Do chemical language models provide a better compound representation?

Mirko Torrisi
Bristol Myers Squibb
mirko.torrisi@bms.com

Saeid Asadollahi
Bristol Myers Squibb
saeid.asadollahi@bms.comm

Antonio de la Vega de León
Bristol Myers Squibb
antonio.delavegadeleon@bms.comm

Kai Wang
Bristol Myers Squibb
kai.wang1@bms.com

Wilbert Copeland
Bristol Myers Squibb
wilbert.copeland@bms.com

Abstract

In recent years, several chemical language models have been developed, inspired by the success of protein language models and advancements in natural language processing. In this study, we explore whether pre-training a chemical language model on billion-scale compound datasets, such as Enamine and ZINC20, can lead to improved compound representation in the drug space. We compare the learned representations of these models with the de facto standard compound representation, and evaluate their potential application in drug discovery and development by benchmarking them on biophysics, physiology, and physical chemistry datasets. Our findings suggest that the conventional masked language modeling approach on these extensive pre-training datasets is insufficient in enhancing compound representations. This highlights the need for additional physicochemical inductive bias in the modeling beyond scaling the dataset size.

1 Introduction

Language models have revolutionized the field of natural language processing, and protein representation [11, 17, 22]. This has stimulated the development of numerous chemical language models (CLMs) [5, 8, 10, 26], on increasingly large portions of public compound databases, such as ChEMBL [16] and ZINC20 [9]. CLMs are pre-trained using masked language modeling, and then fine-tuned on down-stream tasks of interest, often resulting in state-of-the-art predictive performance [2, 23].

The availability of multi-billion scale compound databases, such as ZINC20 [9] and Enamine [21], prompted us to investigate whether pre-training a CLM on billion-scale compound datasets can suffice to outperform the de facto standard compound representation for drug discovery, i.e., Extended-Connectivity Fingerprints (ECFP) [19]. Additionally, this study is the first to investigate the potential impact of using the Enamine REAL Database for pre-training CLMs, as most recent studies have focused on exploiting ZINC20 or ChEMBL.

Our investigation aims to explore the effectiveness of pre-training CLMs on various compound databases of up to 2 billion compounds. We evaluate whether scaling the pre-training dataset can improve the ability of CLMs to represent the drug space compared to ECFP. To achieve this, we

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

Table 1: SELFIES symbols shared by pre-training and benchmark datasets.

	Pre-training datasets				
Name	ChEMBL	Enamine	ZINC20	All	*Tot*
MoleculeNet	147	60	108	56	*224*
Papyrus1k	71	35	60	35	*73*
Tot	*341*	*67*	*277*		

Table 2: Pre-training datasets: number of compounds per dataset split.

	Dataset Split			
Name	Training	Validation	Test	Tot
ChEMBL-2M Enamine-2M ZINC20-2M	2,348,552	11,861	11,862	2,372,275
Enamine-2B ZINC20-2B	1,845,171,532	9,319,049	9,319,049	1,863,809,630

benchmark the predictive performance of these representations on biophysics, physiology, and physical chemistry datasets. By conducting these experiments, we aim to determine whether CLMs can provide a richer representation of the drug space than ECFP without fine-tuning.

2 Methods

2.1 Datasets standardization

All compounds in both pre-training and benchmarking datasets are originally represented using SMILES. We standardize all SMILES using DataMol 0.10 [15], and encode them as SELFIES using selfies 2.1.1 [12]. We opt for using SELFIES to facilitate the learning process, since it was specifically developed for generative modeling and provides a guarantee of molecular validity. We remove all compounds that do not pass the standardization procedure, or are not encodable as SELFIES, or are found to be duplicates.

2.1.1 Pre-training datasets

The pre-training datasets in this study are derived from 3 public compound databases: ChEMBL 33 [16], Enamine REAL Database [21], and ZINC20 [9]. After being standardized, ChEMBL, Enamine, and ZINC20 are described by 341, 67, and 277 distinct SELFIES symbols, respectively (see Table 1).

The standardized databases are split assigning 99%, 0.5%, and 0.5% of the compounds to the training, validation, and test set, respectively. We obtain ChEMBL-2M splitting ChEMBL at random. Enamine-2M and ZINC20-2M are randomly sampled from Enamine and ZINC20, respectively, while keeping the same amount of compounds in their training, validation, and test sets as in ChEMBL-2M.

Similarly, we obtain ZINC20-2B splitting ZINC20 at random, and Enamine-2B sampling from Enamine at random (i.e., about $1.9B$ compounds in total). The size of each dataset split is reported in Table 2.

2.1.2 Papyrus1k

Papyrus1k is derived from Papyrus, a public database consisting of nearly $60M$ compound-protein pairs, $1.2M$ of which are considered of high quality, i.e., "representing exact bioactivity values measured and associated with a single protein or complex subunit" [4]. We focus on these high quality pairs, and keep only those involving proteins observable in at least $1k$ bioactivities. We binarize all bioactivities by setting a pIC_{50} threshold value of 6.

Table 3: Number of unique compounds and tasks in the benchmark datasets, and their sparsity.

Name	Tasks	Compounds	Sparsity
Papyrus1k	280	489,402	99%
	Physiology		
BBBP	1	1,885	0%
Tox21	12	7,586	10%
ToxCast	379	7,653	65%
SIDER	27	1,280	0%
ClinTox	2	845	0%
	Physical chemistry		
ESOL	1	1,117	0%
FreeSolv	1	642	0%
Lipophilicity	1	4,189	0%

After standardizing and encoding all compounds as SELFIES, we find 73 distinct SELFIES symbols in Papyrus1k. Out of these, 35 symbols are found in all pre-training datasets (see Table 1). Thus, we remove 1,698 compounds (4,031 bioactivities) containing SELFIES symbols not found in at least one pre-training dataset. We randomly split compounds per protein, assigning 80%, 10%, and 10% for training, testing, and validation, respectively. The final benchmark dataset contains 489,402 compounds, 280 proteins, and 881,081 bioactivities (see Table 3).

2.1.3 MoleculeNet

MoleculeNet [25] is a collection of benchmark datasets for assessing machine learning methods for characterizing compounds. To expand into more areas related to drug discovery and development, we focus on two categories of the datasets in MoleculeNet: physiology and physical chemistry. We do not include the biophysics datasets because of their similarity to Papyrus1k, and the quantum mechanism datasets as we feel it is less related to drug discovery. The physical chemistry datasets are the only regression datasets in this study.

After standardizing, removing duplicates, and encoding the compounds in these datasets as SELFIES, we find 224 distinct SELFIES symbols. Out of these, only 56 are in the pre-training datasets (see Table 1), yet describe 17,329 compounds (25,197 pairs). We, thus, remove 1,787 compounds (3,275 pairs) for containing SELFIES symbols not found in at least one pre-training dataset. We also remove 3 compounds for requiring more than 512 symbols, i.e., the maximum supported length by our CLMs. The composition of the final benchmark datasets, and their sparsity, is in Table 3.

For ToxCast, i.e., one of the physiology datasets, we keep only 379 tasks (out 617) associated to at least $1k$ compounds. For each dataset, we randomly split compounds per task, assigning 80%, 10%, and 10% for training, testing, and validation, respectively.

2.2 Chemical language modeling

To construct the character-based tokenizer and the CLM architecture, specifically Bert-Large [6], we employ Transformers 4.28.1 [24]. We opt for Bert-Large as our model architecture because it is recognized as a ubiquitous baseline for masked language modeling [18]. We utilize Lightning 2.0.1 [7] to facilitate multi-node training on 2 NVIDIA DGX-A100 machines.

We pre-train all CLMs with masked language modeling (15%), Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-8$), learning rate of $5e-5$, weight decay set to 0.1, and a cosine scheduler. We set the warm up steps to $1k$ and the epoch to 1 for ZINC20-2B and Enamine-2B. We reduce the warm up steps to 100 and increase the epochs to 5 for ChEMBL-2M, Enamine-2M, and ZINC20-2M.

2.3 Benchmark modeling

We utilize Scikit-learn 1.2.2 [3] to train either a Random Forest (RF) classifier or regressor for each task in the benchmark datasets. Each RF is trained on either ECFP or the average pool of the last hidden layer of a CLM, i.e., the CLM embedding.

To generate traditional fingerprints, we employ the Morgan fingerprints algorithm in RDKit 2022.9.5 [13] with $2,048$ bits and radius 0, 1, or 2, corresponding to ECFP0, ECFP2, and ECFP4, respectively. For generating CLM embeddings, we use the final model checkpoint for each CLM.

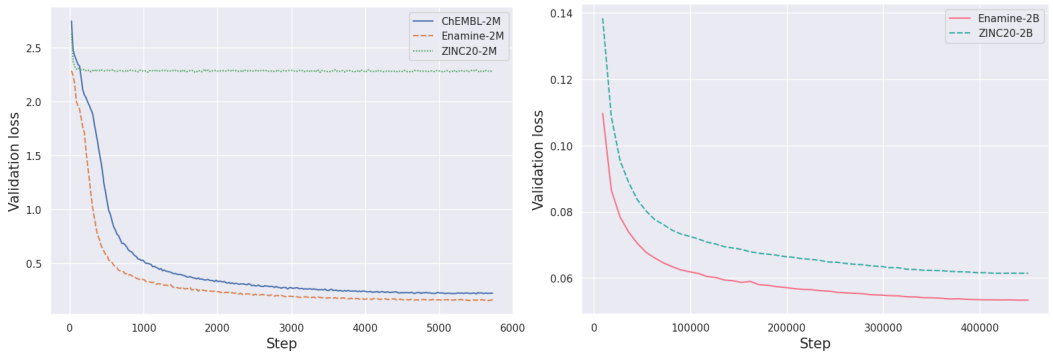
To evaluate 30 combinations of hyperparameters, we conduct a Bayesian hyperparameter optimization using Optuna [1]. We guide the optimization process using either the Matthew correlation coefficient (MCC), or the root-mean-square error (RMSE) as the evaluation metric. Specifically, we let Optuna suggests the number of estimators (ranging from 100 to 5k), the depth (ranging from 5 to 100), and the number of features to consider ("sqrt" and "log2").

3 Results

We successfully pre-train a CLM model for each pre-training dataset, except for ZINC20-2M, which did not converge. It is worth noting that while all validation losses approach zero, the losses of Enamine-2B and ZINC20-2B, i.e., the CLMs pre-trained on $2B$ compounds, are lower yet continue to decrease (see Figure 1). Previous studies have used similar validation loss values as an indication of effective pre-training [20, 14].

In the following 2 sections, we present benchmark results for CLM embeddings obtained through pre-training on ChEMBL, Enamine, and ZINC20. Surprisingly, we find that CLM embeddings perform similarly, with no clearly noticeable difference scaling the pre-training dataset from $2M$ to $2B$ compounds. We also find that CLM embeddings perform worse than ECFP4 on Papyrus1k, our benchmark dataset of high quality bioactivities (see Section 3.1). Finally, we observe more on par performance on MoleculeNet (see Section 3.2).

To further investigate why CLMs do not provide a better compound representation than ECFP4, we report results for ECFPs with smaller radius. A ECFP with radius 0 represents the initial atom identifier and simple physicochemical properties of that atom (ECFP0), with radius 1 includes the information of the immediate neighbors (ECFP2), and with radius 2 encompasses the neighbors up to 2 bonds away (ECFP4). Notably, we find that ECFP0 obtains comparable predictive performance to CLMs on both Papyrus1k and MoleculeNet. We also observe that larger radius ECFPs achieves superior predictive performance, especially on Papyrus1k. This suggests that CLMs may not capture the topological information that is beneficial for biophysical applications, which is represented by larger radius ECFPs. Analogous observations can be made on MoleculeNet, although a larger radius seems to be not always advantageous.



(a) Pre-training losses for the 2M compounds regime. (b) Pre-training losses for the 2B compounds regime.

Figure 1: Pre-training validation losses for (a) 2M compounds regime, and (b) 2B compounds regime. The pre-training for (a) lasts for 5 epochs, while for (b) it only lasts for one epoch.

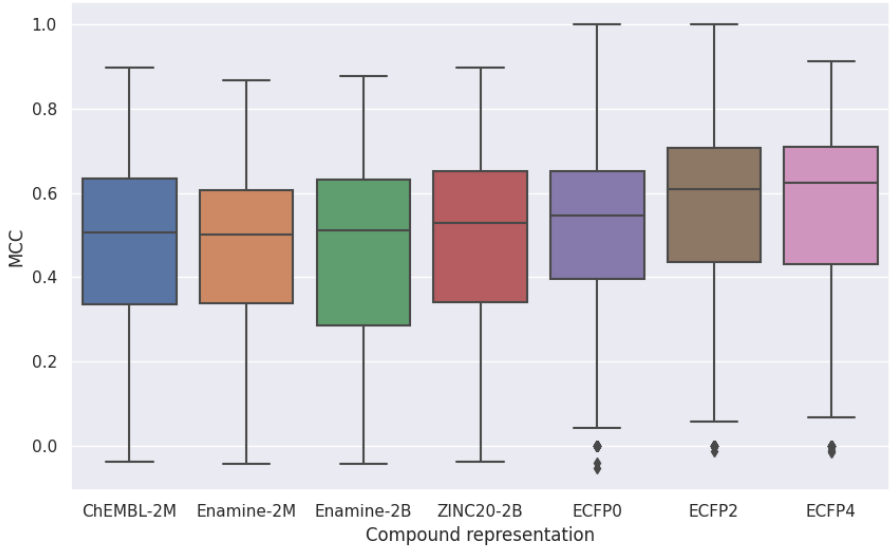


Figure 2: MCC on the high quality test set of Papyrus1k. Each box plot represents the distribution of individual MCC values for the test set of all proteins in the Papyrus1k dataset. ECFP0 seems comparable to CLM embeddings, while ECFP2 seems to provide sufficient biophysics information.

3.1 Biophysics benchmark

The benchmark results on Papyrus1k show similar results for all compound representations, with ECFP2 and ECFP4 showing an advantage over both ECFP0 and CLM embeddings (see Figure 2). Interestingly, ECFP0 seems to perform on par with CLM embeddings, with a Pearson correlation coefficient of at least 0.86 (see Figure 3a). This may suggest that CLMs are not capturing topological information, and that neighbors up to 1 bond away may provide sufficient biophysics information.

Moreover, CLMs show an even greater correlation among themselves (see Figure 3b), with no clearly noticeable difference between CLMs pre-trained on $2M$ or $2B$ compounds. This finding may suggest that masked language modeling of Bert-Large on $2M$ compounds suffice to capture atomic identifiers, but may not capture additional valuable insights.

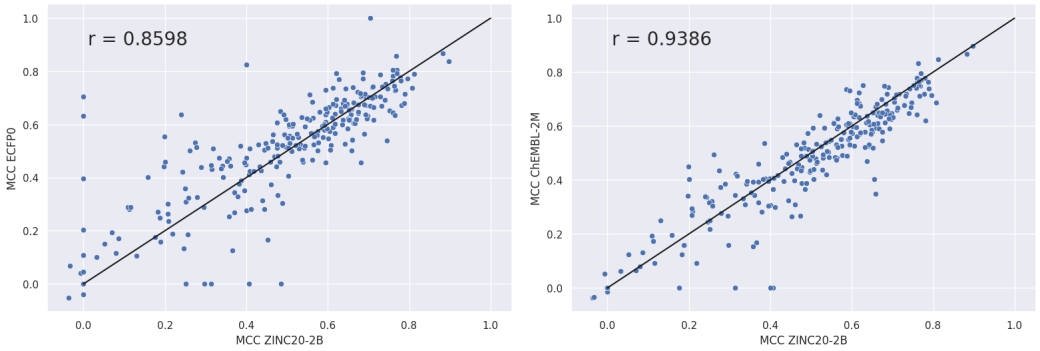
Finally, removing from the pre-training set of ChEMBL-2M the compounds in the validation and test set of Papyrus1k does not appear to affect the benchmark results (results not shown).

The RF hyperparameters and validation MCC are reported in the Supplementary material (see Figure 5).

3.2 Physiology and physical chemistry benchmark

The predictive performance of ECFPs on MoleculeNet is somewhat comparable to CLM embeddings, with a less clear benefit from neighbor information compared to Papyrus1k (see Figure 4). Moreover, Enamine-2B appears to be on par with ZINC20-2B in terms of predictive performance for physiology (classification) datasets, while exhibiting a slight disadvantage for physical chemistry (regression) datasets. Overall, ChEMBL-2M seems to provide the best CLM embeddings, which may be attributed to the larger set of SELFIES symbols it contains.

BBBP (binary labels of blood-brain barrier penetration) is the only physiology dataset with a single task, similarly to all physical chemistry (regression) datasets. All compound representations show comparably strong predictive performance, with ChEMBL-2M and Enamine-2M having a slight advantage over Enamine-2B and ZINC20-2B. BBBP is the only MoleculeNet dataset exhibiting some correlation between radius size and predictive performance for ECFPs.



(a) MCC values of ZINC20-2B and ECFP0.

(b) MCC values of ZINC20-2B and ChEMBL-2M.

Figure 3: Comparison of two compound representations at a time, and their Pearson correlation coefficient: (a) shows the minimum agreement between a CLM embedding and ECFP0, while (b) an even stronger agreement between CLMs pre-trained on $2M$ or $2B$ compounds.

Tox21 (qualitative toxicity measurements on 12 proteins) and ToxCast (toxicology data on 379 tasks) lead to similarly low predictive performance for all compound representations. ECFPs show a slight advantage with increasing radius when looking at the median MCC. Similarly to all physiology datasets except BBBP, ECFP2 performs at least as well as ECFP4.

SIDER (marketed drugs and adverse drug reactions on 27 system organ classes) exhibits low predictive performance for all compound representations. Though, ECFP0 performs slightly better than any other compound representation.

ClinTox (qualitative data of drugs approved and of those that failed clinical trials for toxicity) is the only physiology dataset with less than $1k$ compounds per task. All CLMs leads to 0 MCC except for ChEMBL-2M, which outperforms all compound representations in terms of predictive performance. This result may indicate that Enamine-2B and ZINC20-2B may have limited generalizability to the broader drug space represented in ClinTox (and ChEMBL).

ESOL (water solubility data for $1k$ compounds) leads to generally strong regression performance, with ECFP0 performing on par with ZINC20-2B and slightly ahead of the remaining compound representation.

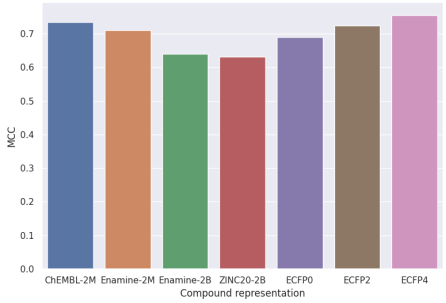
FreeSolv (hydration free energy for 642 compounds) is the only physical chemistry dataset where CLM embeddings exhibit superior predictive performance compared to ECFPs. Nonetheless, increasing the maximum depth and number of estimators for the RFs may further enhance the predictive performance on this dataset (see Figure 7 in the Supplementary material).

Lipophilicity (octanol/water distribution coefficient for $4k$ compounds) is the only physical chemistry dataset where ECFP2 and ECFP4 exhibit slightly better performance compared to all other compound representation.

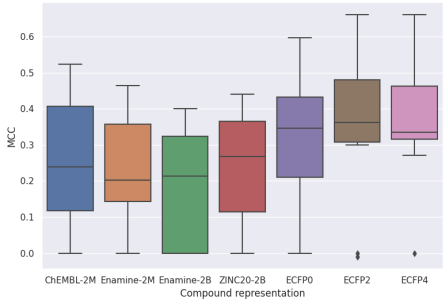
4 Conclusion

In this study, we explore whether CLMs yield representations that facilitate accurate predictions on the physical, physiological, and biophysical characteristics of small drug-like molecules. We systematically assess the impact of pre-training data origin and size. Our results show that simply scaling the pre-training of CLMs from millions to billions of example compounds does not lead to molecular representations that outperform ECFP4 on downstream drug prediction tasks. Furthermore, we do not observe substantial differences in predictive performance of CLMs when pre-trained on distinct biomolecule databases, including ChEMBL 33, Enamine REAL Database, and ZINC20. The only notable exception to these observations was the CLM trained on ChEMBL, which performed better than both ECFP4 and other CLMs on a single physiology benchmark dataset.

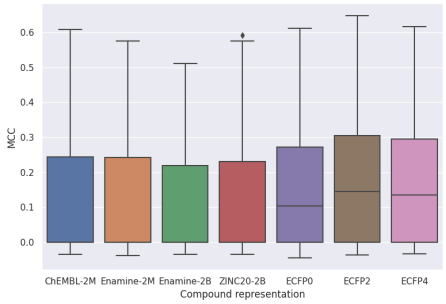
We further observed that, for most tests, CLMs – regardless of size or origin of the pre-training dataset – appear to have predictive performance comparable to ECFP with a radius of zero. ECFP0



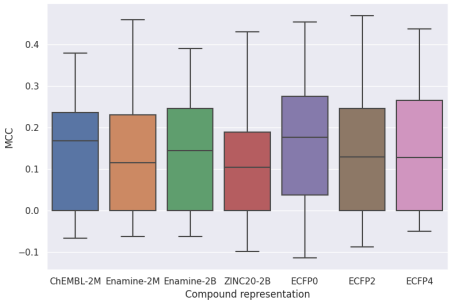
(a) MCC on BBBP.



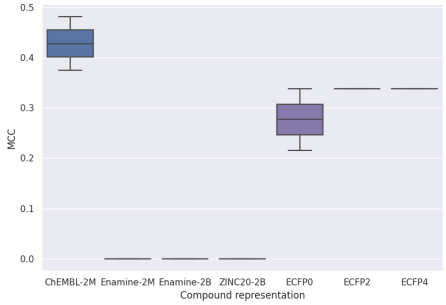
(b) MCC on Tox21.



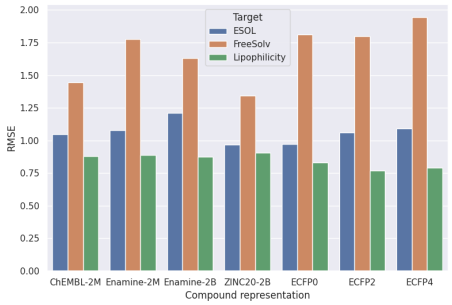
(c) MCC on ToxCast.



(d) MCC on SIDER.



(e) MCC on ClinTox.



(f) RMSE on test set.

Figure 4: MCC on (a-e) each physiology dataset, and (f) RMSE physical chemistry datasets of MoleculeNet. No clear superiority for ECFPs or CLMs is observable on (a), (d), (e) and (f).

encodes only basic physicochemical properties and the initial atom identifier, while ECFP4 includes topological information in addition to these properties. Based on these observations, we hypothesize that CLMs may not comprehensively learn topological information through masked language modeling. To address this limitation, we propose that future experiments with CLMs applied towards drug discovery and development should explore strategies that embed additional physicochemical inductive biases into the learning procedure.

Acknowledgments and Disclosure of Funding

The authors are grateful to Remco Loos, Giorgio Tamo, and Matthew Trotter from BMS for useful comments and suggestions.

Antonio de la Vega de Leon is funded thanks to project PTQ2021-011654 by MCIN/AEI/10.13039/501100011033.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] J. Born and M. Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence 2023*, pages 1–13, 4 2023.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[4] O. J. Béquignon, B. J. Bongers, W. Jespers, A. P. IJzerman, B. van der Water, and G. J. van Westen. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *Journal of Cheminformatics*, 15:1–11, 12 2023.

[5] S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] W. A. Falcon. Pytorch lightning. *GitHub*, 3, 2019.

[8] Y. Fang, N. Zhang, Z. Chen, X. Fan, and H. Chen. Molecular language model as multi-task generator. *arXiv preprint arXiv:2301.11259*, 2023.

[9] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, and R. A. Sayle. Zinc20 - a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60:6065–6073, 12 2020.

[10] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3:015022, 1 2022.

[11] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[12] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[13] G. Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

[14] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. Dos, S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, and M. Ai. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, page 2022.07.20.500902, 7 2022.

[15] H. Mary, E. Noutahi, DomInvivo, M. Moreau, L. Zhu, S. Pak, D. Gilmour, t, Valence-JonnyHsu, H. Hounwanou, I. Kumar, S. Maheshkar, S. Nakata, K. M. Kovary, C. Wognum, M. Craig, and D. Bot. datamol-io/datamol: 0.11.4, Sept. 2023.

[16] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. D. Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47:D930–D940, 1 2019.

[17] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118:e2016239118, 4 2021.

[18] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021.

[19] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 5 2010.

[20] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[21] A. Shivanyuk, S. Ryabukhin, A. Tolmachev, A. Bogolyubsky, D. Mykytenko, A. Chupryna, W. Heilman, and A. Kostyuk. Enamine real database: Making chemical diversity real. *Chemistry today*, 25(6):58–59, 2007.

[22] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

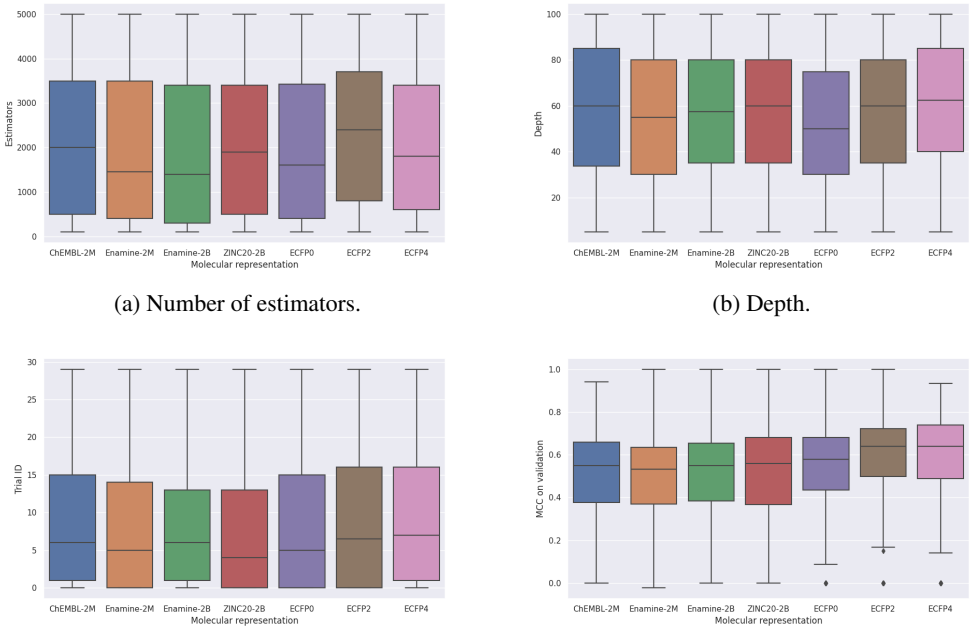
[23] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436, 9 2019.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[25] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[26] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan. Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4:025035, 6 2023.

5 Supplementary material



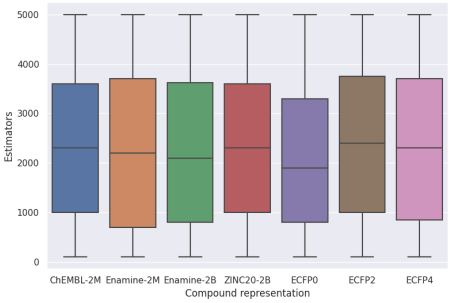
(a) Number of estimators.

(b) Depth.

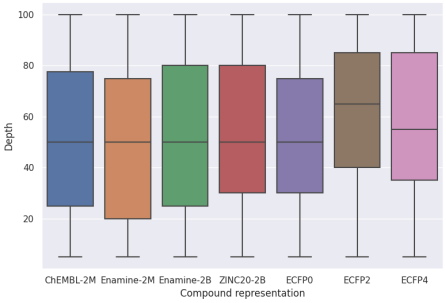
(c) Best trial ID.

(d) MCC on validation set.

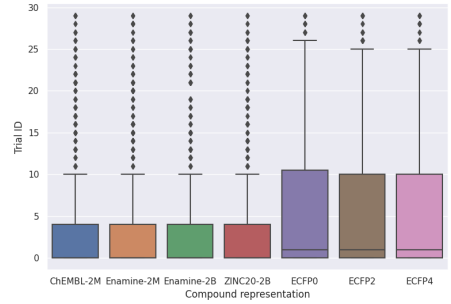
Figure 5: Overview of the best hyperparameters for Papyrus1k (a-b), the number of trials to find them (c), and their MCC on validation set (d).



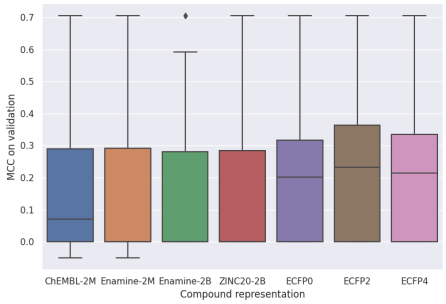
(a) Estimators.



(b) Depth.

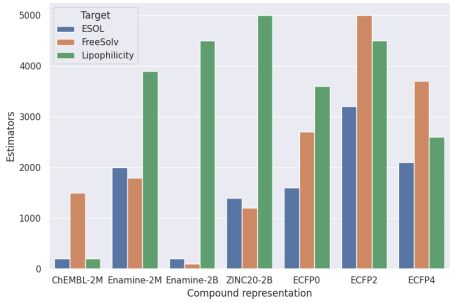


(c) Best trial ID.

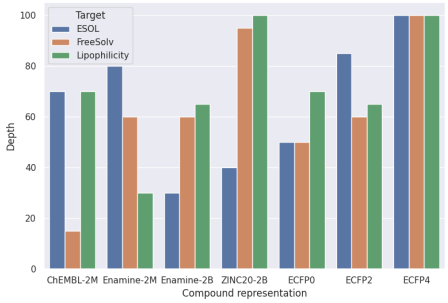


(d) MCC on validation set.

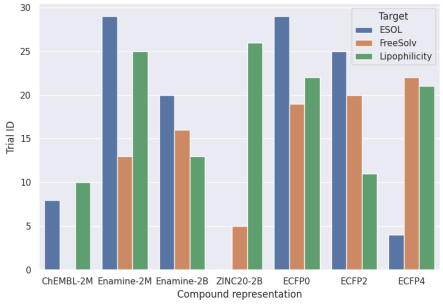
Figure 6: Overview of the best hyperparameters for MoleculeNet - physiology (a-b), the number of trials to find them (c), and the resulting MCC on validation (d).



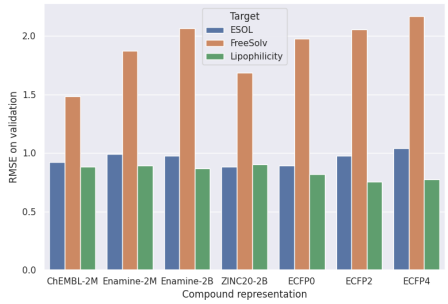
(a) Estimators.



(b) Depth.



(c) Best trial ID.



(d) RMSE on validation set.

Figure 7: Overview of the best hyperparameters for MoleculeNet - physical chemistry (a-b), the number of trials to find them (c), and the resulting RMSE on validation (d).