Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Multi-task Proteochemometric Modelling

Anastasia Pentina [1,*] and Djork-Arné Clevert [1]

[1] Machine Learning Research, Bayer AG, Berlin, 10117, Germany

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: In silico prediction of protein-ligand binding is a hot topic in computational chemistry and machine learning-based drug discovery, as an accurate prediction model could reduce the time and resources required to detect and identify and prioritize potential drug candidates. Proteochemometric modelling (PCM) is a promising approach for in-silico protein-ligand binding prediction that utilises both compound and target descriptors. However, in its original form PCM model cannot separate multiple assays associated with the same target. Therefore, a practitioner applying PCM approach to modelling experimental data has either to select only one assay for each target, and thus exclude potentially significant amount of data, or pull measurements from different assays together effectively mixing possibly very different functional dependencies between (protein, ligand) pairs and experimental measurements.
Results: We describe two modifications of PCM models that increase its flexibility allowing to separate multiple assays associated with the same target. Evaluated on a subset of internal Bayer dose-response data and ChEMBL, these approaches result in improved performance compared to standard PCM models. Our results demonstrate importance of disentangling multiple assays associated with the same target when using PCM methodology in pharmaceutical environment.
Availability: Source code is made publicly available on GitHub for non-commercial usage after publication.
Contact: anastasia.pentina@bayer.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Computational methods play an important role in improving efficiency of the drug discovery process. Quantitative Structure-Activity Relationship (QSAR) models allow estimating interactions between ligands and biological targets without the need for performing in vitro experiments and are therefore a promising approach to reduce time and costs of identifying active compounds in early drug discovery.

Building a reliable QSAR model from scratch requires a significant amount of experimental data which is problematic when working on new targets. The multi-task approach [5] compensates for limited data available for any individual target by modelling multiple targets jointly and thus sharing information between them. A typical example of this approach, which was a part of the winning solution to Merck challenge [16], is a feed-forward neural network that takes as input a compound descriptor and has as many outputs as there are targets being modelled. Other examples of using multi-task learning in QSAR modelling include linear models for drug response prediction [11], graph convolutional networks for ADMET property modelling [17, 10] and graph-regularized support-vector regression for kinase models [19].

Multi-task learning reduces the amount of data required per target for building a reliable model by taking advantage of correlations between compounds' affinities to different proteins that are present in the data. However, it doesn't account for known structural similarities between targets. Proteochemometric (PCM) approach [20, 8], on the other hand, models compound-protein interactions directly in the ligand-target space. By using as inputs not only descriptors of compounds, but also feature representations of targets, it allows modelling multiple targets using a single-output function. In addition to efficient utilisation of data for multiple targets, PCM approach provides other advantages, such as improved interpretability - usage of target's features allows tracking which of them were most important for concluding compound's (in)activity. Moreover, PCM models can be used to predict activity on completely new targets for which no experimental data is available. These properties make PCM approach an attractive modelling tool that has been applied to different targets using various machine learning methods ranging from random forests [4] to neural networks [15].

In practice PCM and QSAR models are built using historical experimental data and as such merely model assays' outcomes rather than true compound's activity. In many cases information about the target is not sufficient to fully identify an assay, because, for example, employed experimental protocols have been changing over the years, or because there are multiple assays each addressing a different aspect of compounds-target interaction. This forces a researcher wishing to employ PCM modelling to select which of potentially many assays corresponding to one target to include in the model and thus to ignore substantial amounts of knowledge

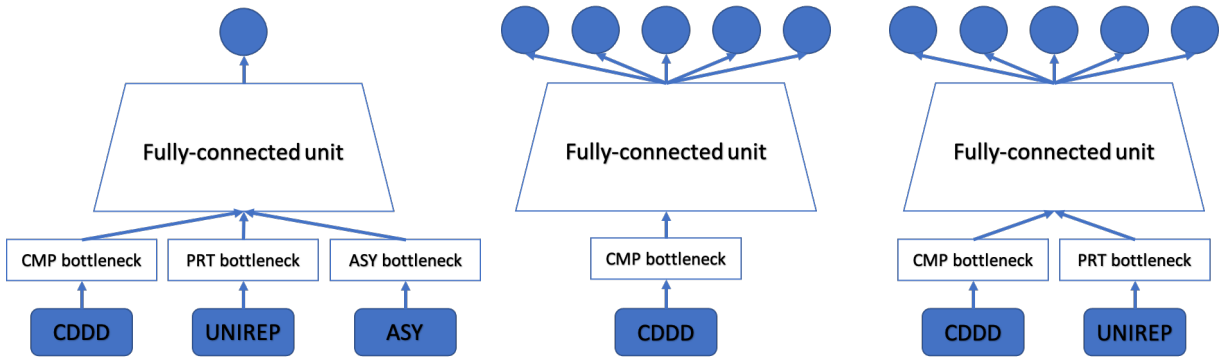


Fig. 1. A schematic representation of proteochemometric modeling PCM/PCM-ext (left), multi-task MT (middle) and multi-task proteochemometric modeling MT-PCM (right) architectures for predicting protein–ligand binding. The input units at the bottom receive combinations of compounds, proteins or assays embeddings, which first goes through a dedicated bottleneck layer (CMP, PRT, ASY), then concatenated and fed into a fully connected unit. Single- or multi-task output units are shown on the top.

present in the data for the left out assays. An alternative solution is to use all available data ignoring assays' specifics, effectively modelling all assays associated with one target as one. This is a viable strategy if merged assays are only marginally different, however, in practice multiple assays associated with one target are typically designed so that to provide maximal information about the studied protein-ligand interaction and thus are expected to have minimal correlations, if any.

In this work we study hybrid approaches that combine multi-task learning and PCM modelling using fully-connected neural networks. We demonstrate importance of being able to disentangle multiple assays associated with the same target and propose two approaches to achieve it. The first one is a modification of PCM model that in addition to compound and target embeddings takes as input a one-hot-encoding of assay identifier. The second approach is a multi-task PCM model that utilises compound and target embeddings as inputs and has a separate output for every modelled assay. These methods, like a PCM model, benefit from information about structural similarities between targets. At the same time, like a multi-task model, they have the flexibility to simultaneously model multiple assays per target. We test these approaches on internal Bayer data of 1364 dose-response assays and analyse under which circumstances they are more effective than the individual parts - MTL and PCM.

2 Data

For experimental evaluation we use a subset of internal Bayer dose-response data and ChEMBL. We convert endpoints values to pIC50 (negative log of the IC50 concentration) and keep only those between 0 and 15. We aggregate repetitive measurements by keeping the maximum value and the corresponding qualifier. If measurements with qualifier $<$ are present, only they are used for aggregation. All considered assays have at least 100 uncensored data points. For validation purposes we cluster compounds into 5 groups. The clusters were obtained by first computing the MACCSkey fingerprint [9] using RDKit, and then utilizing sklearn's KMeans clustering implementation[18] on the MACCSkey fingerprints. Finally, we ensure that every assay contains measurements for compounds in each of 5 clusters. Further statistics about the dataset can be found in Table 1.

Kim et al. [12] have shown that using unsupervised-learned representations for both compounds and targets leads to superior performance compared to hand-crafted features. Therefore, in this work we also employ this approach.

We represent every compound using Continuous and Data Driven Molecular Descriptors (CDDD) [21]. These are 512-dimensional descriptors learned in an unsupervised way using a recurrent autoencoder to translate between non-canonical SMILES and their canonical form. They have been shown to be very effective in QSAR modeling, as well as PCM modelling [21, 12], inverse molecular problems such as optical chemical structure recognition [7] or reverse-engineering of molecular structures [14].

For protein representation we use UniRep [3] embeddings as they performed the best in previous work [12]. They are trained using LSTM on predicting the next amino acid in the sequence given the previous ones. A fixed-length embedding for a given protein is obtained by averaging the hidden states of the model during the forward pass. Depending on the architecture, embeddings of different dimensionality are available. In our experiments we use the 256-dimensional version.

Table 1. Data used in this study were extracted from the ChEMBL25 database and PubChem. The final number of compounds in each task after preprocessing is mentioned.

Total number of data points	3,003,764
Number of uncensored data points	1,569,629
Number of assays	1,364
Number of targets	534
Number of compounds	561,495

3 Methods

All used methods are feed-forward neural networks with ELU activations [6]. Every type of input - CDDD embedding of compounds, UniRep embedding of proteins or assay description - first goes through a dedicated bottleneck (see CMP, PRT, ASY in Table 2). The outputs of the bottlenecks are concatenated and fed into a fully connected unit. A dropout is applied after every layer with a fixed rate.

Our starting point are a proteochemometric and a multi-task learning models. The first one, PCM, takes 3 types of inputs - CDDD embedding of a compound, UniRep embedding of a protein and a description of an assay - biochemical assays are encoded as $(1, 0)$ and all the others as $(0, 1)$. MTL uses only CDDD embedding of compounds and has 1364 outputs, one for each modelled assay.

The first proposed modification is PCM-ext - it is analogous to PCM, but uses a representation of every modelled assay by a one-hot-encoding.

Table 2. Specifications of the network architecture and hyperparameters of the models were selected using Optuna that was run for 150 trials with median pruner and 10 warm-up steps.

Method	Bottlenecks			Layers	Dropout	Learn rate
	CMP	PRT	ASY			
PCM	1024	1024	8	4096, 4096, 16	0.3	10^{-4}
MT	256	-	-	4096, 4096, 512, 256, 256	0.2	10^{-4}
PCM-ext	1024	1024	8	4096, 4096, 2048	0.1	10^{-4}
MT-PCM	1024	8	-	4096, 2048, 64	0.2	10^{-4}

The second modification - MT-PCM - is a combination of PCM and MTL in that it utilises both compound and target embeddings (like PCM) and has as many outputs as there are assays being modelled (like MTL). A schematic representation of all considered models can be found in Figure 1.

All models were trained using Adam optimiser [13] for 50 epochs with batch size of 1024. Multi-task models are based on censored squared loss that takes into account prefixes of censored value[1] :

$$\ell(y, \hat{y}) = \begin{cases} (y - \hat{y})^2, & \text{if prefix is } = \\ (y - \hat{y})\,\mathrm{ReLU}(y - \hat{y}), & \text{if prefix is } > \\ (\hat{y} - y)\,\mathrm{ReLU}(\hat{y} - y), & \text{if prefix is } < \end{cases}$$

For training PCM models we used regular squared loss as it gave better results.

All inputs and outputs, except for CDDD embedding and assay embeddings in PCM-ext were whitened. During training, we applied $\mathcal{N}(0, 0.05)$ Gaussian noise to protein embeddings and to CDDD embeddings we applied the following transformation to preserve their geometry:

$$\mathrm{cddd} = \tanh(\mathrm{arctanh}(\mathrm{cddd}) + \xi), \qquad \xi \sim \mathcal{N}(0, 0.05). \quad (1)$$

Hyperparameters of the models were selected using Optuna [2] that was run for 150 trials with median pruner and 10 warm-up steps. Widths of compound and protein bottlenecks were sampled uniformly from the set $\{2^2, 2^3, \ldots, 2^{10}\}$ and assay bottleneck width was fixed to 8. Number of layers in the fully connected unit was selected from 1 to 6 and their widths were sampled uniformly from $\{2^2, 2^3, \ldots, N\}$, where N is the minimum value between the width of the previous layer and 2^{12}. This ensured a pyramidal structure of the networks. Dropout rate was selected uniformly from $\{0, 0.1, \ldots, 0.5\}$ and the learning rate from $\{10^{-5}, 10^{-4}, 10^{-3}\}$. As training data for Optuna trials we used 4 folds and the fifth was used as the validation set. Quality of a trial was estimated using R^2 measure based on uncensored values from the validation fold. Resulting architectures are summarised in Table 2.

4 Results

4.1 Overall comparison

We evaluate all methods using leave-one-fold-out evaluation and only use uncensored data for metric computation. During the training of all methods we weight all data points equally (without any normalisation with respect to assay/task size) and the evaluation is done in the similar manner - by computing performance on test data as if it is all coming from one task. Results are summarised in Table 3. Note that we used one of the clusters for selecting hyper-parameters with Optuna. Potentially this could lead to overfitting on that cluster 4, however Figure 2 shows that this is not the case. Performances across all clusters are comparable and ranking of different methods is generally stable.

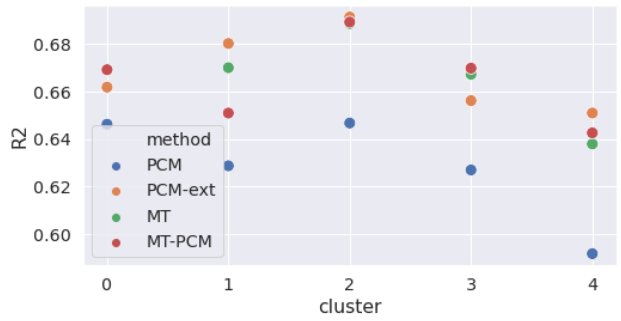


Fig. 2. Spearman correlation as a function of compound cluster. Despite the fact that cluster 4 was used to select models' hyperparameters with Optuna, one doesn't observe any overfitting effects on that cluster and ranking of considered methods is stable across all folds.

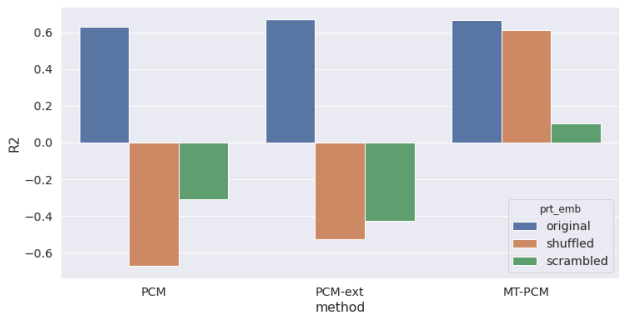


Fig. 3. Performance of PCM and MT-PCM for different protein embeddings used in test set. "Scrambled" embeddings were obtained by randomly shuffling coordinates of UniRep protein embeddings, while "shuffled" correspond to randomly substituting correct embedding with embedding of another target.

Results reported in Table 3 and Figure 2 demonstrate that three methods - MT, MT-PCM and PCM-ext - perform superior to PCM across all considered performance measures - root mean squared error (RMSE), R^2, Spearman and Pearson correlations, as well as fraction of well-modelled assays for which $R^2 > 0.5$. This shows the importance of being able to disentangle multiple assays associated with the same target. Indeed, this flexibility is the distinct feature of these three methods, not shared by the standard proteochemometric approach.

Table 3. Performance measures at computed using all uncensored data from test fold. From left to right: root mean squared error (RMSE), coefficient of determination (R^2), Spearman and Pearson correlations, as well as fraction of well-modelled assays for which $R^2 > 0.5$

Method	RMSE	R^2	SpearCorr	PearCorr	% assays $R^2 > 0.5$
PCM	0.665	0.632	0.768	0.816	9.9
MT	0.629	0.670	0.789	0.834	11.4
PCM-ext	0.628	0.672	0.792	0.835	12.8
MT-PCM	0.632	0.667	0.788	0.832	11.5

In contrast to MT, MT-PCM and PCM-ext models have access to information about the structure of the targets provided by UniRep embeddings. This allows them to use structural similarities between proteins to model their binding behaviours. As a result, these models are able to exploit not only correlations in compounds' affinities to various targets, but also the intuition that small molecules interact similarly with

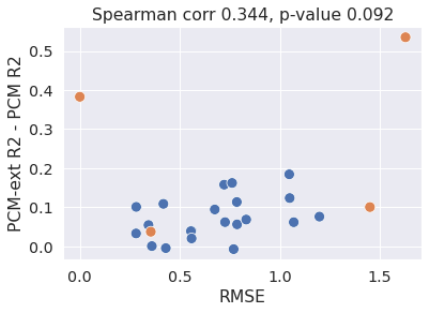


Fig. 4. Difference in PCM and PCM-ext performances as a function of difference between pairs of assays, associated with the same target. Orange markers correspond to selected targets analysed in details in Figure 6.

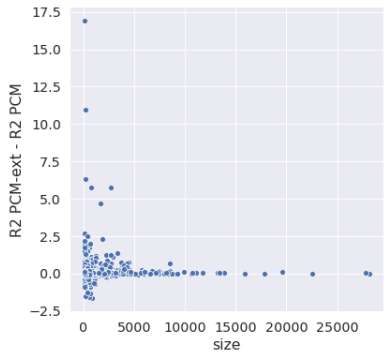


Fig. 5. Delta in R^2 performance as a function of assay size. Every dot corresponds to one modelled assay and only assays with standard deviation of at least 0.05 units based on uncensored pIC50 measurements are depicted.

structurally similar targets. In principle, MT-PCM and PCM-ext could use to not utilise this information, by either selecting a very narrow protein embedding layer during the Optuna run, or by assigning negligibly low weights to it during the final training. In our experiments this is observed for MT-PCM approach: as reported in Table 2, Optuna run for this method resulted in the smallest possible bottleneck for protein embeddings. We additionally asses the role of target information on performance of all three models - PCM, PCM-ext and MT-PCM - by evaluating them using modified protein embeddings. Results in Figure 3 demonstrate that while performance of all methods deteriorates when protein embeddings are modified, MT-PCM exhibits a relatively low drop, especially for "shuffled" modification.

Note that non of the three methods - MT, MT-PCM and PCM-ext - can be expected to always provide superior performance compared to the other ones, because they are based on different relatedness assumptions about the data. MT method assumes that there exists a feature representation, under which all targets can be modelled using a linear function and its last layer provides such an embedding. MT-PCM, on the other hand, assumes that there is such a beneficial embedding for (compound, target) pairs. PCM-ext represents in fact a very similar relatedness assumption. Consider a splimplified scenario, where instead of all bottleneck and fully-connected layer on uses just one linear layer. Then MT-PCM would result in independent solving of every task n with a weight vector $w_n = (w_n^{cmp}, w_n^{prt})$ as:

$$(w_n^{cmp})^T x_{cmp} + (w_n^{prt})^T x_{prt}. \tag{2}$$

At the same time PCM-ext would correspond to solving all tasks with the same weight vector, but different biases:

$$(w^{cmp})^T x_{cmp} + (w^{prt})^T x_{prt} + (w^{asy})^T e_n = \tag{3}$$

$$(w^{cmp})^T x_{cmp} + (w^{prt})^T x_{prt} + w_n^{asy}, \tag{4}$$

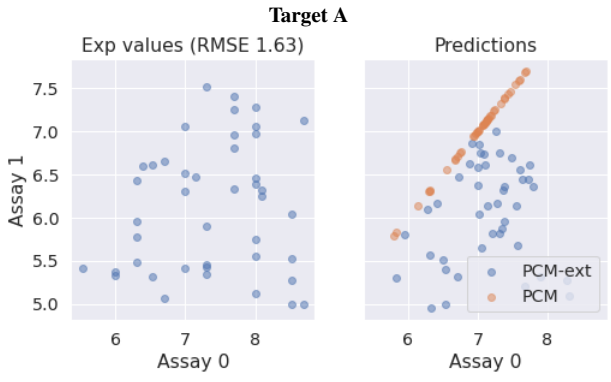
where $e_n = (0, \ldots, 0, 1, 0 \ldots, 0)$ is a vector with only one non-zero element in position n. In this simplistic case it's evident that PCM-ext relies on a stronger dependence between tasks, however in non-linear case the differences between MT-PCM and PCM-ext are more subtle.

4.2 Per assay performance comparison

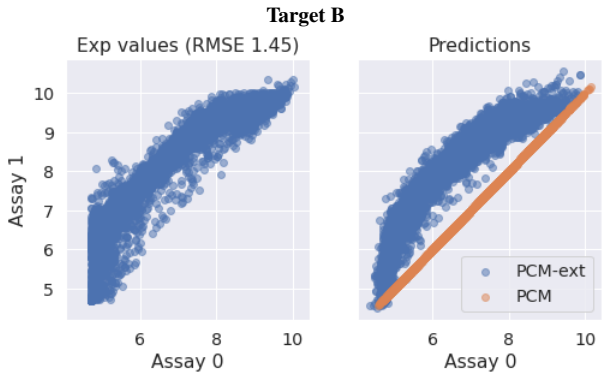
The main advantage of MT, MT-PCM and PCM-ext over PCM is in that they have access to assay identifiers and are capable of modelling multiple assays associated with the same target. If in the data being modelled, every target is associated with only one assay/task, this flexibility might not bring any benefit and even hurt the performance by making the training data more sparse. However, in the data used in this work 52% of targets are associated with more than one assay and they account for 88% of all data points. PCM model attempts to overcome this limitation by taking into account assay type, however, in the dataset we consider it is not sufficient, as 47% of (target, assay type) pairs are associated with more than one assay. Therefore a clear boost in performance of MT, MT-PCM and PCM-ext compared to PCM, as reported in Table 3, is predictable. At the same time the differences in modelling quality between these three methods are rather limited. For simplicity in our subsequent discussion of model performances on individual assays we will be focusing on PCM-ext as it has the best overall performance.

Intuitively, one expects PCM-ext to outperform PCM on assays which are not uniquely identifiable by their target protein. To quantify this intuition we selected targets which are associated with at least 2 different assays and those assays overlap on at least 10 compounds. For each of the resulting 25 targets we compute average R^2 over two assays. Figure 4 shows that there is a monotonous dependence (except for one outlier) between difference in assays measured by RMSE on overlapping compounds and benefits that PCM-ext demonstrates compared to PCM. On the other hand, if two assays correspond to the same target and at the same time can be well modelled using just one function, modelling them using PCM-ext might have no benefit compared to PCM. In fact, merging data for such assays, as de facto is done in PCM, could even be advantageous, especially in small data regime. This intuition is supported by per-assay performance analysis demonstrated in Figure 5: all the endpoints for which PCM provides better predictions than PCM-ext are small in size.

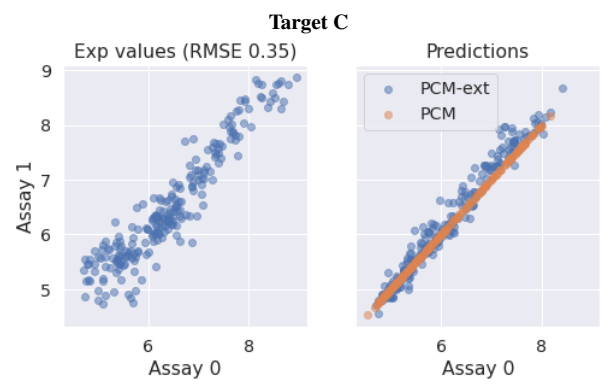
For further illustration we select 4 targets that are highlighted by orange color on Figure 4. We report differences between original measurements for pairs of tasks corresponding to these targets, as well as predictions of PCM and PCM-ext on overlapping compounds in Figures 6. The first one - Target A (Figure 6) - is an example of a case in which there is no clear dependence between measurements of two assays. Performance of PCM demonstrates that one of the assays dominated its learning process, leaving the second one very poorly modelled (with negative R^2). In contrast, PCM-ext has the capacity to model both assays reasonably well. The second example - Target B (Figure 6) - illustrates a situation in which there is clearly a dependence between two assays, but it is not an identity. As a result, PCM-ext results in overall better model quality compared to PCM. The last two examples - Target C and Target D (Figure 6) - illustrate a situation in which difference between two assays is quite small. For Target C, PCM was able to produce reasonable predictions, at the same time Target D is an outlier in the overall trend, for which PCM predictions are poor for no obvious reason. We attribute superiority of PCM-ext in this case to overall different model specifications.



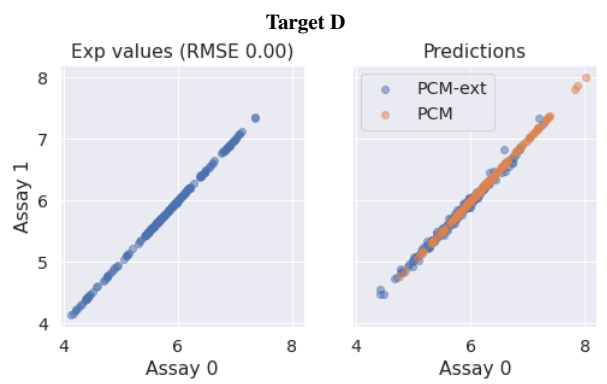
Method	Assay 0	Assay 1
PCM	0.494	-0.476
PCM-ext	0.532	0.556



Method	Assay 0	Assay 1
PCM	0.616	0.710
PCM-ext	0.708	0.819



Method	Assay 0	Assay 1
PCM	0.606	0.717
PCM-ext	0.629	0.768



Method	Assay 1	Assay 2
PCM	0.204	0.195
PCM-ext	0.585	0.580

Fig. 6. Illustration of 4 selected targets (A-D) that are highlighted by orange color on Figure 4. Every dot corresponds to one compound, representing either the true experimental pIC50 across two assays (left plots) or the pIC50 model predictions (right plots) for PCM-ext (blue) or PCM (orange). The RSME between the experimental pIC50 values of the two assays is reported on the left plot. Tables report performance of PCM and PCM-ext on these two assays as measured by R^2 on uncensored data only.

5 Conclusion

In this work we examined proteochemometric modelling on internal Bayer data. We demonstrated that in realistic scenarios, when multiple assays in historical experimental data might correspond to the same target, PCM models suffer from inability to disentangle such endpoints. Our results show that usage of more flexible models - either through multi-task approach (like MT-PCM) or by encoding assays identifiers in the model input (PCM-ext) - leads to superior modelling quality. Our in-depth analysis of four exemplar targets shows that the benefits of more flexible models are most pronounced in case readouts from multiple assays associated with the same target have a complex dependence, if any.

In silico protein–ligand binding prediction might further be improved by developing more powerful protein descriptors that contain binding site information, which we believe are relevant directions for future research in this area.

Acknowledgements

This project was received financial support from European Commission grant numbers 831472, 963845 and 956832 under the Horizon2020 Framework Program for Research and Innovation.

References

[1]MELLODDY Consortium. https://www.melloddy.eu. Accessed: 2022-01-12.

[2]Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3]Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, 2019.

[4]Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26 9:1169–75, 2010.

[5]Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6]Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR*

2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

[7]Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2mol – accurate smiles recognition from molecular graphical depictions. *Chem. Sci.*, 12:14174–14181, 2021.

[8]Isidro Cortés-Ciriano, Qurrat Ul Ain, Vigneshwari Subramanian, Eelke B. Lenselink, Oscar Méndez-Lucio, Adriaan P. IJzerman, Gerd Wohlfahrt, Peteris Prusis, Thérèse E. Malliavin, Gerard J. P. van Westen, and Andreas Bender. Polypharmacology modelling using proteochemometrics (pcm): recent methodological developments, applications to target families, and future prospects. *Med. Chem. Commun.*, 6:24–50, 2015.

[9]Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002. PMID: 12444722.

[10]Evan N. Feinberg, Elizabeth Joshi, Vijay S. Pande, and Alan C. Cheng. Improvement in admet prediction with multitask deep featurization. *Journal of Medicinal Chemistry*, 63(16):8835–8848, 2020. PMID: 32286824.

[11]Yuan H, Paskov I, Paskov H, González AJ, and Leslie CS. Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, 6:31619, 2016.

[12]Paul T. Kim, Robin Winter, and Djork-Arné Clevert. Unsupervised representation learning for proteochemometric modeling. *International Journal of Molecular Sciences*, 22(23), 2021.

[13]Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14]Tuan Le, Robin Winter, Frank Noé, and Djork-Arné Clevert. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ecfps) to their molecular structures. *Chem. Sci.*, 11:10378–10389, 2020.

[15]Eelke Lenselink, Niels Dijke, Brandon Bongers, George Papadatos, Herman Vlijmen, Wojtek Kowalczyk, Adriaan IJzerman, and Gerard Westen. Beyond the hype: deep neural networks outperform established methods using a chembl bioactivity benchmark set. *Journal of Cheminformatics*, 9:45, 08 2017.

[16]Junshui Ma, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure activity relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274, 2015. PMID: 25635324.

[17]Floriane Montanari, Lara Kuhnke, Antonius Ter Laak, and Djork-Arné Clevert. Modeling physico-chemical admet endpoints with multitask graph convolutional networks. *Molecules*, 25(1), 2020.

[18]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19]Lars Rosenbaum, Alexander Dörr, Matthias R. Bauer, Frank M. Boeckler, and Andreas Zell. Inferring multi-target QSAR models with taxonomy-based multi-task learning. *J. Cheminformatics*, 5:33, 2013.

[20]Gerard J. P. van Westen, Jörg K. Wegner, Adriaan P. IJzerman, Herman W. T. van Vlijmen, and A. Bender. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, 2:16–30, 2011.

[21]Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.*, 10:1692–1701, 2019.