ChemBERTa-2: Fine-Tuning for Molecule's HIV Replication Inhibition Prediction

Sylwia Nowakowska*[a b]

Two versions of Large Language ChemBERTa-2 models, pre-trained with two different methods, were fine-tuned in this work for HIV replication inhibition prediction. The best model achieved AUROC of 0.793. The changes in distributions of molecular embeddings prior to and following fine-tuning reveal models' enhanced ability to differentiate between active and inactive HIV molecules.

AI-based tools hold considerable promise in improving and accelerating the drug discovery process[1–3]. Among the various applications in that field, these tools can be employed to predict key molecular properties[4–11] . To facilitate model comparison and benchmarking, MoleculeNet datasets have been released[12]. The collection contains datasets for the molecular properties' prediction on different levels: quantum, physical chemistry, biophysics, and physiology. Among many AI approaches, Graph Neural Networks (GNNs) have been used for such tasks, due to their ability to encode the inherent graph-like nature of molecules, enabling the incorporation of spatial and contextual information[13–16].

An alternative method for representing molecular structures is a string-based SMILES notation, which can serve as a direct input to Large Language Models (LLMs)[8,17]. This approach was used among others by Chithrananda *et al.*, who pre-trained RoBERTa transformer[18] on 10M SMILES curated from PubChem with the use of the masked-language modelling (MLM) method. The pre-trained model, ChemBERTa, was fine-tuned for the prediction of molecule toxicity, blood-brain barrier penetration (BBBP), and HIV replication inhibition on the MoleculeNet datasets, achieving promising results, with AUROC for the HIV inhibition task of 0.622. In the later work, the authors pre-trained the RoBERTa transformer on bigger dataset containing 77M SMILES from PubChem with the use of either MLM or multi-task regression (MTR) method. During MLM pre-training 15% of the tokens in each input SMILES string was masked and the model was trained to identify them correctly. The MTR pre-training relied on simultaneous prediction of 200 molecular properties. The obtained ChemBERTa-2 model variations were fine-tuned for the downstream tasks of predicting various molecular properties, achieving competitive outcomes with

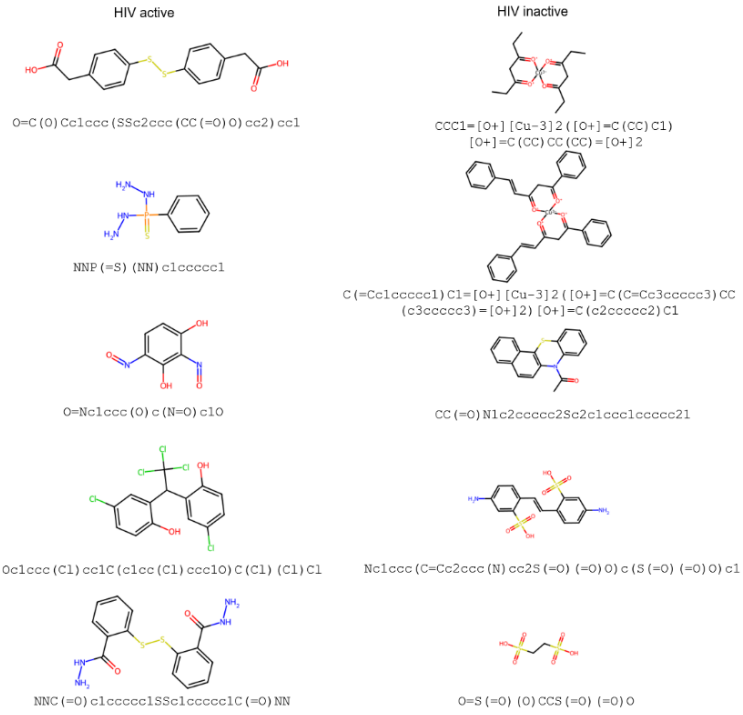


Figure 1) Examples of molecules from the HIV dataset with the corresponding SMILES representation.

state-of-the-art architectures. In that context, the authors reported that the models pre-trained with MTR tended to perform better highlighting the criticality of comprehensive

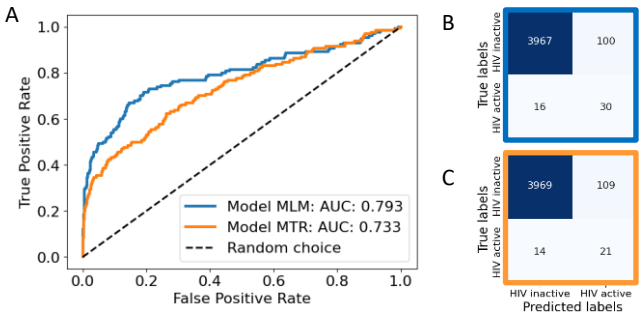


Figure 2) The comparison of performance of ChemBERTa-2 model pre-trained with either MLM or MTR after fine-tuning on HIV inhibition prediction task A) receiver operating characteristics (ROC) curves with reported area under the curve (AUC); B-C) corresponding confusion matrices.

[a] *University of Zurich, Rämistrasse 71, 8006 Zürich, Switzerland.*
[b] *University Hospital Zurich, Rämistrasse 100, 8006 Zürich, Switzerland.*

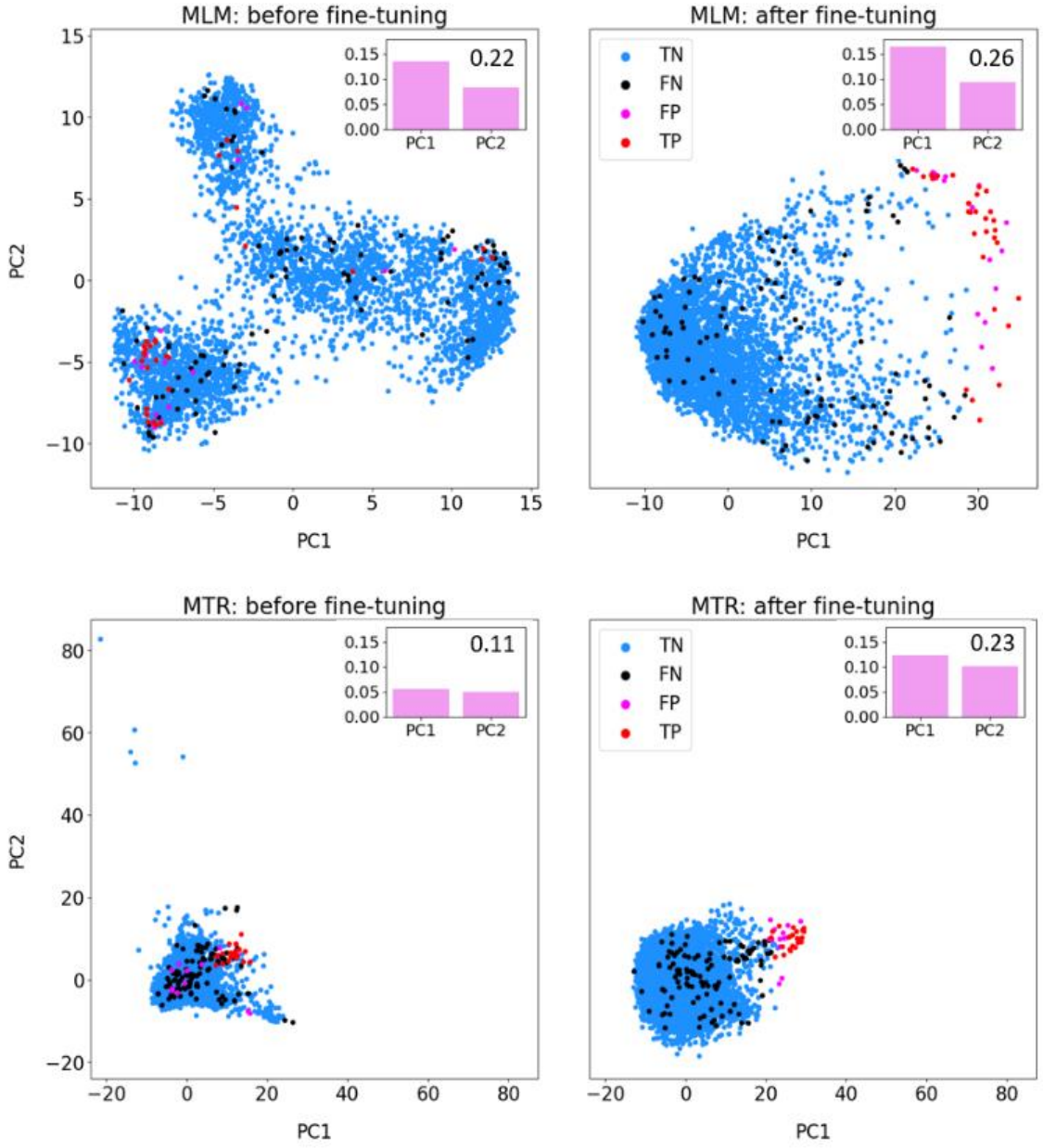
assessment of pre-training approaches. The performance of the HIV inhibition task was not reported in that work[5]. The MTR pre-trained ChemBERTa-2 model was also fine-tuned for aqueous solubility prediction achieving comparable results with the published models[6].

In the view of the importance of the pre-training method on the models' performance on the downstream task after fine-tuning, the present study puts an emphasis on the evaluation of the MLM/MTR approaches using ChemBERTa-2 models for HIV inhibition prediction.

In this work, the benchmark MoleculeNet HIV dataset containing 39684 inactive and 1443 active molecules was used. Examples of molecules drawn with RDKit‡ are shown in Fig. 1.

The default training/validation/test 80/10/10 scaffold splitting from DeepChem library was utilized[19]. The ChemBERTa-2-77M-MLM and ChemBERTa-2-77M-MTR from Hugging Face[20] were fine-tuned using T4 GPU on Google Colab using the simpletransformers library[21]. To compensate for the class imbalance, automatic weight balancing was used. To prevent overfitting, early stopping based on validation loss was used during hyperparameter tuning. Each model was evaluated on the holdout test set with AUROC and confusion matrix. The code for the model training is released in a GitHub repository§.

In Fig. 2 the performance of best MLM and MTR models after hyperparameter tuning is illustrated. The MLM model exhibited



Fi **ning.** In the in

better performance characterized by AUROC of 0.790, while the MTR scored 0.733.

For deeper understanding of the pre-training approach, the embeddings of the molecules from the test set were computed for the MLM and MTR models. These embeddings were derived from the final hidden layer, both before and after the fine-tuning process. As each molecule's representation consisted of 384 embeddings, Principal Component Analysis (PCA) was performed subsequently for dimensionality reduction. The obtained representations in 2D latent space together with the explained variance ratio are illustrated in Fig. 3. In case of the MLM model prior to fine-tuning, the representations are dispersed randomly within the latent space. However, after fine-tuning, the representations corresponding to molecules classified by the model as HIV inactive exhibit lower values of PC1 (cf. light and dark blue points), while representations of molecules classified as HIV active are having higher PC1 values (cf. magenta and red points). In the case of the MTR model prior to fine-tuning, molecules correctly classified as HIV active tend to cluster together in their representations. On the other hand, the remaining points are scattered randomly with several outliers. Following fine-tuning, the distribution of representations becomes similar to that of the MLM model. The changes in the distributions of the embeddings before and after fine-tuning indicate that the models become more adept at distinguishing between HIV active and inactive molecules.

The outcomes of this study align with the notion that transformer models hold significant promise for predicting molecular properties. Zhu *et al.* showcased an effective strategy employing the RoBERTa transformer architecture integrated with a Graph Neural Network (GNN)[7]. This approach enabled inputting to the model two distinct molecular representations - the chemical structure and SMILES notation - concurrently during the pre-training phase, a concept termed dual-view pre-training. The training objective for the transformer branch was to predict masked token, *i.e.,* MLM method. Moreover, a third objective was to maximize the consistency between the two high-level representations produced by each branch. After pre-training, the fine-tuned transformer branch achieved state-of-the are performance on different molecular property prediction tasks, with HIV activity precision task characterized by AUROC of 0.810[22]. Further research into impact of various pre-training methods on molecular property prediction in downstream task, employing either the transformer alone or in conjunction with GNN has the potential to drive additional progress in these domains. This could result in the development of robust models capable of revolutionizing the drug discovery process.

In conclusion, the analysis presented in this study showcases how both pre-training methods of the RoBERTa architecture, namely MLM and MTR, empower the models to grasp significant representations of molecular structures. Notably, the MLM model attains a performance level nearly on par with the current best results in predicting HIV activity in a downstream task.

Conflicts of interest

There are no conflicts to declare.

Notes and references

‡ RDKit: Open-source cheminformatics. https://www.rdkit.org

§https://github.com/SylwiaNowakowska/LLM_Fine_Tuning_Molecular_Properties

1 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat Rev Drug Discov*, 2019, **18**, 463–477.

2 P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos and C. Fernandez-Lozano, *Computational and Structural Biotechnology Journal*, 2021, **19**, 4538–4558.

3 P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann and G. Schneider, *Nat Rev Drug Discov*, 2020, **19**, 353–364.

4 S. Chithrananda, G. Grand and B. Ramsundar, 2020.

5 W. Ahmad, E. Simon, S. Chithrananda, G. Grand and B. Ramsundar, 2022.

6 A. S. Lang, *ACSR*, , DOI:10.31031/ACSR.2023.04.000578.

7 J. Zhu, Y. Xia, T. Qin, W. Zhou, H. Li and T.-Y. Liu, 2021.

8 R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez and R. Stojnic, 2022.

9 J. Born, G. Markert, N. Janakarajan, T. B. Kimber, A. Volkamer, M. R. Martínez and M. Manica, *Digital Discovery*, 2023, **2**, 674–691.

10 J. Chen, Y.-W. Si, C.-W. Un and S. W. I. Siu, *Journal of Cheminformatics*, 2021, **13**, 93.

11 D. Zhuang and A. K. Ibrahim, *Applied Sciences*, 2021, **11**, 7772.

12 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.

13 S. Zhu, B. H. Nguyen, Y. Xia, K. Frost, S. Xie, V. Viswanathan and J. A. Smith, *Green Chem.*, 2023, **25**, 6612–6617.

14 Y. Wu, C. Zhang, L. Wang and H. Duan, *Chem. Commun.*, 2021, **57**, 4114–4117.

15 Z. Yang, M. Chakraborty and A. D. White, *Chem. Sci.*, 2021, **12**, 10802–10809.

16 O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel and T. Langer, *Drug Discovery Today: Technologies*, 2020, **37**, 1–12.

17 C. Qian, H. Tang, Z. Yang, H. Liang and Y. Liu, 2023.

18 Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, 2019.

19 B. Ramsundar, P. Eastman, P. Walters and V. Pande, *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*, O'Reilly Media, Sebastopol, CA, 1st edition., 2019.

20 DeepChem (DeepChem), https://huggingface.co/DeepChem, (accessed August 25, 2023).

21 T. Rajapakse, Simple Transformers, https://simpletransformers.ai/, (accessed August 28, 2023).

22 Papers with Code - HIV dataset Benchmark (Molecular Property Prediction), https://paperswithcode.com/sota/molecular-property-prediction-on-hiv-dataset, (accessed August 25, 2023).