Corresponding Projections for Orphan Screening

Sven Giesselbach[1,2], Katrin Ullrich[3], Michael Kamp[2,3], Daniel Paurat[1,2], and Thomas Gärtner[4]

[1]Fraunhofer IAIS
[2]Competence Center Machine Learning Rhine-Ruhr
[3]University of Bonn
[4]University of Nottingham
{sven.giesselbach,daniel.paurat}@iais.fraunhofer.de
{ullrich@iai,kamp@cs}.uni-bonn.de
thomas.gaertner@notthingham.ac.uk

Abstract

We propose a novel transfer learning approach for orphan screening called corresponding projections. In orphan screening the learning task is to predict the binding affinities of compounds to an orphan protein, i.e., one for which no training data is available. The identification of compounds with high affinity is a central concern in medicine since it can be used for drug discovery and design. Given a set of prediction models for proteins with labelled training data and a similarity between the proteins, corresponding projections constructs a model for the orphan protein from them such that the similarity between models resembles the one between proteins. Under the assumption that the similarity resemblance holds, we derive an efficient algorithm for kernel methods We empirically show that the approach outperforms the state-of-the-art in orphan screening.

1 Introduction

This paper proposes an approach to predicting binding affinities of compounds to a protein without training data. In biological organisms the bindings of small compounds to proteins induce subsequent cellular reactions. The strength of this binding is expressed via a real-valued *affinity*. In the context of affinity prediction, we refer to compounds as *ligands*. Protein-ligand-complexes regulate a variety of biochemical processes, e.g., the effectiveness of transporters, ion channels, hormones, receptors, and enzymes. To know whether or how strong a ligand binds to a protein is crucial for *drug discovery* and *design*. The identification of ligands with high affinity for new drug substances is a central concern in medicine [8]. Via high-throughput screening (HTS) machinery in laboratories one is able to determine ligand affinities practically. In order to supplement this time-consuming and cost-intensive procedure, molecular databases can be screened with computational methods [1, 6, 11, 24], denoted *in-silico virtual screening* [19].

Ligands can be represented by *molecular fingerprints* [2], i.e., vectorial representations that comprise their structural or physico-chemical information. For a protein with a training set of ligands and their binding affinities, this allows to train a prediction model using similarity search [5] and various machine learning approaches, e.g., random forests or neural networks [3, 9, 10, 12]. The most prominent and successful methods so far are *support vector machines (SVM)* using molecular fingerprints [5, 13, 20, 21, 22].

This paper tackles an even more challenging task called *orphan screening* [23]. It describes the prediction of ligand affinities for proteins without known ligand affinity values (*orphan targets*), such as

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

G-protein coupled receptors (GPCRs) which are popular drug targets with few or no identified ligands [4, 8]. Note that in previous work on orphan screening (see, e.g., [3, 4, 5, 8, 14, 23]), only those compounds with an affinity above a predifined threshold are called ligands and the task is to classifiy a compound as ligand or non-ligand. In the present work, we consider the regression case, i.e., the direct prediction of the binding affinities. Thus, for simplicity we refer to every compound as ligand.

The state-of-the art in orphan screening is the application of support vector machines with *target-ligand-kernels* (TLK) [7]. The TLK is a tensor product of a target kernel and a ligand kernel. Thus, it serves as similarity measure for target-ligand-pairs and allows for a simultaneous screening of proteins and ligands in *chemogenomics* [8].

We present a novel transfer learning approach [15] called *corresponding projections* (CP) for orphan screening. The idea behind CP is to relate the projections of proteins to the projections of the corresponding prediction models. Their relationship is used to generate a model for the orphan target. Just like the TLK approach, the CP requires that labelled training information for other proteins (*supervised targets*) is available for which a prediction model can be trained. Subsequently, in the actual CP optimisation step a prediction model is assigned to the orphan target such that its relationship to the given models resembles the relationship of the orphan protein to the given ones (see Fig. 1). The following section defines CP for affinity prediction using *support vector regression* (SVR) as supervised training method.

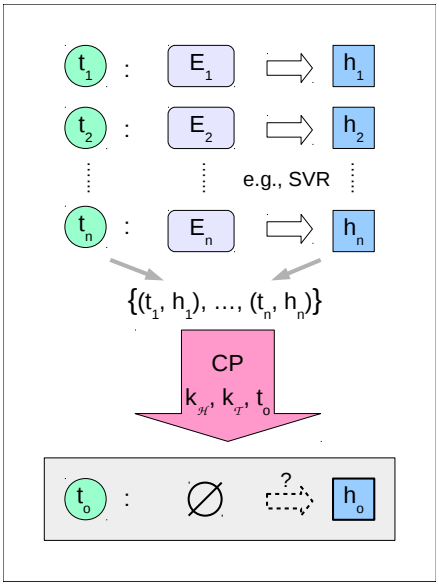


Figure 1: Overview of the corresponding projections approach (CP). The goal is to find a hypothesis h_o for the orphan target t_o using a set of supervised targets $t_1, \ldots, t_n$ with labelled training sets $E_1, \ldots, E_n$. For each t_i with $i \in [n]$, a model h_i is trained using E_i, e.g., by SVR. The CP algorithm uses these proteins and models, as well as a similarity measure $k_{\mathcal{T}}$ between targets and a similarity measure $k_{\mathcal{H}}$ between models to construct h_o.

In Sec. 3 the CP approach is empirically evaluated on an orphan screening task and compared to TLK and other baselines, including a simplified variant of CP, i.e., a target similarity-weighted sum of supervised models introduced in [5]. Sec. 4 concludes the paper. The solutions of CP variants with proof and further practical result can be found in the appendix.

2 Corresponding Projections

Given a set of $n \in \mathbb{N}$ target proteins $t_1, \ldots, t_n \in \mathcal{T}$ with corresponding training sets $E_1, \ldots, E_n$, each consisting of labelled examples

$$E_i = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$$

with $i \in [n]$, a model $h_i : \mathcal{X} \mapsto \mathcal{Y} \in \mathcal{H}$ can be trained for each target, e.g., via SVR. Thus, we obtain pairs $(t_1, h_1), \ldots, (t_n, h_n) \in \mathcal{T} \times \mathcal{H}$. Furthermore, let $t_o \in \mathcal{T}$ denote the orphan protein for which we want to infer a model $h_o \in \mathcal{H}$.

For that, let $\langle , \rangle_{\mathcal{T}}$ and $\langle , \rangle_{\mathcal{H}}$ be inner products with associated norms for targets $t \in \mathcal{T}$ and models $h \in \mathcal{H}$. We assume that the projection of proteins resembles the projection of models, i.e.,

$$\frac{\langle t_i, t_o \rangle_{\mathcal{T}}}{\|t_i\|_{\mathcal{T}}} \approx \frac{\langle f(t_i), f(t_o) \rangle_{\mathcal{H}}}{\|f(t_i)\|_{\mathcal{H}}} \tag{1}$$

for every supervised target t_i. With this, finding h_o can be solved by a least squares approach.

Definition 1. *The model for the orphan protein h_o is given by $h_o = \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{Q}_o(h)$, where*

$$\mathcal{Q}_o(h) = \nu \|h\|_{\mathcal{H}}^2 + \sum_{i=1}^{n} \left| \langle h, h_i \rangle_{\mathcal{H}} \|t_i\|_{\mathcal{T}} - \langle t_o, t_i \rangle_{\mathcal{T}} \|h_i\|_{\mathcal{H}} \right|^2, \tag{2}$$

with trade-off parameter $\nu \geq 0$. The optimisation in (2) is called corresponding projections (CP).

If $\mathcal{H}$ is a Hilbert space, the model h_o lies in the span of the supervised hypotheses $h_1, \ldots, h_n$, i.e., $h_o = \sum_{i=1}^{n} \beta_{oi} h_i$ for appropriate $\beta_o \in \mathbb{R}^n$. This parametrisation of h_o leads us to a solution of CP.

Lemma 2. *Let $\mathcal{H}$ be a Hilbert space and $k_\mathcal{T}$ a similarity measure on the target space $\mathcal{T}$. CP in (2) can be solved via*

$$\beta_o = [\nu G + GNG]^\dagger G \rho_o,$$

where $[M]^\dagger$ denotes the Moore-Penrose inverse of matrix M and

$$\begin{aligned} N &= \mathrm{diag}(\{k_\mathcal{T}(t_i, t_i)\}_{i=1}^n) \\ G &= \{\langle h_i, h_j \rangle_\mathcal{H}\}_{i,j=1}^n \\ \rho_o &= \{\sqrt{k_\mathcal{T}(t_i, t_i)}\, k_\mathcal{T}(t_o, t_i) \|h_i\|_\mathcal{H}\}_{i=1}^n \quad . \end{aligned}$$

The proof of Lemma 2 and more theoretical details can be found in Appendix A.1.

3 Empirical Evaluation

We evaluate the proposed CP approach on an orphan screening task and compare it to state-of-the art baselines. For that, we use 9 protein-ligand datasets extracted from BindingDB (bindingdb.org). Each dataset corresponds to a human protein with peptidase domain and comprises between 240 and 2649 ligands with affinity labels ($\mathrm{p}K_i$-values) towards the respective protein. We utilise the standard molecular fingerprint ECFP4 [16] for the representation of ligands. Target similarities were calculated from the amino acid sequence similarity of the peptidase domain. We normalise the similarities of all targets except for the respective orphan to sum to 1.

The parameter ν of the optimisation problem (Eq. 10) was determined on an independent training set and fixed to $\nu = 5$ for all orphan targets. To improve the numerical stability of inversion of $\nu G + GNG$, we add a constant $\lambda \in \mathbb{R}_+$ to the diagonal of the matrix. That is, we consider a slightly different CP optimisation

$$\beta_o = [\nu G + \lambda \mathbf{I}_n + GNG]^\dagger G \rho_o \quad ,$$

with $\lambda = 1$ for all orphan targets [1].

In order to perform orphan screening, we perform a leave-one-out cross validation over all proteins, hence considering each protein as orphan once. We report the *root mean squared error* (RMSE) of the predicted affinities. In order to test for stability of the method, we create 10 distinct draws of the dataset by randomly sampling 240 ligands per protein. The reported RMSE is an average over the 10 draws per orphan protein. In accordance with state-of-the-art approaches we use the linear kernel for our experiments. The models for the supervised targets are trained using SVR. A 3-fold cross-validation is used to obtain the best hyperparameters for each target with parameter ranges $\epsilon \in \{0.1, 0.01, 0.001\}$ and regularisation parameter $C \in \{2^{-i} : i \in \{-5, -4, \ldots, 4, 5\}\}$. More details on the experimental setup can be found in section A.2.

Fig. 2 shows the RMSE of all approaches averaged over all orphan proteins and all draws. CP achieves a median RMSE of 2.197. We compare CP with trivial and state-of-the-art baselines:

A naive way of combining target models without considering similarities of targets is to build a simple average over them and use the averaged model (Avg) to predict the ligand affinities for the orphan target. However, we find that Avg performs significantly worse than CP with a median RMSE of 3.610.

The most straight-forward way of using protein similarities is to choose models of other targets according to their similarity to the orphan target. To understand the range of how proximity of proteins affects the prediction quality we evaluate the performance of models from the targets closest to and farthest from the orphan. The model of the most similar target (Closest Protein) outperforms Avg, while its median RMSE of 2.427 is significantly higher than the median RMSE of CP. Note that using the model of the least similar protein (Farthest Protein) yields a median RMSE of 3.203, worse than CP but still better than Avg.

The fact that Closest Protein performs better supports the intuition that the orphan and targets closer to it share similar traits which determine the affinities of ligands. Reducing Avg to a model using just

[1] The code can be found at bitbucket.org/grumpy_kat/corresponding-projections.

the 3 closest targets (Avg-Clo-3) yields a significant performance boost, indicating that it benefits from the focus on closer proteins. With a median RMSE of 2.260 it outperforms Closest Protein.

Rather than completely omitting the targets which are not among the 3 closest, the state-of-the-art approach TLK incorporates target similarities via features which are composed of joint target and ligand kernels. The hyperparameters for TLK are again optimised via grid search and 3-fold crossvalidation on all supervised targets using the same parameter ranges as above. TLK achieves a median RMSE of 2.934, placing it between Closest and Farthest. The TLK-variant (TLK-Clo-3) which also uses just the 3 closest proteins suffers from reduced training data size and achieves a median RMSE of 3.708, worse than the original TLK and even worse than Avg.

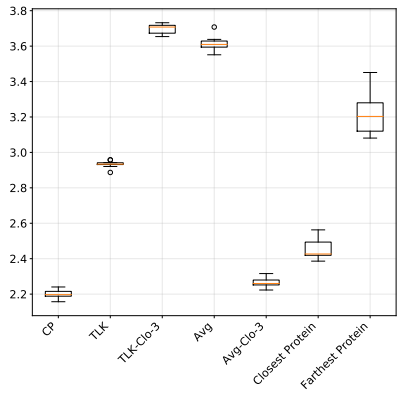


Figure 2: RMSEs of the proposed CP approach for all 9 proteins and 10 draws in comparison to the baselines.

Other than the previous approaches, CP optimizss the weights of each model to resemble the similarities of the targets, thereby making use of all available models. The results suggest that incorporating the similarities in the target and hypothesis space is beneficial: Not only is CP's median RMSE lower that that of simpler schemes like Avg-Clo-3, it also outperforms the state-of-the-art approach TLK.

The approach closest to CP is a simplified weighted average of models (for further details see Def. 6 in Sec. A.1 of the appendix). Simplified outperforms the other baseline approaches, but Fig. 3 shows that it performs worse than CP, indicating that the optimisation step of CP yields a better model than the plain integration of the similarities.

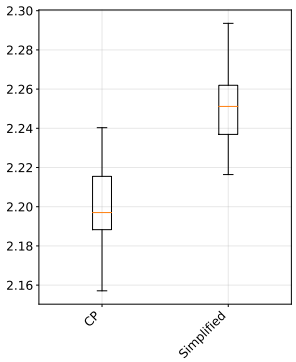


Figure 3: The RMSE of Simplified and CP averaged over all proteins and draws.

4 Conclusion and Future Work

We introduced corresponding projections, a transfer learning based approach to hypothesis finding for unlabelled target domains that incorporates target and hypothesis similarities to derive a model for an orphan target. We applied the algorithm to the problem of target-ligand affinity prediction and empirically demonstrated its superior performance over the current state-of-the-art approach of using support vector regression with target-ligand-kernels.

For future work we will expand our evaluation of corresponding projections to further tasks, such as domain adaptation in natural language processing. We will also investigate how the proximity of targets affects the quality of predictions.

Acknowledgments

We want to thank Stefan Rüping and Stefan Wrobel for their valuable input and fruitful discussions. This research has been funded by the German Federal Ministry of Education and Research, Foerderkennzeichen 01S18038B.

References

[1] Qurrat U. Ain, Antoniya Aleksandrova, Florian D. Roessler, and Pedro J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 5(6):405–424, 2015.

[2] Andreas Bender, Jeremy L. Jenkins, Josef Scheiber, Sai Chetan K. Sukuru, Meir Glick, and John W. Davies. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.*, 49(1):108–119, 2009.

[3] Joel R. Bock and David A. Gough. Virtual Screen for Ligands of Orphan G Protein-Coupled Receptors. *J. Chem. Inf. Model.*, 45(5):1402–1414, 2005.

[4] Dumitru Erhan, Pierre-Jean L'Heureux, Shi Yi Yue, and Yoshua Bengio. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.*, 46(2):626–635, 2006.

[5] Hanna Geppert, Jens Humrich, Dagmar Stumpfe, Thomas Gärtner, and Jürgen Bajorath. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.*, 49(4):767–779, 2009.

[6] Hanna Geppert, Martin Vogt, and Jürgen Bajorath. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.*, 50(2):205–216, 2010.

[7] Laurent Jacob and Jean-Philippe Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.

[8] Laurent Jacob, Brice Hoffmann, Véronique Stoven, and Jean-Philippe Vert. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinformatics*, 9(1):363–379, 2008.

[9] José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis. K_{DEEP}: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.*, 58(2):287–296, 2018.

[10] Indra Kundu, Goutam Paul, and Raja Banerjee. A machine learning approach towards the prediction of protein-ligand binding affinity based on fundamental molecular properties. *RSC Adv.*, 8(22):12127–12137, 2018.

[11] Liwei Li, Bo Wang, and Samy O. Meroueh. Support Vector Regression Scoring of Receptor-Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *Journal of chemical information and modeling*, 51(9):2132–2138, 2011.

[12] Yu-Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altmann. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1583–1546, 2018.

[13] Andreas Maunz and Christoph Helma. Prediction of chemical toxicity with local support vector regression and activity-specific kernels. *SAR QSAR Environ. Res.*, 19(5-6):413–431, 2008.

[14] Xia Ning, Huzefa Rangwala, and George Karypis. Multi-Assay-based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets. *Journal of chemical information and modeling*, 49(11): 2444–2456, 2009.

[15] Sinno J. Pan, Qiang Yang, et al. A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[16] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

[17] Bernhard Schölkopf, Ralf Herbrich, Alex J. Smola, and Robert Williamson. A Generalized Representer Theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[18] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

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

[20] Nobuyoshi Sugaya. Ligand Efficiency-Based Support Vector Regression Models for Predicting Bioactivities of Ligands to Drug Target Proteins. *J. Chem. Inf. Model.*, 54(10):2751–2763, 2014.

[21] Katrin Ullrich, Jennifer Mack, and Pascal Welke. Ligand Affinity Prediction with Multi-Pattern Kernels. *Proceedings of the International Conference on Discovery Science*, pages 474–489, 2016.

[22] Katrin Ullrich, Michael Kamp, Thomas Gärtner, Martin Vogt, and Stefan Wrobel. Co-regularised support vector regression. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 338–354, 2017.

[23] Anne M. Wassermann, Hanna Geppert, and Jürgen Bajorath. Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels Is Dominated by Nearest Neighbor Effects. *J. Chem. Inf. Model.*, 49(10):2155–2167, 2009.

[24] Hongyi Zhou and Jeffrey Skolnick. FINDSITE^X: A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Mol. Pharmaceutics*, 9(6):1775–1784, 2012.

A Appendix

A.1 Corresponding Projections Theory

In this theoretical part of the appendix we present variants of the CP optimisation as defined in Def. 1 as well as their corresponding solutions. We start with a linear version of corresponding projections, then we derive a simplified version and lastly we introduce the more general non-linear corresponding projection which we evaluated in this paper.

Linear and Simplified Algorithm

At first we turn towards the case $\mathcal{H} = \mathbb{R}^d$ with the canonical inner product, i.e., we consider linear functions of d-dimensional input vectors. We denote this CP version *linear corresponding projections* (LCP). As a preparation for the following result we define the matrices $H \in \mathbb{R}^{d\times n}$ and $N \in \mathbb{R}^{n\times n}$ via

$$H = (h_1|\cdots|h_n), \quad N = \operatorname{diag}(k_{\mathcal{T}}(t_i, t_i)), \tag{3}$$

as well as the vectors $\rho_o, \delta_o \in \mathbb{R}^n$ with

$$\{\delta_o\}_i = k_{\mathcal{T}}(t_o, t_i)\|h_i\|_d, \quad \{\rho_o\}_i = \sqrt{k_{\mathcal{T}}(t_i, t_i)}\{\delta_o\}_i, \tag{4}$$

where $\|\cdot\|_d$ is the Euclidean norm in $\mathbb{R}^d$.

Lemma 3. *Let $(t_1, h_1), \ldots, (t_n, h_n) \in \mathcal{T}\times\mathcal{H}$ be examples of targets and corresponding hypotheses. If $\mathcal{H} = \mathbb{R}^d$ and $k_{\mathcal{T}}$ a similarity measure, LCP can be solved as follows*

$$f(t_o) = \left[\nu\mathbf{I}_d + \sum_{i=1}^{n} h_i k_{\mathcal{T}}(t_i, t_i) h_i^T\right]^{\dagger} \cdot \sum_{i=1}^{n} h_i \|h_i\|_d \sqrt{k_{\mathcal{T}}(t_i, t_i)} k_{\mathcal{T}}(t_o, t_i), \tag{5}$$

where † is the Moore-Penrose inverse of a square matrix.

Proof. We formulate the objective $\mathcal{Q}_o(h)$ in Def. 1 with H, N, ρ_o, and δ_o from above.

$$\mathcal{Q}_o(h) = \nu h^T h + h^T HNH^T h - 2h^T H\rho_o + {\delta_o}^T\delta_o$$

The solution of LCP in (5) can be derived by setting the gradient of $\mathcal{Q}_o(h)$

$$\frac{\partial\mathcal{Q}_o}{\partial h} = 2\nu h + 2HNH^T h - 2H\rho_o$$

equal to zero. We obtain $h_o = [\nu\mathbf{I}_d + HNH^T]^{\dagger}H\rho_o$. □

As the matrix HNH from above is positive semi-definite, the inverse $[\nu\mathbf{I}_d + HNH]^{-1}$ always exists if ν is positive. Otherwise, the more general $[\cdot]^{\dagger}$ will be applied.

Subsequent to the linear version LCP we define a simplified CP variant.

Definition 4. *Let $(t_1, h_1), \ldots, (t_n, h_n) \in \mathcal{T}\times\mathcal{H}$ be supervised targets and their hypotheses. For an arbitrary similarity function $k_{\mathcal{T}}$ on targets we define simplified corresponding projections (SCP) according*

$$f(t_o) = \sum_{i=1}^{n} h_i \frac{k_{\mathcal{T}}(t_o, t_i)}{\sqrt{k_{\mathcal{T}}(t_i, t_i)}}. \tag{6}$$

In the naïve SCP approach the orphan hypothesis h_o is a linear combination of supervised hypotheses h_i with coefficients that have not to be learned beforehand. The coefficients in (6) are essentially the left hand side of the CP initial equation in (1). Therefore, the complexity of SCP is only $\mathcal{O}(|T|d\kappa)$ if the cost for $k_{\mathcal{T}}$ is bounded by κ. In contrast, the complexity for the calculation of (5) is $\mathcal{O}(|T|d^2\kappa)$. Actually, for SCP the candidate space $\mathcal{H}$ is not necessarily equal to $\mathbb{R}^d$, but an arbitrary function

space. A similar approach to SCP for classification already appeard in [5], where the authors also applied a weighted sum of predictors denoted as *SVM linear combination* (SVM-LC).

Non-Linear Corresponding Projections

In the last section we considered a linear as well as a simplified version of CP. Now we want to exploit that $\mathcal{H}$ is a Hilbert space with general inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$. In this scenario we conclude a representation of h_o as linear combination

$$h_o = \sum_{i=1}^{n} \beta_{oi} h_i \quad , \quad \beta_o \in \mathbb{R}^n, \tag{7}$$

i.e, h_o lies in the span of the supervised hypotheses $h_1, \ldots, h_n$. This can be shown with an argumentation similar to the proof of the *representer theorem* (RT) [17]. To this aim, let us consider the decomposition $h_o = s + g$, where $s \in span\{h_1, \ldots, h_n\}$ and $\langle g, s' \rangle_{\mathcal{H}} = 0$ for all $s' \in span\{h_1, \ldots, h_n\}$. Then we obtain

$$\begin{aligned} &= \nu \|s+g\|^2 + \sum_{i=1}^{n} [\langle s+g, h_i \rangle \sqrt{k_{\mathcal{T}}(t_i, t_i)} \\ &\qquad - k_{\mathcal{T}}(t_o, t_i) \|h_i\| \,]^2 \\ &\geq \nu \|s\|^2 + \sum_{i=1}^{n} [\langle s, h_i \rangle \sqrt{k_{\mathcal{T}}(t_i, t_i)} - k_{\mathcal{T}}(t_o, t_i) \|h_i\| \,]^2, \end{aligned}$$

which shows the claim. Analogous to (3) and (4), we consider matrices $G, N \in \mathbb{R}^{n \times n}$ with general inner product

$$\{G\}_{i,j} = \langle h_i, h_j \rangle_{\mathcal{H}}, \quad N = \mathrm{diag}(k_{\mathcal{T}}(t_i, t_i)) \tag{8}$$

and vectors $\rho_o, \delta_o \in \mathbb{R}^n$

$$\{\delta_o\}_i = k_{\mathcal{T}}(t_o, t_i) \|h_i\|_{\mathcal{H}}, \quad \{\rho_o\}_i = \sqrt{k_{\mathcal{T}}(t_i, t_i)} \, \{\delta_o\}_i. \tag{9}$$

We identify h_o with its defining vector $\beta_o \in \mathbb{R}^n$.

Lemma 5. *Let $\mathcal{H}$ be a Hilbert space and $k_{\mathcal{T}}$ a similarity measure on the target space $\mathcal{T}$. With the representation of h_o in (7) CP from Def. 1 can be solved as*

$$\beta_o = [\nu G + GNG]^{\dagger} G \rho_o \ , \tag{10}$$

where N, G, and ρ_o are defined as in (8) and (9).

Proof. With (7), (8), and (9) the objective in Def. 1 can be written

$$\mathcal{Q}_o(\beta) = \nu \beta^T G \beta + \beta^T G N G \beta - 2 \beta^T G \rho_o + (\delta_o)^T \delta_o.$$

Its gradient $\partial \mathcal{Q}_o / \partial \beta$ set to zero shows $\beta_o = [\nu G + GNG]^{\dagger} G \rho_o$. □

This constitutes the approach used throughout the paper. Solving Eq. 10 requires inverting an $n \times n$ matrix and thus has a computational complexity of $\mathcal{O}(n^3)$.

Note that in practice it can happen that the the union of ligands for all supervised targets q is smaller than the number supervised targets n. This can happen, e.g., if every supervised target has the same small training set of ligands and n is larger than this training set. For this case, the CP solution according to Eq. 10 can be rewritten such that it can be solved in time $\mathcal{O}(q^3)$. For that we require that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS)

$$\mathcal{H} = \left\{ h(\cdot) = \sum_{i=1}^{\infty} \pi_i k_{\mathcal{H}}(x_i, \cdot) : x_i \in \mathcal{X}, \alpha_i \in \mathbb{R} \right\},$$

where $k_{\mathcal{H}}$ defined on $\mathcal{X} \times \mathcal{X}$ is the reproducing kernel of $\mathcal{H}$ (for more details see, e.g., [17, 18]). Assume now that each hypothesis h_i arised from a training process with training examples from $\mathcal{X} \times \mathcal{Y}$ solving a regularised cost function like the one applied for *regularised empirical risk minimisation*. Actually, the set of training instances and according labels depends on the respective *supervised*

target t_i. Let $\{x_1, \ldots, x_q\}$ be the union of the training instances for all targets t_i, $i = 1, \ldots, n$, and K the Gram matrix of kernel $k_{\mathcal{H}}$. The parameterised representation of each hypothesis h_i

$$h_i(x) = \sum_{j=1}^{q} \pi_{ij} k_{\mathcal{H}}(x_j, x) \quad , \quad x \in \mathcal{X}, \pi_i \in \mathbb{R}^q, \tag{11}$$

exists according to the RT. If x_j was not in the set of the original training instances of t_i the parameter π_{ij} is just equal to zero. We cannot apply the RT for the orphan target t_o and its hypothesis h_o directly because of the lack of training examples. However, h_o can be represented equivalently with coefficients $\pi_o \in \mathbb{R}^q$ as we have the representation in (7)

$$\begin{aligned} h_o(x) &= \sum_{i=1}^{n} \beta_{oi} h_i(x) = \sum_{i=1}^{n} \beta_{oi} \left(\sum_{j=1}^{q} \pi_{ij} k_{\mathcal{H}}(x_j, x) \right) \\ &= \sum_{j=1}^{q} \left(\sum_{i=1}^{n} \beta_{oi} \pi_{ij} \right) k_{\mathcal{H}}(x_j, x) = \sum_{j=1}^{q} \pi_{oj} k_{\mathcal{H}}(x_j, x), \end{aligned} \tag{12}$$

where $\beta_o \in \mathbb{R}^n$ and $\pi_i, \pi_o \in \mathbb{R}^q$. Hence, with $\Pi = (\pi_1 | \cdots | \pi_n)$ the coefficients of the orphan target are $\pi_o = \Pi \beta_o$. Analogous to the CP solution in (10) we define the matrices $\tilde{G} \in \mathbb{R}^{q \times n}$ and $N \in \mathbb{R}^{n \times n}$

$$\tilde{G} = K\Pi, \quad N = \operatorname{diag}(k_{\mathcal{T}}(t_i, t_i)) \tag{13}$$

and vectors $\tilde{\rho}_o, \tilde{\delta}_o \in \mathbb{R}^n$

$$\{\tilde{\delta}_o\}_i = k_{\mathcal{T}}(t_o, t_i) \sqrt{\pi_i^T K \pi_i}, \quad \{\tilde{\rho}_o\}_i = \sqrt{k_{\mathcal{T}}(t_i, t_i)} \, \{\tilde{\delta}_o\}_i \tag{14}$$

Again, we identify h_o with its vector of coefficients π_o.

Lemma 6. *Let h_i, $i = 1, \ldots, n$, and h_o have the representations (11) and (12) from above with $\pi_i, \pi_o \in \mathbb{R}^q$. With $k_{\mathcal{H}}$ and $k_{\mathcal{T}}$ we denote the reproducing kernel of $\mathcal{H}$ and a similarity measure on the target space $\mathcal{T}$, as well as K be the Gram matrix with respect to $x_1, \ldots, x_q$. With (11) and (12) CP can be solved as*

$$\begin{aligned} \pi_o = &\left[\nu K + \sum_{i=1}^{n} K \pi_i k_{\mathcal{T}}(t_i, t_i) \pi_i^T K \right]^{\dagger} \\ &\cdot \sum_{i=1}^{n} \left(\sqrt{\pi_i^T K \pi_i} \sqrt{k_{\mathcal{T}}(t_i, t_i)} k_{\mathcal{T}}(t_o, t_i) \right) K \pi_i, \end{aligned} \tag{15}$$

where $\nu \geq 0$ and t_i, $i = 1, \ldots, n$, are the supervised targets. We call this approach kernel corresponding projections (KCP).

Proof. The proof is again a consequence of the stationarity of the gradient of the parameterised objective

$$\mathcal{Q}_o(\pi) = \nu \pi^T K \pi + \pi^T \tilde{G} N \tilde{G}^T \pi - 2\pi^T \tilde{G} \tilde{\rho}_o + \tilde{\delta}_o{}^T \tilde{\delta}_o$$

from Def. (1). □

Solving Eq. 15 requires inverting a $q \times q$ matrix and thus has a computational complexity in $\mathcal{O}(q^3)$. As mentioned above, this is preferable to solving Eq. 10 for $n \gg q$.

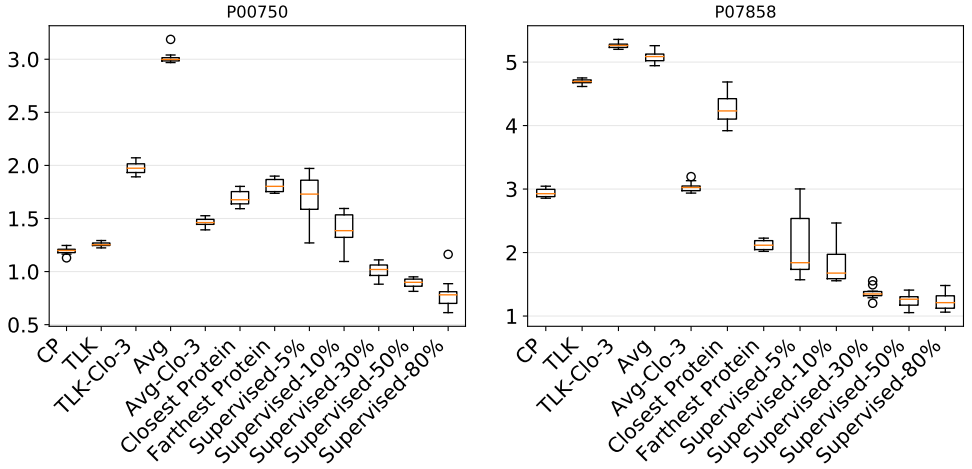
A.2 Extended Experimental Results

In this section we provide information on the results with supervised baselines.

We also evaluated a hypothetical supervised case in which we trained an SVR on several fractions ($5\%, 10\%, 30\%, 50\%$, and 80%) of the available ligands, testing it on remaining ones. This yields an upper bound for the RMSE. On average all supervised models clearly outperform CP. However,

Method	Median RMSE
CP	2.197
Supervised-5%	1.797
Supervised-10%	1.440
Supervised-30%	1.140
Supervised-50%	1.038
Supervised-80%	0.946

Table 1: Median RMSE of CP and the supervised approaches over all draws, averaged over all proteins.



(a) (b)

Figure 4: The RMSE of the unsupervised and supervised approaches over all draws for two proteins protein *P00750* and *P07858*.

they solve a different task, since CP assumes no labelled data for the orphan target. Notably, the performance of CP is a lot closer to the supervised approaches, than TLK. For Supervised 50% the median RMSE of CP is larger than Supervised 50% by a factor of 2, while TLKs median RMSE is larger by a factor of almost 3. A more detailed investigation of the results, shows that CPs performance varies strongly depending on the orphan target. Fig. 4(a) shows the performance for the protein named *P00750*. We can observe that CP outperforms both the supervised model trained on 5% and on 10% of the data, reaching a performance close to a supervised model trained on 30% of the data. Contrary Fig. 4(b) shows that for protein *P07858* CP is worse than all supervised approaches and the model of the farthest protein performs best out of all unsupervised approaches, nearly as good as the supervised model trained on 10% of the data. In total CP performs better than Supervised-5% for 4 proteins, better than Supervised-10% for 2 proteins, comparable to Supervised-5% for 1 protein and significantly worse than all supervised approaches for 4 proteins.

This raises open research questions left for future work. One question is whether the chosen similarity measure for target similarities is suitable or whether other similarity measures would perform better. Another one is whether the performance of CP increases with the number of ligands per protein or the total number of proteins. Since the performance of CP varies between proteins, another question is whether this performance can be related to the similarity of the orphan target and supervised ones.