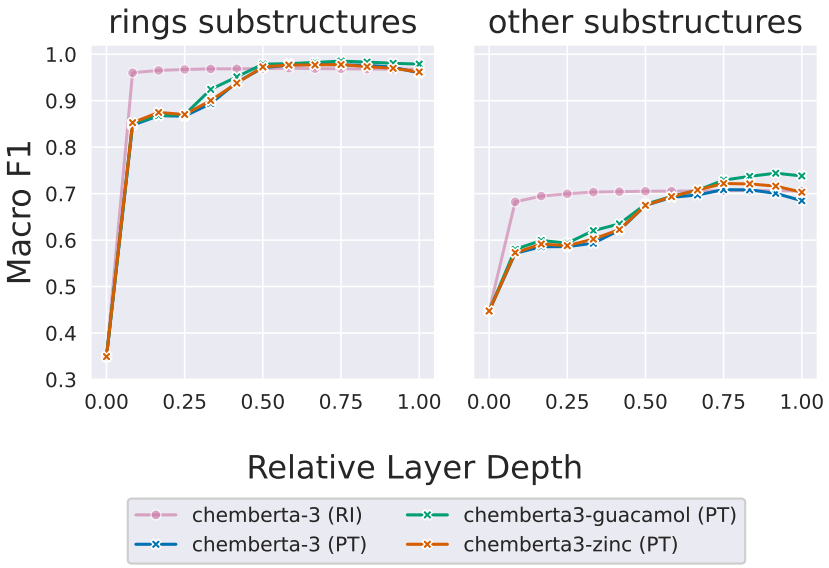


Figure 17: Effect of further pre-training `chemberta-3` on different data. `chemberta-3 (PT)` denotes the pre-trained model, `chemberta-3-zinc (PT)` is the model further pre-trained on a 100k subset of Guacamol, while `chemberta-3-guacamol (PT)` has been further pre-trained on a subset of ZINC-100M. `chemberta-3 (RI)` is the randomly initialized `chemberta-3`.

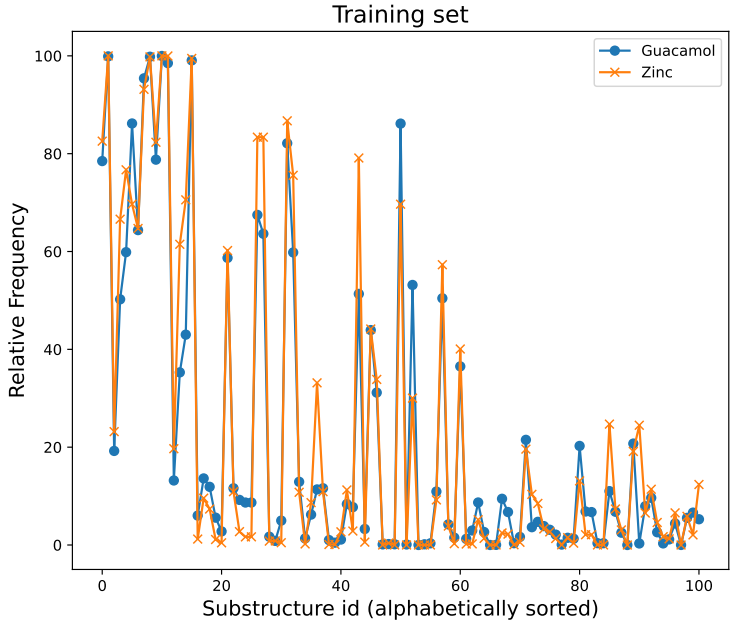


Figure 18: Comparison of relative frequency distributions of molecular substructures in the training splits of the subsampled ZINC and Guacamol datasets used for further pre-training `chemberta-3`.

Model	dim	MFP	RDFP	ATFP	TTFP	Avg
LR	300	0.84	1.02	0.88	0.90	0.91
	512	0.84	1.00	0.91	0.87	0.91
	1024	0.85	0.98	0.95	0.95	0.93
	2048	1.09	1.20	1.27	1.05	1.15
	Avg	0.91	1.05	1.00	0.94	-
SVM	300	0.72	0.94	0.75*	0.77	0.80
	512	0.71	0.89*	0.73*	0.73	0.77
	1024	0.68	0.75	0.68	0.70	0.70
	2048	0.67	0.72	**0.64***	0.69	0.68
	Avg	0.70	0.83	0.70	0.72	-
XGB	300	0.76	0.97	0.80	0.80	0.83
	512	0.73	0.91	0.76	0.80	0.80
	1024	0.72	0.83	0.74	0.74	0.76
	2048	0.72	0.79	0.69	0.73	0.73
	Avg	0.73	0.88	0.75	0.77	-
Avg	-	0.78	0.92	0.82	0.81	-

Table 16: Test set performance [RMSE (↓)] on **lipophilicity**. We evaluate four different methods for generating fingerprints (MFP, RDFP, ATFP, TTFP) and evaluate dimension sizes of 300, 512, 1024, and 2048. For the SVM, c=4 worked best except for the ones marked with * (c=16). Interestingly, we find that both XGB and SVM perform better for larger dimensions, while LR performs better for lower dimensions.

Model	dim	MFP	RDFP	ATFP	TTFP	Avg
LR	300	1.81	1.63	1.44	1.39	1.57
	512	2.06	1.93	2.11	1.87	1.99
	1024	4.74	6.87	367.63	11.18	97.61
	2048	2.53	2.61	5.84	4.03	3.75
	Avg	2.79	3.26	94.26	4.62	-
SVM	300	$1.19_{c=256}$	$1.02_{c=64}$	$0.93_{c=16}$	$0.99_{c=16}$	1.03
	512	$1.06_{c=256}$	$0.91_{c=256}$	$1.00_{c=4}$	$1.14_{c=4}$	1.03
	1024	$1.06_{c=64}$	$0.90_{c=1024}$	$0.92_{c=4}$	$1.20_{c=16}$	1.02
	2048	$1.00_{c=256}$	$0.88_{c=1024}$	$\mathbf{0.83}_{c=16}$	$1.11_{c=16}$	0.96
	Avg	1.08	0.93	0.92	1.11	-
XGB	300	1.23	1.16	1.01	1.08	1.12
	512	1.12	1.06	1.01	1.14	1.08
	1024	1.13	1.01	1.04	1.16	1.09
	2048	1.17	0.86	0.93	1.17	1.03
	Avg	1.16	1.02	1.00	1.13	-
Avg	-	1.66	1.74	32.06	2.29	-

Table 17: Test set performance [RMSE (↓)] on **ESOL** (solubility prediction). We evaluate four different methods for generating fingerprints (MFP, RDFP, ATFP, TTFP) and evaluate dimension sizes of 300, 512, 1024, and 2048. For the SVM, we report the best-performing c for each configuration. We find that both XGB and SVM maintain a robust performance around 1.00 across all dimensions, while LR performance varies substantially, especially for dim $= 1,024$.

Model	reasoning	basic	expl	hint	both
$SVM_{(c=16)}$ + ATFP			0.64		
molformer			0.57		
Llama-3.2-3B-Instruct	-	188.24	61.13	42.06	25.93
gpt-oss-20b$_{4,096}$	low	2.89	2.66	2.82	2.84
gpt-oss-20b$_{4,096}$	medium	3.11	2.80	**2.53**	2.94
gpt-oss-20b$_{4,096}$	high	237,981.69	246.42	19.95	98.97
gpt-oss-20b$_{8,192}$	high	139.36	43.04	27.67	24.60
gpt-oss-20b$_{16,384}$	high	61.84	376.55	337.18	15.45

Table 18: Test set performance [RMSE (↓)] on lipophilicity. We find that the gpt-oss-20b$_{4,096}$ model using the medium reasoning level together with hints performs best out of all LLMs and that providing explanations as well as hints can improve the performance. Nonetheless, all LLMs are substantially outperformed by the SVM using fingerprints and the molformer model.

Model	reasoning	basic	expl	hint	both
$SVM_{(c=16)}$ + ATFP			0.83		
molformer			0.59		
Llama-3.2-3B-Instruct	-	233.11	213.79	477.07	940,720,868.84
gpt-oss-20b$_{4,096}$	low	9.27	31.97	9.51	**8.96**
gpt-oss-20b$_{4,096}$	medium	48.95	51.68	14.11	23.31
gpt-oss-20b$_{4,096}$	high	193.50	41.91	95.61	36.32
gpt-oss-20b$_{8,192}$	high	23.53	44.42	42.34	193.03
gpt-oss-20b$_{16,384}$	high	32.94	83.57	19.39	60.72

Table 19: Test set performance [RMSE (↓)] on esol (solubility prediction). We find that the gpt-oss-20b$_{4,096}$ model using the low reasoning level and both (hints and explanations) performs best. Interestingly, we find that especially for larger contexts, providing explanations or hints can deteriorate the performance. Again, all LLMs are substantially outperformed by the SVM using fingerprints and the molformer model.

We need to predict logD value (probably logD at physiological pH, e.g., logD7.4). This is a computational estimate. We are given a SMILES: C#Cc1cccc(Nc2nc3cc(C(=O)O)ccc3c3cncnc23)c1

Let's parse: It's a heteroaromatic system fused? Actually: C#C-? Let's analyze. The base skeleton: phenyl ring (c1cccc...) with a -C#C substituent at position 1? The SMILES: C#C c1 c c c c (N c2 n c3 c c (C(=O)O) c c c3 c3 c n c n c2 3) c1. So there is a phenyl ring with a -C#C group (propyne). At the para position (maybe 4) there is an anilide linking to a bicyclic heteroaromatic (imidazo[...])? Actually c2nc3cc(C(=O)O)ccc3c3cncnc23. That seems to be a fused bicyclic heteroaromatic containing two rings: a benzodiazine? It appears to be a quinazoline ring (a benzene fused with a diazine). Quinazoline: fused benzene and pyrimidine. So we have a 4-? It attaches via an aniline NH to the quinazoline ring.

Also there is a carboxylic acid group (C(=O)O) attached to the quinazoline ring.

The entire molecule contains one carboxylic acid (pKa 4.5). At physiological pH, it will be deprotonated (anionic). This will significantly reduce logD. Also there is one NH (aniline). There are hetero nitrogens: 4? In quinazoline ring: two ring nitrogens (positions 1 and 3). Additional ring may have two more? Wait: c3cncnc23: That is a fused pyrimidine? Actually c3 c n c n c 2 3. So ring 3 has two nitrogens. So total hetero atoms: 4 ring nitrogens, 1 amine N, 1 carboxylate O's (2). So 5 nitrogens, 3 oxygens. So high hetero content.

Also there's a nitrile (C#C) group: actually it's an alkyne (C#C) group attached to phenyl; no hetero.

So logP (neutral) might be moderate-high due to aromatic rings ($\sim$3?). But logD7.4 will be lower due to deprotonated carboxylate (negative). Many hetero nitrogens may also be basic? In quinazoline, the ring nitrogens are not protonated at physiological? They are aromatic; they can act as weak bases (pKa 2-3). So mostly neutral. The aniline NH is weakly basic; likely neutral.

So overall, the logD7.4 may be around 1.5? Actually carboxylate reduces lipophilicity drastically. For example, carboxylate-containing molecules like ibuprofen logD7.4 $\sim$3.5? Wait ibuprofen is acid (pKa 4.5). Ibuprofen has logP 3.5, logD7.4 $\sim$3.5? Actually at neutral, logD decreases a bit; but still high due to hydrophobic. But our molecule has 3 aromatic rings (two fused hetero aromatic + phenyl). That gives high aromatic surface. Carboxylate reduces logP by $\sim$2 units maybe. For a 3-ring acid, logD7.4 maybe $\sim$2.5-3.0. Example: diclofenac (two phenyl rings + acid) has logP 4.5, logD7.4 4.0? Actually diclofenac logP 4.5, logD7.4 4.0? Not sure. Another acid with hetero: warfarin logP 3.1, logD7.4 $\sim$2.5. So our compound may be similar to warfarin.

Thus I'd guess logD7.4 $\sim$2.0-2.5. Let's pick 2.3.

But maybe due to hetero nitrogens, more hydrophilic, so logD lower $\sim$1.8.

Let's compute approximate cLogP using fragments: Each phenyl ring $\sim$2.3; hetero ring reduces; carboxylate reduces $\sim$1.5. Hard.

I'll answer: logD $\approx$ 1.9.

Given the question: P̈redict the lipophilicity the following molecule.S̈o I should provide a single numeric value. I'll answer 2.0 ($\pm$0.5). Provide explanation.

Let's produce final.assistantfinal**Predicted logD (pH $\approx$ 7.4)** $\approx$ **2.0**

The molecule contains three aromatic rings (one phenyl and one fused benzodiazine) that give it a relatively high baseline lipophilicity, but the presence of a carboxylic acid (which will be largely anionic at physiological pH) and several hetero nitrogen atoms dramatically reduces its distribution into the octanol phase. A logD value around 2.0 is consistent with similar tricyclic heteroaromatic acids (e.g., warfarin, diclofenac-like scaffolds) when measured at neutral to slightly basic conditions.

Figure 19: Response of `gpt-oss-20b`$_{4,096}$ with medium reasoning level and preprocessed functional groups as hints. While the model seemingly "reasons" about possible implications of different molecular subgroups, it predicts a logD value of 2.0 while the true value lies at -1.067.