Orphan Principal Component Analysis

Diogo Braga
University of Technology
Vienna, Austria
e12007525@student.tuwien.ac.at

Maximilian Thiessen
University of Technology
Vienna, Austria
maximilian.thiessen@tuwien.ac.at

Abstract

In orphan screening, the learning task is to predict binding affinities between the components of a protein for which there are no supervised data available. With a set of proteins with labelled data and respective predictive models, it is possible to build a model for the orphan protein through transfer learning techniques. The method presented in this paper is based on the principal component analysis algorithm with must-link constraints between similar instances, in which the proteins and models are embedded in the same space and try to find the closest model for the orphan protein. With this algorithm, we have an alternative to solve problems related to drug discovery and design.

1 Introduction

Orphan screening addresses the problem of searching for new ligands in orphan proteins, i.e., the prediction of affinities for a target protein with no labelled examples. To compensate for the lack of data, labelled information of other targets and relations between them is used to infer an appropriate predictive model for the orphan target. This corresponding hypothesis can be found via a kernel method, in this case, a similarity measure, using the labelled examples for training. On the opposite, a target without labelled training information is called orphan target, and we try to learn an orphan hypothesis in order to be able to predict it.

Different molecules have different shapes, and these shapes usually play a crucial role in the behaviour of the molecules in the human body and other living things. The strength of the bindings of compounds is expressed via a real value affinity. In the medical context, identifying compounds with high affinity is a central concern since it can be used for drug discovery and design. Although large molecular compounds are already known, and their protein-ligand information is described in molecular databases, orphan proteins still exist, as the number of functional proteins in biological organisms is enormous and, therefore, can lead to discoveries of new proteins.

This paper will present a brief introduction to state-of-art approaches for solving this problem, mainly, the *Corresponding Projections* (CP) [Giesselbach et al., 2018]. After reflecting on this approach, we present *Orphan Principal Component Analysis* (OPCA) as a variant of interactive knowledge-based PCA [Oglic et al., 2014]. The focus of this paper is to compare this new OPCA approach with state-of-art approaches in practical terms and to check if improvements are taking into account the distinct theoretical foundations on which they are based. The theoretical part of this paper is based on the research developed by Katrin Ulrich.

2 Corresponding Projections

Our goal is to search for a predictor h from an hypothesis space $\mathcal{H}$ that maps instances from $\mathcal{X}$ to labels from $\mathcal{Y}$. For predict the ligand affinity, the main learning task is to find the binding affinity model $h_t : \mathcal{X} \to \mathcal{Y}$ with respect to a protein target t in the target space $\mathcal{T}$. For that, we consider

TUML Project (2020/21).

learning a function $f : \mathcal{T} \to \mathcal{H}$ that assigns a binding model h_t to each target t. A target $t \in \mathcal{T}$ is called *supervised target* if there is labelled training data from $\mathcal{X} \times \mathcal{Y}$ to solve the main learning task for t. The corresponding *supervised hypothesis* h_t can be found via a kernel method using the labelled examples for training. On the opposite, we have a target $t_0 \in \mathcal{T}$ without label information (orphan target) and we use transfer learning to learn an orphan hypothesis $h_0 \in \mathcal{H}$ for t_0. Knowing f, the orphan hypothesis h_0 can be defined via

$$h_0 = f(t_0). \tag{1}$$

This process is demonstrated in figure 1. In this algorithm, transfer learning proves to be a fundamental technique in solving the problem. Therefore, the corresponding projections (CP) are essential because they store the labelled information from the existing targets and then apply that information to the orphan targets, those without label information. The empirical risk objective, in this particular case, is defined by the comparison of the projections of targets and hypothesis, in order to find the best orphan hypothesis h_0. A good orphan hypothesis h_0 can be found if we search the corresponding projections to be nearly equal in the sense of

$$\frac{k_{\mathcal{T}}(t_i, t_0)}{\sqrt{k_{\mathcal{T}}(t_i, t_i)}} \approx \frac{\langle f(t_i), f(t_0)\rangle_{\mathcal{H}}}{\|f(t_i)\|_{\mathcal{H}}} = \frac{\langle h_i, h_0\rangle_{\mathcal{H}}}{\|h_i\|_{\mathcal{H}}} \tag{2}$$

for $i = 1, ..., n$. Since we want to explore the similarities between the targets and hypothesis, we try to minimise the loss term for the left and right hand side of Equation (2)

$$\boldsymbol{\ell}\big(k_{\mathcal{T}}(t_i, t_0)\,\|f(t_i)\|_{\mathcal{H}}\,, \langle f(t_i), f(t_0)\rangle_{\mathcal{H}} \sqrt{k_{\mathcal{T}}(t_i, t_i)}\big) \tag{3}$$

for $i = 1, ..., n$, where $\boldsymbol{\ell} : \mathbb{R}^2 \to \mathbb{R}^+$. Therefore, considering the loss function (in this case, the squared loss $\boldsymbol{\ell}_2$), we want to find a hypothesis h_t from a hypothesis space $\mathcal{H}$ concerning a target t. Consider $(t_1, h_1), ..., (t_n, h_n) \in \mathcal{T} \times \mathcal{H}$ as the supervised targets and the respective supervised hypothesis, and $t_0 \in \mathcal{T}$ as the orphan target. Assume $k_{\mathcal{T}}$ to be a kernel function for targets and $\mathcal{H}$ as a Hilbert space with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The following optimization indicates the corresponding projections algorithm that solves orphan screening for an orphan target:

$$f(t_0) = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, v\,\|h\|^2_{\mathcal{H}} + \sum_{i=1}^{n} \Big(\langle h, h_i\rangle_{\mathcal{H}} \sqrt{k_{\mathcal{T}}(t_i, t_i)} - k_{\mathcal{T}}(t_0, t_i)\,\|h_i\|_{\mathcal{H}} \Big)^2 \tag{4}$$

where $v > 0$ is a hyperparameter.

The different variants developed based on the corresponding projections can be found in the work developed in Giesselbach et al. [2018]. The corresponding projections are described here to provide the idea of the learning scenario present in the orphan screening problem. In this paper, we are going to focus and carry out experiments for the OPCA, the following approach presented.

3 Orphan Principal Component Analysis

In this scenario, $\mathcal{T}$ is a target space with a similarity measure (kernel function) $k_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and $\mathcal{H}$ is a hypothesis space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. This approach applies *principal component analysis* (PCA) for dimensionality reduction to infer a suitable hypothesis for the orphan target t_0. More precisely, this algorithm is related to *Interactive knowledge-based kernel principal component analysis* (IPCA) [Oglic et al., 2014] to facilitate the integration of knowledge from a focused domain to a visualisation process. This knowledge can be inserted through control points, classification constraints, must-link constraints, and cannot-link constraints into the PCA optimisation, resulting in the projection's calculation. As in CP, projections are essential in OPCA because that is where the labelled information is transferred to targets without labelled information. To find a hypothesis $h_0 \in \mathcal{H}$ for the orphan target t_0, we will use the IPCA algorithm with must-link constraints.

In this algorithm, the instance space $\mathcal{X}$ is defined as the union of targets and hypothesis

$$\mathcal{X} = \mathcal{T} \cup \mathcal{H}. \tag{5}$$

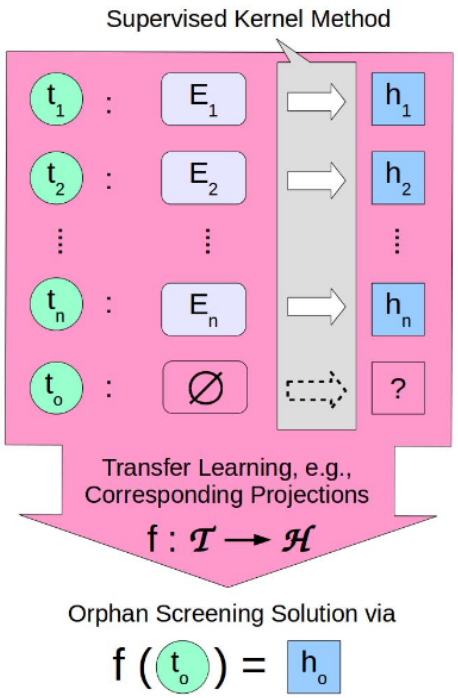

Figure 1: Overview of the orphan screening's learning scenario [from Giesselbach et al., 2018]

We define a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ via

$$k(x, x') = \begin{cases} k_{\mathcal{T}}(x, x') & : x, x' \in \mathcal{T} \\ \langle x, x' \rangle_{\mathcal{H}} & : x, x' \in \mathcal{H} \\ 0 & : \text{otherwise} \end{cases} \tag{6}$$

The kernel k is positive semi-definite as $k_{\mathcal{T}}$, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and the constant 0 are positive semi-definite as well. We fix a set

$$X_{\cup} = \{x_1, ..., x_p, x_{p+1}, ..., x_{p+q}\} \subseteq \mathcal{X}, \tag{7}$$

where $\{x_1, ..., x_p\} \subseteq \mathcal{T}$ are targets and $\{x_{p+1}, ..., x_{p+q}\} \subseteq \mathcal{H}$ are hypothesis. The set $X_{\cup}$ comprise the supervised targets t_i, the supervised hypothesis h_i, $i = 1, ..., n$, and the orphan target t_0. Let $K_{\mathcal{T}}$ be the Gram matrix of $k_{\mathcal{T}}$ with respect to $x_1, ..., x_p$ and $K_{\mathcal{H}}$ be the matrix

$$K_{\mathcal{H}} = (\langle x_i, x_j \rangle \mathcal{H})_{i,j=p+1}^{p+q}.$$

Joining both, the Gram matrix K of k with respect to $x_1, ..., x_p, x_{p+1}, ..., x_{p+q}$ equals

$$K = \begin{Bmatrix} K_{\mathcal{T}} & 0_{p \times q} \\ 0_{q \times p} & K_{\mathcal{H}} \end{Bmatrix} \subseteq \mathbb{R}^{D \times D}, \tag{8}$$

where $D = p + q$.

The idea of the IPCA is to make the dimensionality reduction technique of PCA more significant, with the addition of knowledge related to the domain. In this particular case, the *must-link constraints* describe the request that similar instances are supposed to have a small distance in the optimized IPCA projection's image space. In the orphan screening scenario, the must-link constraints require supervised targets t_i and their corresponding supervised hypothesis h_i to have a small distance

after projection for all $i = 1, ..., n$. These n constraints are represented with the following set $\mathcal{C}$ of instances index pairs

$$\mathcal{C} = \{(l, l') : x_l \text{ is a supervised target with supervised hypothesis } x_{l'}\}.$$

The must-link constraints related to IPCA are included in the definition of the optimal projection as follows [Oglic et al., 2014]

$$\max_{\Pi \in \mathbb{R}^{D \times d}} \text{tr}\Big(\frac{1}{D}\Pi^T K H_D K \Pi\Big) - v\text{tr}\Big(\frac{1}{\mathcal{C}}\Pi^T K L K \Pi\Big), \quad \text{s.t. } \Pi^T K \Pi = \mathbf{I}_d \tag{9}$$

where $v > 0$ is a trade-off hyperparameter. Matrices H_D and L are defined as

$$H_D = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_D(\mathbf{1}_D)^T$$

as well as

$$L = \sum_{(l,l') \in \mathcal{C}} (e_l - e_{l'})(e_l - e_{l'})^T,$$

where e_l is the unit vector in $\mathbb{R}^D$ such that the l-th component is equal to 1.

Let $\mathcal{X}$ and $K \in \mathbb{R}^{D \times D}$ be defined as in Equations (5) and (8). Furthermore, we fix the vector

$$K(x) = (k(x, x_1), ..., k(x, x_D)),$$

where $x_1, ..., x_D \in \mathcal{X}$ are the instances of $X_\cup$ according to Equation (7). Let $\Pi \in \mathbb{R}^{D \times d}$ be the solution of the optimisation in Equation (9). The *orphan principal component analysis* (OPCA) determine the orphan hypothesis h_0 for the orphan target t_0 via

$$h_0 = \underset{h \in \mathcal{H}'}{\text{argmin}} \left\| (K(h) - K(t_0))\Pi \right\|^2. \tag{10}$$

Concisely, OPCA solves the orphan screening problem by searching for the hypothesis in $\mathcal{H}$ with the shortest distance to the projection of the orphan target t_0. Apart from the supervised proteins' hypothesis $h_1, ..., h_n$, we can include more hypothesis, whether they are random hypothesis or linear combinations of the proteins' hypothesis. Thus, in the equation (10), $\mathcal{H}'$ appears instead of $\mathcal{H}$, since $\mathcal{H} \subset \mathcal{H}'$ and $\mathcal{H}'$ can include also include other hypotheses.

4 Experimental Results

This section presents the evaluation carried out on the new OPCA algorithm compared with the state-of-art approaches. The evaluation performed is based on the following research question:

- The unsupervised orphan screening task can be solved in corresponding projections through the assumption that similarities in the target space are the same in the hypothesis space. The same task can be solved using the dimensionality reduction technique of PCA with must-link constraints between similar instances (OPCA), where the targets and hypothesis are embedded in the same space and we try to find the closest hypothesis. Does this new method achieve smaller RMSE?

For carrying out the experiments, we use one dataset containing protein similarities, and we use 9 datasets regarding the protein-compound information extracted from *BindingDB* (`bindingdb.org`). Each of these 9 datasets (one for each protein) corresponds to a human protein with a peptidase domain and comprises between 240 and 2649 compounds. For the representation of the compounds,

we utilise the standard molecular fingerprint ECFP4 [Pan and Yang, 2010]. We associate a feature vector with the similarities calculated from the amino acid sequence similarity of the peptidase domain for each compound and, at the end of each line, the affinity label regarding the respective protein.

The experimental process has three main phases: the dataset creation, the training of the models, and the evaluation of the algorithms. In the first phase, we create 10 distinct draws of the dataset through a process of combining compounds, randomly choosing 240 for each protein. At this point, we have new combinations of compounds for each protein ready to be incorporated into the models. In the second phase, for each draw, we create a gram matrix using the linear kernel on the compound matrix, which represents the junction of the compounds of all proteins. The result of the linear kernel is the gram matrix, which considers the similarities between all compounds at the level of features; in this case, the number of features they have in common. These considered features are the amino acid sequence similarity of the peptidase domain. From the gram matrix, we can reduce the similarities between the compounds only to the selected protein and generate a sub-matrix. Using this new matrix and the compounds' affinities of the selected protein, we establish a coefficient to each compound by training the models through the *Support Vector Regression* algorithm. 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 regularization parameter $C \in \{2^{-i} : i \in \{5, 4, ..., 4, 5\}\}$. In the end of this phase, we have the models associated with each protein. In the third and last phase, we evaluate the algorithms that can solve the orphan screening problem. Therefore, we perform a leave-one-out cross-validation over all proteins, considering each protein as an orphan once. This method is the same as the one applied in the paper from Giesselbach et al. [2018], on which this work is based. As a comparison measure, we take into account the Root-mean-squared error (RMSE) of the predicted affinities, as well as the execution time of each algorithm. The reported RMSE is an average over the 10 draws per orphan protein.

Bearing in mind that OPCA is the focus of this paper, it makes sense to refer to the most significant differences between this and the other algorithms in practical terms. As mentioned earlier, this algorithm has an instance space different from all other approaches, where it embed targets and hypotheses in the same space and requires supervised targets to be close to the corresponding supervised hypothesis. Thus, the kernel gram matrix present in this algorithm includes targets and hypotheses, and both are treated as instances of the same type for calculating the optimal projection. In the end, the orphan target is compared with the hypotheses, choosing the one that minimizes the error. The search hypothesis space used in equation 10 was selected as the supervised hypotheses of the supervised proteins.

Due to some programming complications, the experiments we present do not use the optimization function 9, but an approximation indicated as:

$$\max_{\Pi \in \mathbb{R}^{D \times d}} \text{tr}\Big(\frac{1}{D}\Pi^T K H_D K \Pi\Big) - v\text{tr}\Big(\frac{1}{\mathcal{C}}\Pi^T K L K \Pi\Big) + \text{tr}\Big(Q \times \big(\Pi^T K \Pi - \mathbf{I}_d\big)\Big),$$

where Q is a large constant, in this case with the value 100, and v is the trade-off hyperparameter that was established to be 1. We also used the gradient of the optimization above in the solver and, in this way, the procedure can work adequately to find a local maximum of the differentiable function. The gradient ∇f, the derivative of the objective function, is defined as:

$$\nabla f = \frac{\partial f}{\partial \Pi} = \frac{1}{D} K H_D K \Pi + \frac{1}{D}(KHK)^T \Pi - (\frac{v}{C} K L K \Pi + \frac{v}{C}(KLK)^T \Pi) + K \Pi Q + K^T \Pi Q^T.$$

This optimization was performed through the *scipy* library with the solver method *BFGS*, which uses the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno. We use 1000 as the maximum number of iterations performed. To obtain the absolute error of the OPCA algorithm, we search for the best hypothesis in the equation 10, then find the protein associated with this hypothesis, produce the predictions and perform the RMSE against the actual values of the orphan protein.

Figure 2 shows the RMSE of all approaches averaged over all draws for the protein P00750 and P07858. Our main comparison algorithms will be the NLCP, as this is the one that is presented as the best in state-of-art. In the first image, concerning protein P00750, it is possible to visualize a case in

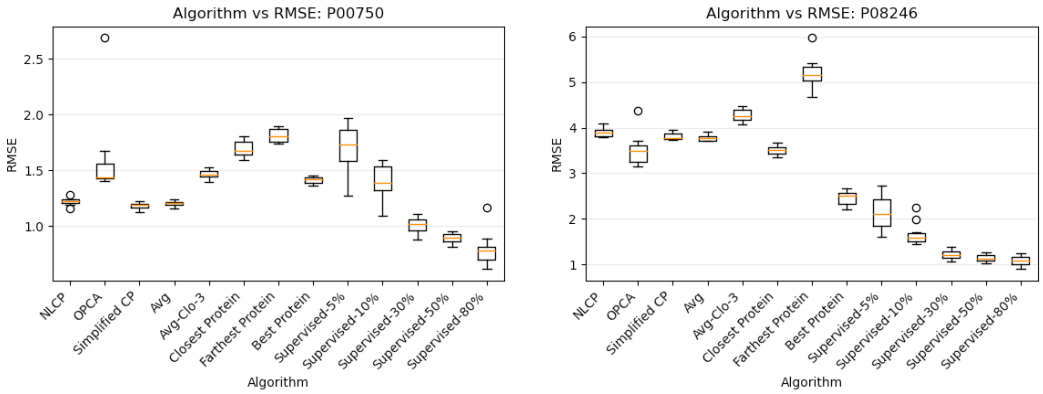

Figure 2: RMSE of the approaches over all draws for two proteins: P00750 and P08246.

which the OPCA has an error more significant than the NLCP. However, in the second image, relative to protein P08246, OPCA already presents a minor error than NLCP, showing itself, therefore, to be more effective. These 2 figures show that the performance varies strongly depending on the orphan target. Therefore, as an OPCA comparison term for all tested algorithms, we will use the results in figure 3, where the RMSE of all approaches averaged over all orphan proteins and all draws are shown.

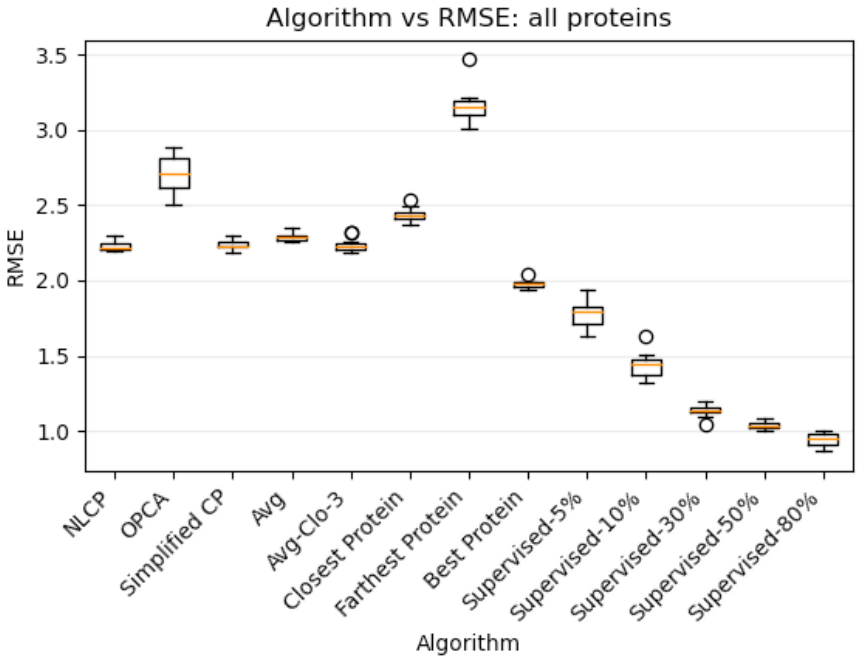

Figure 3: RMSEs of the proposed OPCA approach for all 9 proteins and 10 draws in comparison to the baselines.

Comparing and explaining the algorithms present, a direct way to combine target models without considering similarities of targets is to average them and use the average model (Avg) to predict the orphan target's composite affinities. Avg has a median RMSE of 2.285. Reducing Avg to a model using only the 3 closest targets (Avg-Clo-3) produces an increase in performance, showing that it benefits from a focus on the closest proteins, with a median RMSE of 2.231.

If we use protein similarities, the most direct way is to choose models from other targets according to their similarity to the orphan target and understand the influence of the proximity between proteins in the prediction. For this purpose, we evaluated the performance of the models of the closest and

farthest targets to the orphan. The closest has a median RMSE of 2.428, and the farthest has a median RMSE of 3.155. The fact that the closest protein performs better confirms that the orphan and the closest targets share similar features that determine the affinities of the compounds.

The Best-Protein calculates the orphan's RMSE for all proteins and selects the one with the lowest value: the protein with the most similarity. This process has a median RMSE of 1.977.

The NLCP optimizes the weights of each model to match the similarities of the targets, making use of all possible models. The results suggest that including the similarities in the target space and hypothesis is profitable, presenting a median RMSE of 2.214. The simplified CP is the other variable in Corresponding Projections and also has a good result with a median RMSE of 2,229. This algorithm is not as good as the NLCP, which is expected not to include as much complexity.

We also evaluate a hypothetical supervised case in which we trained an SVR in various fractions (5%, 10%, 30%, 50% and 80%) of the available compounds, testing it on the others. On average, all supervised models outperform the rest of the algorithms. However, they solve a different task since the other algorithms do not assume any labelled data for the orphan target.

Table 1 presents the execution times associated with each algorithm, making available the RMSEs already discussed. Regarding the execution times, it is possible to confirm that the simplicity of the algorithms naturally brings a faster execution, unlike those that require more complex processes, as is the case of NLCP, OPCA and Best-Protein. The worst cases are related to the supervised models since they require training of the algorithm. Naturally, the process appears to be more time-consuming in cases with more supervised data present, noting an increasing trend related to the increase in data used for training.

Algorithm	**RMSE**	**Execution Time**
NLCP	2.214	22.06 ms
OPCA	2.711	847.74 ms
Simplified CP	2.229	3.59 ms
Avg	2.285	3.55 ms
Avg-Clo-3	2.231	3.52 ms
Closest Protein	2.428	6.08 ms
Farthest Protein	3.155	5.15 ms
Best Protein	1.977	28.63 ms
Supervised-5%	1.797	392.37 ms
Supervised-10%	1.440	393.69 ms
Supervised-30%	1.140	845.40 ms
Supervised-50%	1.037	2395.78 ms
Supervised-80%	0.946	6324.41 ms

Table 1: Algorithm vs Median RMSE vs Mean Execution Time: Cross-validation over all proteins

5 Conclusion

This report describes a new algorithm to solve the orphan tracking problem, OPCA. In this algorithm, we embed targets and hypotheses in the same space, which is new compared with the other approaches. We try to find the closest hypothesis after applying the dimensionality reduction technique of PCA with optimized must-link constraints projections between similar instances. The results presented consider a hypothesis space in equation 10 as being the supervised hypothesis of the supervised proteins. Thus, the algorithm returns a median close to the Closest Protein, as we are finding the closest hypothesis of the supervised hypothesis space. As future work, this search space can be

expanded to random hypothesis and linear combinations, with results expected to improve as there are more variants for the optimization to influence. Regarding this optimization, it is essential to mention that this is an approximation. The final median returned makes sense, but the fact that it is not the exact value of the complex optimization may explain some outliers that exist in the execution of this process. The instability of the OPCA results can also be explained by the unusual combination of targets and hypotheses in the same instance space, performing calculations between them when they do not have the same origin.

Regarding the execution time, the OPCA, in fact, takes more time in comparison to the other approaches. This delay is explained by the process of maximizing the optimal projection. Using only the supervised hypothesis in the search space, this optimization process is not as important if we also consider the time spent, but it is expected to be more efficient and effective for more expansive search spaces. The final conclusion is that, although not all the capabilities this approach can provide are explored, a baseline has already been carried out, and the algorithm meets the expectations in the theoretical foundation. Referring to the research question, the new model did not achieve better RMSE considering all proteins. However, it is important to refer that OPCA's hypothesis space does not consider combinations of proteins, unlike NLCP, which performs these non-linear combinations matching the similarities of proteins and using all models.

References

Sven Giesselbach, Katrin Ullrich, Michael Kamp, Daniel Paurat, and Thomas Gärtner. Corresponding projections for orphan screening, 2018.

Dino Oglic, Daniel Paurat, and Thomas Gärtner. Interactive knowledge-based kernel pca. volume 8725, pages 501–516, 2014.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.