BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions

Andreu Vall, Sepp Hochreiter, Günter Klambauer
Institute for Machine Learning
Johannes Kepler University
Linz, Austria
klambauer@ml.jku.at

Abstract

Screening molecules for desired biological activities with bioassays is at the core of the drug discovery process. The data produced by bioassays enable building quantitative structure-activity relationship (QSAR) models that are fundamental components of computer-aided drug discovery. Despite the advances brought by Deep Learning-based QSAR models, it is still unclear how to build these models for new bioassays for which no active nor inactive molecules are known. To ameliorate this problem, we propose BioassayCLR, a machine learning method that leverages rich textual bioassay descriptions for modeling. Our model takes as input both the chemical structure of a molecule and the textual description of the bioassay and outputs the predicted activity for this pair. The approach can be viewed as a contrastive learning approach in which representations of both molecules and bioassays should be learned, which are similar if the molecule-bioassay pair is active and dissimilar if the pair is inactive. We perform experiments on bioassay descriptions and molecules from PubChem with 223,219,241 records of molecule-bioassay activity, corresponding to 2,120,811 unique molecules and 21,002 unique bioassays. On a strict temporal hold-out set with 615 unseen bioassays and 248,290 unseen molecules, BioassayCLR reaches an AUROC of 63.97 ± 0.47 outperforming the baselines using simple textual similarity by a margin, whereas all other QSAR methods yield random performance of 50.00. To our knowledge, this is the first time that a textual representation of a bioassay is directly fed into a QSAR model and, thus, the first method that can produce accurate predictions for bioassays that are only described by natural language. Because of these properties, our method allows for zero-shot transfer learning in drug discovery.

1 Introduction

Selecting compounds to be screened by a new bioassay is the critical step in drug discovery. The use, development and improvement of biological assays (bioassays) is at the heart of drug discovery. In this field, bioassays take the central role to determine the biological properties of a small molecule, such as inhibitory activity on a drug target in a wet-lab test. New bioassays are often developed with the aim to screen a library of molecules for a particular activity on a target. At this initial phase, when a new bioassay has been developed, the library design problem emerges in all drug discovery projects [36]. The library design problem concerns how to select molecules to be screened without previous experience about the new bioassay [15, 10, 19]. A good selection of molecules will lead to a high number of active molecules, which can potentially be further developed into a drug. Therefore, this initial phase critically determines the success of a drug discovery project and is usually both time- and cost-intensive. The drug discovery process could be made more effective by improving the selection of molecules to be tested in a newly developed bioassay (Section A.1)

ELLIS Machine Learning for Molecules workshop (ML4Molecules) 2021.

Bioassays provide data to build quantitative structure activity relationship (QSAR) models. As soon as molecules have been selected and measured with the new bioassay, data is available which can be used for data-driven and machine learning approaches [58, 23]. With these approaches, QSAR models can be built, which can then be used to virtually screen databases of molecules with high predicted activities. These molecules can then be further tested in a wet-lab [32, 42] to increase the set of active molecules, which is a determining factor in drug discovery [17, 3]. This process of acquiring bioassay data and improving the QSAR model depends strongly on the initial selection of molecules [15, 42].

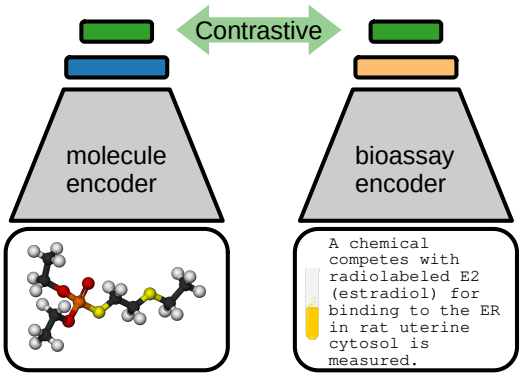

Figure 1: Schematic overview of our approach. BioassayCLR learns similar representations for active molecule-bioassay pairs.

QSAR modeling has been strongly improved by Deep Learning which needs large amounts of labelled data. Since the advent of Deep Learning methods in drug discovery, QSAR models have been strongly improved with respect to predictive quality and thus ranking and selection of molecules with desired activity [48, 6, 23, 53, 56, 9]. Machine learning methods, including Deep Learning, rely on training data, that is, a set of molecules that has already been tested in the assay and hence each molecule has an associated activity score. Usually several tens of active and inactive molecules are necessary to yield models with a good predictive quality [32, 46, 56]. To this end, recent efforts have been undertaken to make Deep Learning models more efficient with respect to the necessary training data [2, 35], an area which is called *few-shot learning* or *low-resource drug discovery*. Despite the recent progress in Deep Learning and low-resource drug discovery, the problem of the selecting molecules for a bioassay without any known actives or inactives, the so-called *zero-shot learning problem*, has not been solved.

Information from the textual description of the bioassay can be leveraged with contrastive learning. Despite the lack of known active and inactive molecules for novel bioassays, there is information available that could potentially be used for machine learning: the textual description of the bioassay. For each bioassay, the procedure in the wet-lab, their endpoint, and the substrate, is usually described in textual form. Their have even been efforts to semantically describe such bioassays using an ontology [52]. With the recent advances of machine learning methods for natural language [47, 51, 57], it has become evident that information in textual form can be leveraged for predictive models. As a prominent example, the BioBERT model [25] has been trained on biomedical texts and has been shown to be highly effective for biomedical named entity recognition, biomedical relation extraction, and biomedical question answering. In light of these results, we hypothesize that a meaningful representation of the textual description of the bioassay might be learned in a contrastive learning [13] approach and be used for QSAR models.

To solve the library design problem for new bioassays and the zero-shot learning problem in drug discovery, we propose a new machine learning model that takes as input both the chemical structure of a molecule and a textual description of a bioassay (Fig. 1). To this end, a bioassay encoder and a molecule encoder are trained in a contrastive learning approach. Contrastive learning methods have recently had a profound impact on machine learning and computer vision, because they offer a way to learn powerful, transferable representations [13, 8, 14]. We introduce a new method called BioassayCLR, which procures meaningful representations of both molecules and bioassays, and which can predict bioassay activity even when no actives and inactives are known (Fig. 2). Thus, we solve the problem of selecting compounds for a new bioassay and, equivalently, the zero-data or zero-shot problem in drug discovery.

2 BioassayCLR

Bioactivity prediction is usually considered as a classical supervised, binary prediction prediction task. For a given bioassay or drug target, a machine learning model $\hat{y} = g(m)$ can be trained on a set of available measurement pairs of molecules and activity labels $\{(m_1, y_1), \ldots, (m_N, y_N)\}$.

The problem has also been treated as a multi-task learning problem [49, 9, 41, 31, 32], in which several types of activity labels are available for a molecule $\{(m_1, \boldsymbol{y}_1), \ldots, (m_N, \boldsymbol{y}_N)\}$, where $\boldsymbol{y}_n$ are vectors containing activity values of different bioassays or drug targets. The advantage of multi-task learning over single-task is that a learned molecule encoder $\boldsymbol{m} = \boldsymbol{h}(m)$ can be shared across prediction tasks. However, multi-task deep networks (MT-DNN) cannot be used for zero-shot transfer learning, when predictions should be made for a new bioassay for which no training data is available.

Contrastive learning of molecules and bioassays. To allow for meaningful predictions of bioassays, for which no training data is available, we propose a contrastive learning approach in which molecule representations are learned together with representations of bioassays. Our machine learning model uses both a molecule m and a textual description of the bioassay a as input. To train this model, we consider the training data as triplets $\{(\boldsymbol{m}_1, \boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{m}_N, \boldsymbol{a}_N, y_N)\}$ of molecule embeddings from a molecule encoder $\boldsymbol{m} = \boldsymbol{f}(m)$, bioassay embeddings from a bioassay encoder $\boldsymbol{a} = \boldsymbol{g}(a)$, and a binary activity label y. A scoring function $k(\boldsymbol{m}, \boldsymbol{a})$ should return high values if a molecule $\boldsymbol{m}$ is active on a bioassay $\boldsymbol{a}$ and low values otherwise. The contrastive learning approach equips our model with the potential for zero-shot transfer learning, that is, supplying meaningful predictions for unseen bioassays.

The BioassayCLR model has the following structure:

$$\hat{y} = k(\boldsymbol{m}, \boldsymbol{a}) = k(\boldsymbol{f}(m), \boldsymbol{g}(a)), \qquad (1)$$

where $\hat{y}$ is the predicted activity, $\boldsymbol{f}(.)$ and $\boldsymbol{g}(.)$ are neural networks serving as the *molecule encoder* and the *bioassay encoder*, $k(.,.)$ is a scoring function that should approximate the targeted distribution $p(y = 1 \mid \boldsymbol{m}, \boldsymbol{a})$. In practice, we use the following: $k(\boldsymbol{m}, \boldsymbol{a}) = \frac{\exp(\tau^{-1} \boldsymbol{m}^T \boldsymbol{a})}{\exp(\tau^{-1} \boldsymbol{m}^T \boldsymbol{a}) + 1}$, where τ^{-1} can either be a hyperparameter or a learned parameter (Section A.3).

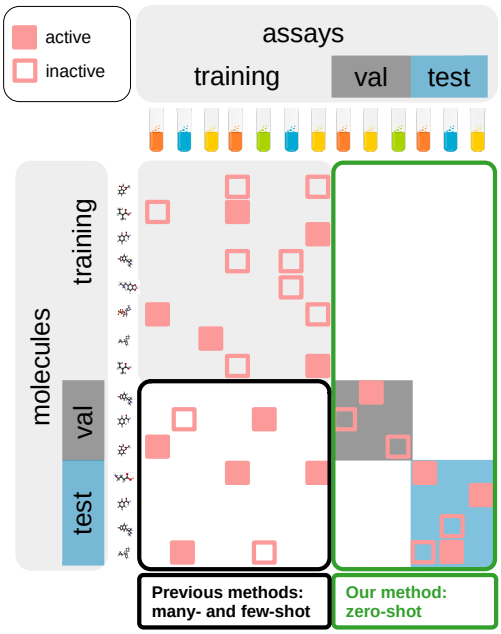

Figure 2: **Predictive ability of methods:** While previous methods, e.g. multi-task deep networks, make predictions on unseen molecules and known bioassays (black box), BioassayCLR allows to make predictions for molecules on unseen bioassays (green box). **Strict hold-out evaluation:** To assess the ability of our method to predict unseen bioassays, we use a strict hold-out setting: methods are assessed on their ability to correctly predict the activity of unseen molecules in unseen bioassays (blue area). Both molecules and bioassays are sorted increasingly by their PubChem identifier, given at the time of their entry in the database, which leads to an approximate temporal validation strategy. BioassayCLR allows for accurate predictions of assays without any training data, that is zero-shot transfer learning.

The objective of our model is to minimize the following contrastive loss function [13, 34, 29, 20]:

$$\mathrm{L_{NCE}} = -\frac{1}{N} \sum_{n=1}^{N} y_n \log(k(\boldsymbol{m}_n, \boldsymbol{a}_n)) + (1 - y_n) \log(1 - k(\boldsymbol{m}_n, \boldsymbol{a}_n)). \qquad (2)$$

The loss function encourages that molecules that are active on a bioassay have correlated representations, whereas inactive molecules have decorrelated representations, to the given bioassay. In contrast to our approach, recent prominent contrastive learning approaches [40, 7] only have access to pairs without label. Another difference to these methods is that for zero-shot transfer learning of bioactivity tasks, only a representation of the positive class, but not of the negative class, is available.

Encoders. Both the molecule and the bioassay encoders consist of a feature extraction component and a learned component. For the molecules, we first extract molecular descriptors, which we pass further to a fully connected neural network. For the bioassays, we first process their text descriptions

Table 1: Mean of AUROC, average precision (AVGP) and negative-class average precision (NegAVGP) over 615 test bioassays for zero-shot transfer learning. The table shows the mean and one standard deviation of this mean value over five runs initialized with different random seeds.

Method	Bioassay encoder	AUROC [%]	AVGP [%]	NegAVGP [%]
BioassayCLR (ours)	LSA	**63.97 ± 0.47**	**46.34 ± 0.64**	75.14 ± 0.48
soft-NN (baseline)	LSA	61.99 ± 0.32	43.31 ± 0.81	75.69 ± 0.53
1-NN (baseline)	LSA	57.16 ± 0.92	41.25 ± 1.09	72.43 ± 0.52
BioassayCLR (ours)	BioBERT	62.52 ± 0.93	44.93 ± 0.72	75.28 ± 0.45
soft-NN (baseline)	BioBERT	61.71 ± 0.77	42.58 ± 0.70	75.31 ± 0.49
1-NN (baseline)	BioBERT	55.15 ± 0.76	40.89 ± 0.64	72.23 ± 0.56
Global (baseline, 1 run)	–	62.63 ± NA	45.11 ± NA	75.61 ± NA
MT-DNN[1] [9, 49, 31, 41]	–	49.68 ± 0.49	38.48 ± 0.45	69.35 ± 0.17

[1] equivalent to a random classifier, in this case

either using Latent Semantic Analysis (LSA) or using BioBERT as a feature extractor. We then pass these representations to a fully connected neural network. Both fully connected neural networks are learned following the contrastive paradigm described above. Details on feature extraction and on the investigated architectures and hyperparameters can be found in Section A.3.

3 Experiments and results

We use a large public dataset extracted from PubChem [39] with 223,219,241 records of molecule-bioassay activity (Section A.2). In contrast to other benchmarking datasets for molecular property prediction, the PubChem database offers large number of textual descriptions of the bioassays, from which our method can learn. We split the dataset into training, validation and test using an approximate temporal split (Section A.2.2), such that the test set only contains assays and molecules for which not a single activity measurement is contained in the training set. From the machine learning perspective, this represents a zero-shot learning problem.

BioassayCLR is compared to two informative baseline methods derived from MT-DNN. Our baselines could also be considered as new methods, as they have not been suggested before. The 1-nearest neighbour baseline (1-NN) uses the text representation of the new bioassay and selects the most similar bioassay of the training set. Then, MT-DNN predictions for molecules of this bioassay are used as predictions. Similarly, the soft k-nearest neighbours (soft-NN) approach first calculates the textual similarity of the given bioassay with the training set bioassays. Then, these similarity values are used to calculate a weighted average of MT-DNN predictions for new molecules. We test two different bioassay encoders. First, the representations obtained using LSA and, second, the hidden representations obtained from BioBERT [26].

Our method BioassayCLR reaches an AUROC of 63.97 ± 0.47 based on the LSA-encoder, which can be considered high given that not a single activity value of this bioassays was available (Tab. 1). For a comparable study, AUROC values are in the range of 73.10 [32, Tab. 1] for bioassays with 3,900 measurements on average. Our two suggested baselines, soft-NN and 1-NN, developed with the capability of zero-shot learning, reach an AUROC of 61.99 ± 0.32 and 57.16 ± 0.92, whereas MT-DNNs remain at random performance. The learned textual representations of the bioassays yield similarities that are meaningful beyond pure textual similarity (Supplementary Material).

Conclusion. Our results are surprising in the sense that without a single activity measurement and only having the textual description of the bioassay, a predictive model can reach average AUROCs close to 64%. A critical consequence of this finding is that a bioassay need not even exist physically, but only be textually described, and already data-driven virtual screening for active molecules can be performed. The fact that the LSA encoder performs better than BioBERT indicates that the sentence structure and grammar of the bioassay descriptions does not seem to play a major role for predictive modeling, which we also did not expect. However, we see room for improvement of the bioassay encoder. It has not escaped our notice that our trained molecule encoders potentially contain more information than molecule encoders trained purely on activity data, which could make them suitable for transferring to other prediction tasks. We envision that BioassayCLR becomes a useful tool in early-stage drug discovery and that architectural improvements could even boost its performance.

Acknowledgments

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. IARAI is supported by Here Technologies. We thank the projects AI-MOTION (LIT-2018-6-YOU-212), DeepToxGen (LIT-2017-3-YOU-003), AI-SNN (LIT-2018-6-YOU-214), DeepFlood (LIT-2019-8-YOU-213), Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), AIRI FG 9-N (FWF-36284, FWF-36235), ELISE (H2020-ICT-2019-3 ID: 951847), AIDD (MSCA-ITN-2020 ID: 956832). We thank Janssen Pharmaceutica (MaDeSMart, HBC.2018.2287), Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, Software Competence Center Hagenberg GmbH, TÜV Austria, and the NVIDIA Corporation.

References

[1] Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(6):405–424, 2015.

[2] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

[3] J. Arrowsmith and P. Miller. Trial watch: phase ii and phase iii attrition rates 2011-2012. *Nature reviews. Drug discovery*, 12(8):569, 2013.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *arXiv:1607.06450 [cs, stat]*, July 2016.

[5] M.-W. Chang, L.-A. Ratinov, D. Roth, and V. Srikumar. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835, 2008.

[6] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[9] G. E. Dahl, N. Jaitly, and R. Salakhutdinov. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

[10] S. Dandapani, G. Rosse, N. Southall, J. M. Salvino, and C. J. Thomas. Selecting, acquiring, and using small molecule libraries for high-throughput screening. *Current protocols in chemical biology*, 4(3):177–191, 2012.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009.

[12] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[13] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[15] P. J. Hajduk, W. R. Galloway, and D. R. Spring. A question of library design. *Nature*, 470(7332):42–43, 2011.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv:1502.01852 [cs]*, Feb. 2015.

[17] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[19] J. J. Irwin. How good is your screening library? *Current opinion in chemical biology*, 10(4):352–356, 2006.

[20] H. Jiang, R. Wang, S. Shan, and X. Chen. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9765–9774, 2019.

[21] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant. PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, Jan. 2016.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, San Diego, CA, USA, 2015.

[23] G. Klambauer, S. Hochreiter, and M. Rarey. Machine learning in drug discovery. *Journal of chemical information and modeling*, 59(3):945, 2019.

[24] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[27] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073, 2014.

[28] Z. Liu, Y. Ma, Y. Ouyang, and Z. Xiong. Contrastive learning for recommender system. *arXiv preprint arXiv:2101.01317*, 2021.

[29] M. Lopez-Martin, A. Sanchez-Esguevillas, J. I. Arribas, and B. Carro. Supervised contrastive learning over prototype-label embeddings for network intrusion detection. *Information Fusion*, 2021.

[30] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[31] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

[32] A. Mayr, G. Klambauer, T. Unterthiner, M. Steijaert, J. K. Wegner, H. Ceulemans, D.-A. Clevert, and S. Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

[33] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[35] C. Q. Nguyen, C. Kreatsoulas, and K. M. Branson. Meta-learning initializations for low-resource drug discovery. *arXiv preprint arXiv:2003.05996*, 2020.

[36] C. A. Nicolaou and N. Brown. Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies*, 10(3):e427–e435, 2013.

[37] N. S. Pagadala, K. Syed, and J. Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9(2):91–102, 2017.

[38] M. M. Palatucci, D. A. Pomerleau, G. E. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. 2009.

[39] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, and G. Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[41] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[42] D. Reker and G. Schneider. Active-learning strategies in computer-assisted drug discovery. *Drug discovery today*, 20(4):458–465, 2015.

[43] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, 2002.

[44] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[46] N. Sturm, A. Mayr, T. Le Van, V. Chupakhin, H. Ceulemans, J. Wegner, J.-F. Golib-Dzib, N. Jeliazkova, Y. Vandriessche, S. Böhm, et al. Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics*, 12(1):1–13, 2020.

[47] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[48] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter. Deep learning as an opportunity in virtual screening. In *Deep Learning and Representation Learning Workshop, NIPS 2014*, 2014.

[49] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter. Multi-task deep networks for drug target prediction. In *Workshop on Transfer and Multi-task Learning, NIPS2014*, volume 2014, pages 1–4. NeurIPS, 2014.

[50] G. J. van Westen, J. K. Wegner, A. P. IJzerman, H. W. van Vlijmen, and A. Bender. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm*, 2(1):16–30, 2011.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[52] U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon, and S. C. Schürer. Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results. *BMC bioinformatics*, 12(1):1–16, 2011.

[53] W. P. Walters and R. Barzilay. Critical assessment of ai in drug discovery. *Expert Opinion on Drug Discovery*, pages 1–11, 2021.

[54] W. Wang, V. W. Zheng, H. Yu, and C. Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[55] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, Feb. 1988.

[56] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[57] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[58] L. Zhao, H. L. Ciallella, L. M. Aleksunes, and H. Zhu. Advancing computer-aided drug discovery (cadd) by big data and data-driven machine learning modeling. *Drug discovery today*, 2020.

[59] C. Zhou, J. Ma, J. Zhang, J. Zhou, and H. Yang. Contrastive learning for debiased candidate generation in large-scale recommender systems. *arXiv preprint arXiv:2005.12964*, 2020.

Contents

A Appendix

A.1 Related work

Selection strategies for bioassay screening. To increase the chance of success of a drug discovery project, many different strategies on how to select the molecule library have been proposed and tested. A prominent approach is high-throughput screening, in which large parts of physically available molecules are screened at high-throughput [15]. This is possible if the bioassay can be performed in high throughput, the wet-lab facilities and a large molecule library are available. High-throughput screening has been seen as a strong improvement in drug discovery. Naturally, many computational methods have also been suggested to first virtually screen [44] chemical libraries and then perform bioassay screening on the top-ranked molecules. Data-driven strategies, such as machine learning and Deep Learning, have brought a strong improvement of virtual screening methods. However, data-driven strategies are not possible for new bioassays [1] since no data is available, and no actives or inactives are known. To ameliorate this central problem, practitioners and scientists have resorted to using information from similar bioassays, facilitated by efforts to semantically structure the information about bioassays [52]. However, this type of information has not been integrated into machine learning approaches yet. In summary, while data-driven virtual screening strategies have been shown to be highly effective, it is currently unclear how those approaches could be used for designing libraries for newly developed bioassays.

Proteochemometric and molecular docking. Several efforts have been devoted to being able to make predictions for new biological targets, such as proteins. The set of proteochemometric methods [50] use information about the protein, such as its 1D structure, and combine it with information about the molecule. Molecular docking methods use the 3D structure of the protein and search for a conformation of a ligand that fits into a binding pocket [37, 33]. However, many bioassays are not focused on a target, but rather measure a general effect, such as a toxic response or cell proliferation, which limits or prohibits the use of proteochemic or docking methods.

Zero-shot learning problem. From the perspective of machine-learning, the described problem represents a zero-data or zero-shot prediction task [5, 24, 11, 38], for which several methods in the area of computer vision and natural language processing have been developed [54]. The setting is that no training data are available and only a description of the classes or tasks are provided, which in our case is the textual description of the bioassay. In contrast to zero-shot problems in computer vision where a description of each class is available, in the drug discovery setting only a description of the positive class is available. Contrastive learning for zero-data problems has recently been exemplified with the CLIP algorithm [40], in which representations of natural images and language are learned.

Recommender systems. The zero-data problem has earlier been identified by the recommender systems and matrix factorization research community as cold-start problem [43]. The cold-start problem is how to to provide good recommendations for novel users or items. Remedies for the cold-start problem of recommender systems exploit similarities of initial descriptions between users and items [27]. Contrastive learning has recently been suggested to learn the similarities between users and items [28, 59]. From the perspective of recommender systems, our method BioassayCLR

can be understood as having to suggest molecules representing items for a new bioassay representing a new user.

A.2 Data set

A.2.1 Bioactivity dataset

We use a large public dataset extracted from PubChem [39]. It is initially comprised of 224,290,250 records of compound-bioassay activity, corresponding to 2,120,854 unique compounds and 21,003 unique bioassays. We find that some compound-bioassay pairs have multiple activity records, which may not all agree. We reduce every compound-bioassay pair to exactly one activity measurement by applying majority voting. Compound-bioassay pairs with ties are discarded. This step yields our final bioactivity data set, which features 223,219,241 records of compound-bioassay activity, corresponding to 2,120,811 unique compounds and 21,002 unique bioassays.

A.2.2 Data splits

We conduct a temporal split to emulate the situation in which novel compounds and bioassays are not yet know at training time. We approximate this effect by assuming that new compounds and bioassays receive increasingly larger identifiers [21]. We split both the unique PubChem compound identifiers (CIDs) and the unique PubChem bioassay identifers (AIDs) into the oldest 60%, the following 20%, and the most recent 20%. Then, we take the bioactivity records corresponding to the 60% oldest compounds and bioassays for training, the bioactivity records corresponding to the following 20% of compounds and bioassays for validation, and the bioactivity records corresponding to the 20% most recent compounds and bioassays for testing (Fig. 2). Compound-bioassay pairs corresponding to, for example, older compounds tested on newer bioassays, will not be included in any of the splits. This entails a loss of data samples, but we favor the disjoint compound-wise and bioassay-wise splits, in order to conduct a strict evaluation of whether our proposed method can predict bioactivity on novel compounds and bioassays.

A.3 BioassayCLR

A.3.1 Molecule encoder

For each unique compound in the bioactivity data set, we retrieve its SMILES string [55] from PubChem. We use the Python[1] API of the RDKit[2] open-source chemoinformatics software to extract Daylight-like fingerprints, Morgan fingerprints and MACCS keys, which we concatenate obtaining final compound feature vectors of dimension of 2176.

A feed forward neural network takes the compound feature vectors as input and projects them to d-dimensional encodings. Further details on the network architecture and the choice of the dimension d are provided in Section A.3.3.

A.3.2 Bioassay encoder

For each unique bioassay in the bioactivity data set, we retrieve a rich textual description from PubChem, consisting of the concatenation of the title and the description of the bioassay. We process each textual description to obtain a fix-length bioassay feature vector. We follow two different pipelines, which we then assess separately.

We pretrain a Latent Semantic Analysis (LSA) model [30] specialized in PubChem bioassay descriptions. We obtain textual descriptions for 1,252,874 PubChem bioassays. To avoid leaking information, we exclude those corresponding to bioassays present in the validation and test splits of the bioactivity dataset. To train the LSA model, we first compute a bioassay-term matrix of tf-idf coefficients and then compute its truncated SVD decomposition. Finally, we extract the LSA feature vectors for all the bioassay descriptions in our bioactivity dataset. The LSA feature vectors have dimension 2048.

[1] https://www.python.org
[2] http://www.rdkit.org

Table 2: BioassayCLR hyperparameter values explored during model selection.

Hyperparameter	**Explored values**
Encoding dimension d	64, 128, 256
Learning rate	5×10^{-3}, 1×10^{-3}, 5×10^{-4}, 1×10^{-4}, 5×10^{-5}, 1×10^{-5}
Number of hidden layers	1, 2, 3, 4, 8
Number of units per hidden layer	128, 256, 512, 1024
Dropout probability	0.05, 0.1
Parameter τ	Set to 1, Learned
Normalization	Batch-based [18], Layer-based [4]

We use a pretrained instance of the BioBERT model [26],[3] which uses a transformer architecture [51] and has been trained on biomedical text corpora. Each bioassay description is provided as input to BioBERT and we keep the activations at the last layer as the bioassay feature vector. These are of dimension 1024.

The dimension of the BioBERT feature vectors (1024) is given by the architecture of the pretrained BioBERT model. The dimension of the LSA feature vectors (2048) is our hyperparameter choice. While it may seem that the LSA feature vectors are much larger than their BioBERT counterparts, we decided not to reduce them further because they might become too uninformative (the selected 2048 dimensions only explain 65% of the training data variance).

A feed forward neural network takes the assay feature vectors as input and projects them to d-dimensional encodings. Further details on the network architecture and the choice of the dimension d are provided in Section A.3.3.

A.3.3 Feed forward neural networks

The molecule and the bioassay encoders process their feature vectors using each a feed forward neural network. The network architecture on each encoder can be different, except for the output dimensionality d, which must agree. Nevertheless, here we have experimented with both networks having the same architecture.

The input and hidden layers in a network have the following structure

$$\text{dropout}\Big(\text{ReLU}\big(\text{norm}(\mathbf{W}\mathbf{x}+\mathbf{b})\big)\Big),$$

where $\mathbf{x}$ is the input to the layer, and $\mathbf{W}$ and $\mathbf{b}$ are learnable weights. The preactivations are followed by batch normalization [18] or layer normalization [4], a rectified linear unit (ReLU) activation function, and dropout [45]. The output layer does not have normalization, activation function, nor dropout, as it directly serves as the molecule or the bioassay encoding.

A.3.4 Hyperparameters

Models were selected by conducting a hyperparameter manual search (Tab. 2). We explored different configurations for the encoding dimension d, the learning rate, the number of layers, the number of hidden units in each layer, and the dropout probability. We also experimented with the parameter τ, necessary for the scoring function (Eq. 1), being set to 1 or learned, and with using either batch or layer normalization. In total, 115 hyperparameter combinations were investigated.

The search was run separately for models using LSA-based and BioBERT-based bioassay encoders. Model weights were initialized with MSRA [16]. For each hyperparameter configuration, we optimized the objective function (Eq. 2) using Adam [22] with a batch size of 256 samples. For each hyperparameter configuration, a copy of the model weights achieving the highest validation AUROC over 100 training epochs was stored. Upon analysis of the obtained validation metrics, we selected the final models (Tab. 3). We then trained four additional instances of each final model, resulting in five model instances, each having been initialized with a different random seed. Table 1 provides the final test results averaged over the five model instances.

[3] https://huggingface.co/dmis-lab/biobert-large-cased-v1.1

Table 3: BioassayCLR final hyperparameter configurations.

Hyperparameter	**LSA-based**	**BioBERT-based**
Encoding dimension d	128	128
Learning rate	5×10^{-5}	5×10^{-5}
Number of hidden layers	2	3
Number of units per hidden layer	512	256
Input dropout probability	0.1	0.1
Hidden dropout probability	0.05	0.05
Parameter τ	Learned	Set to 1
Normalization	Layer-based	Batch-based

Table 4: Baselines hyperparameter values explored during model selection.

Hyperparameter	**Explored values**
Learning rate	5×10^{-5}, 1×10^{-5}
Number of hidden layers	2, 3
Number of units per hidden layer	256, 512
Dropout probability	0.05, 0.1

A.4 Baselines

We propose baselines, which could be considered two variants of MT-DNN, for the purpose of making activity predictions for novel bioassays. In both cases, for a target novel bioassay, we compute the cosine similarity between its feature vector and the feature vectors of all the training bioassays, thus obtaining a vector of similarities. The first baseline method, 1-nearest neighbour (1-NN), predicts the bioactivity values that MT-DNN would predict for the training bioassay most similar to the target novel bioassay. The second baseline, soft k-nearest neighbours (soft-NN), is a smoother version of the first one. The vector of similarities between the target novel bioassay and the training bioassays is normalized using the softmax function, such that the resulting vector of weights sums up to one. Then, soft-NN predicts the weighted average of the values that MT-DNN would predict for all the training bioassays.

A.4.1 Hyperparameters

We trained a dedicated MT-DNN for each baseline model. Since our training, validation and test splits are bioassay-wise disjoint, we propose the following training procedure. Each MT-DNN visits the training set as usual, but it is then evaluated on the (bioassay-wise disjoint) validation set by using its predictions directly as 1-NN or soft-NN. In this way, we can train MT-DNN models for our baselines using exactly the same splits and information that BioassayCLR used.

Given the results of the hyperparameter search conducted for BioassayCLR, we conducted a hyperparameter search where we explored different configurations for the learning rate, the number of layers, the number of hidden units in each layer, and the dropout probability (Tab. 4). We set the parameter τ to 1 and used layer normalization.

The search was run separately for models using LSA and BioBERT bioassay feature vectors. Model weights were initialized with MSRA [16]. For each hyperparameter configuration, we optimized the multitask masked loss [31] using Adam [22] with a batch size of 256 samples. For each hyperparameter configuration, a copy of the model weights achieving the highest validation AUROC over 100 training epochs was stored. Upon analysis of the obtained validation metrics, we selected the final models (Tab. 5, 6). We then trained four additional instances of each final model, resulting in five model instances, each having been initialized with a different random seed. Table 1 provides the final test results averaged over the five model instances.

Table 5: 1-NN final hyperparameter configurations.

Hyperparameter	**LSA-based**	**BioBERT-based**
Learning rate	1×10^{-5}	1×10^{-5}
Number of hidden layers	3	2
Number of units per hidden layer	255	512
Input dropout probability	0.05	0.5
Hidden dropout probability	0.05	0.05
Parameter τ	Set to 1	Set to 1
Normalization	Layer-based	Layer-based

Table 6: Soft-NN final hyperparameter configurations.

Hyperparameter	**LSA-based**	**BioBERT-based**
Learning rate	5×10^{-5}	1×10^{-5}
Number of hidden layers	2	3
Number of units per hidden layer	512	256
Input dropout probability	0.1	0.5
Hidden dropout probability	0.05	0.05
Parameter τ	Set to 1	Set to 1
Normalization	Layer-based	Layer-based

A.5 Metrics

In this work we report three performance metrics, which we denote AUROC, AVGP and NegAVGP. AUROC is the area under the ROC curve [12]. AVGP is the mean average precision, which is an approximation of the area under the precision-recall curve [30]. The last metric, which we dub "NegAVGP" for negative-class mean average precision, is not standard, but it is very informative. It is simply the mean average precision of the negative class. That is, if the negative class is coded as 0 and the positive class is coded as 1, then NegAVGP is,

$$\text{NegAVGP} = \text{AVGP}(1 - y, 1 - \hat{y}).$$

A.6 Additional results

Table 7: Median of AUROC, AVGP and NegAVGP over 615 test bioassays for zero-shot transfer learning. The table shows the mean and one standard deviation of this median value over five runs initialized with different random seeds.

Method	**Bioassay encoder**	**AUROC [%]**	**AVGP [%]**	**NegAVGP [%]**
BioassayCLR (ours)	LSA	66.72 ± 0.33	40.69 ± 2.83	90.42 ± 0.60
soft-NN (baseline)	LSA	64.25 ± 0.38	33.93 ± 0.85	89.54 ± 0.48
1-NN (baseline)	LSA	59.81 ± 1.16	30.91 ± 2.08	86.64 ± 0.75
BioassayCLR (ours)	BioBERT	66.04 ± 1.06	37.63 ± 1.94	89.73 ± 0.32
soft-NN (baseline)	BioBERT	63.94 ± 1.03	34.34 ± 1.12	89.54 ± 0.61
1-NN (baseline)	BioBERT	56.15 ± 1.74	31.09 ± 1.42	86.99 ± 0.20
MT-DNN[1] [9, 49, 31, 41]	–	49.95 ± 0.26	25.15 ± 0.48	83.19 ± 0.49

[1] equivalent to a random classifier, in this case