Co-Regularised Support Vector Regression

Katrin Ullrich[1], Michael Kamp[1,2], Thomas Gärtner[3], Martin Vogt[1,4], and Stefan Wrobel[1,2]

[1] University of Bonn, Germany
ullrich@iai.uni-bonn.de
[2] Fraunhofer IAIS, Sankt Augustin, Germany
michael.kamp@iais.fraunhofer.de,
stefan.wrobel@iais.fraunhofer.de
[3] University of Nottingham, UK
thomas.gaertner@nottingham.ac.uk
[4] B-IT, LIMES Program Unit, Bonn, Germany
martin.vogt@bit.uni-bonn.de

Abstract. We consider a semi-supervised learning scenario for regression, where only few labelled examples, many unlabelled instances and different data representations (multiple views) are available. For this setting, we extend support vector regression with a co-regularisation term and obtain co-regularised support vector regression (CoSVR). In addition to labelled data, co-regularisation includes information from unlabelled examples by ensuring that models trained on different views make similar predictions. Ligand affinity prediction is an important real-world problem that fits into this scenario. The characterisation of the strength of protein-ligand bonds is a crucial step in the process of drug discovery and design. We introduce variants of the base CoSVR algorithm and discuss their theoretical and computational properties. For the CoSVR function class we provide a theoretical bound on the Rademacher complexity. Finally, we demonstrate the usefulness of CoSVR for the affinity prediction task and evaluate its performance empirically on different protein-ligand datasets. We show that CoSVR outperforms co-regularised least squares regression as well as existing state-of-the-art approaches for affinity prediction.

Keywords: regression, kernel methods, semi-supervised learning, multiple views, co-regularisation, Rademacher complexity, ligand affinity prediction

1 Introduction

We investigate an algorithm from the intersection field of semi-supervised and multi-view learning. In semi-supervised learning the lack of a satisfactory number of labelled examples is compensated by the usage of many unlabelled instances from the respective feature space. Multi-view regression algorithms utilise different data representations to train models for a real-valued quantity. Ligand affinity prediction is an important learning task from chemoinformatics since many drugs act as protein ligands. It can be assigned to this learning scenario in a very natural way. The aim of affinity prediction is the determination of binding affinities for small molecular compounds—the ligands—with respect to a bigger protein using computational methods. Besides a

few labelled protein-ligand pairs, millions of small compounds are gathered in molecular databases as ligand candidates. Many different data representations—the so-called molecular fingerprints or views—exist that can be used for learning. Affinity prediction and other applications suffer from little label information and the need to choose the most appropriate view for learning. To overcome these difficulties, we propose to apply an approach called co-regularised support vector regression. We are the first to investigate support vector regression with co-regularisation, i.e., a term penalising the deviation of predictions on unlabelled instances. We investigate two loss functions for the co-regularisation. In addition to variants of our multi-view algorithm with a reduced number of optimisation variables, we also derive a transformation into a single-view method. Furthermore, we prove upper bounds for the Rademacher complexity, which is important to restrict the capacity of the considered function class to fit random data. We will show that our proposed algorithm outperforms affinity prediction baselines.

The strength of a protein-compound binding interaction is characterised by the real-valued *binding affinity*. If it exceeds a certain limit, the small compound is called a *ligand* of the protein. Ligand-based classification models can be trained to distinguish between ligands and non-ligands of the considered protein (e.g., with support vector machines [6]). Since framing the biological reality in a classification setting represents a severe simplification of the biological reality, we want to predict the strength of binding using regression techniques from machine learning. Both classification and regression methods are also known under the name of *ligand-based virtual screening*. (In the context of regression, we will use the name *ligands* for all considered compounds.) Various approaches like *neural networks* [7] have been applied. However, *support vector regression* (SVR) is the state-of-the-art method for affinity prediction studies (e.g., [12]).

As mentioned above, in the context of affinity prediction one is typically faced with the following practical scenario: for a given protein, only few ligands with experimentally identified affinity values are available. In contrast, the number of synthesizable compounds gathered in molecular databases (such as ZINC, BindingDB, ChEMBL[5]) is huge which can be used as unlabelled instances for learning. Furthermore, different free or commercial vectorial representations or *molecular fingerprints* for compounds exist. Originally, each fingerprint was designed towards a certain learning purpose and, therefore, comprises a characteristic collection of physico-chemical or structural molecular features [1], for example, predefined key properties (Maccs fingerprint) or listed subgraph patterns (ECFP fingerprints).

The canonical way to deal with multiple fingerprints for virtual screening would be to extensively test and compare different fingerprints [6] or perform time-consuming preprocessing feature selection and recombination steps [8]. Other attempts to utilise multiple views for one prediction task can be found in the literature. For example, Ullrich et al. [13] apply multiple kernel learning. However, none of these approaches include unlabelled compounds in the affinity prediction task. The semi-supervised *co-regularised least squares regression* (CoRLSR) algorithm of Brefeld et al. [4] has been shown to outperform single-view *regularised least squares regression* (RLSR) for UCI

[5] zinc.docking.org, www.bindingdb.org, www.ebi.ac.uk/chembl

datasets[6]. Usually, SVR shows very good predictive results having a lower generalisation error compared to RLSR. Aside from that, SVR represents the state-of-the-art in affinity prediction (see above). For this reason, we define *co-regularised support vector regression* (CoSVR) as an ε-insensitive version of co-regularisation. In general, CoSVR—just like CoRLSR—can be applied on every regression task with multiple views on data as well as labelled and unlabelled examples. However, learning scenarios with high-dimensional sparse data representations and very few labelled examples—like the one for affinity prediction—could benefit from approaches using co-regularisation. In this case, unlabelled examples can contain information that could not be extracted from a few labelled examples because of the high dimension and sparsity of the data representation.

A view on data is a representation of its objects, e.g., with a particular choice of features in $\mathbb{R}^d$. We will see that feature mappings are closely related to the concept of *kernel functions*, for which reason we introduce CoSVR theoretically in the general framework of kernel methods. Within the research field of *multi-view learning*, CoSVR and CoRLSR can be assigned to the group of co-training style [16] approaches that simultaneously learn multiple predictors, each related to a view. Co-training style approaches enforce similar outcomes of multiple predictor functions for unlabelled examples, measured with respect to some loss function. In the case of co-regularisation for regression the empirical risks of multiple predictors (*labelled error*) plus an error term for unlabelled examples (*unlabelled error, co-regularisation*) are minimised.

The idea for mutual influence of multiple predictors appeared in the paper of Blum and Mitchell [2] on classification with co-training. Wang et al. [14] combined the technique of co-training with SVR with a technique different from co-regularisation. Analogous to CoSVR, CoRLSR is a semi-supervised and multi-view version of RLSR that requires the solution of a large system of equations [4]. A co-regularised version for support vector machine classification SVM-2K already appeared in the paper of Farquhar et al. [5], where the authors define a co-regularisation term via the ε-insensitive loss on labelled examples. It was shown by Sindhwani and Rosenberg [11] that co-regularised approaches applying the squared loss function for the unlabelled error can be transformed into a standard SVR optimisation with a particular fusion kernel. A bound on the empirical Rademacher complexity for co-regularised algorithms with Lipschitz continuous loss function for the labelled error and squared loss function for the unlabelled error was proven by Rosenberg and Bartlett [9].

A preliminary version of this paper was published at the *Data Mining in Biomedical Informatics and Healthcare* workshop held at ICDM 2016. There, we considered only the CoSVR special case ε-CoSVR and its variants with reduced numbers of variables (for the definitions consult Defs. 1 - 3 below) focusing the application of ligand affinity prediction. The ℓ_2-CoSVR case (see below) with its theoretical properties (Lemmas $1(ii)$ - $3(ii)$, $6(ii)$) and practical evaluation, as well as the faster Σ-CoSVR (Sect. 3.3) variant are novel contributions in the present paper.

In the following section, we will present a short summary of kernels and multiple views, as well as important notation. We define CoSVR and variants of the base algorithm in Sect. 3. In particular, a Rademacher bound for CoSVR will be proven in

[6] UCI machine learning repository, http://archive.ics.uci.edu/ml

Sect. 3.5. Subsequently, we provide a practical evaluation of CoSVR for ligand affinity prediction in Sect. 4 and conclude with a brief discussion in Sect. 5.

2 Kernels and Multiple Views

We consider an arbitrary instance space $\mathcal{X}$ and the real numbers as label space $\mathcal{Y}$. We want to learn a function f that predicts a real-valued characteristic of the elements of $\mathcal{X}$. Suppose for training purposes we have sets $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ of labelled and $Z = \{z_1, \ldots, z_m\} \subset \mathcal{X}$ of unlabelled instances at our disposal, where typically $m \gg n$ holds true. With $\{y_1, \ldots, y_n\} \subset \mathcal{Y}$ we denote the respective labels of X. Furthermore, assume the data instances can be represented in M different ways. More formally, for $v \in \{1, \ldots, M\}$ there are functions $\Phi_v : \mathcal{X} \to \mathcal{H}_v$, where $\mathcal{H}_v$ is an appropriate inner product space. Given an instance $x \in \mathcal{X}$, we say that $\Phi_v(x)$ is the *v-th view* of x. If $\mathcal{H}_v$ equals $\mathbb{R}^d$ for some finite dimension d, the intuitive names *(v-th) feature mapping* and *feature space* are used for Φ_v and $\mathcal{H}_v$, respectively. If in the more general case $\mathcal{H}_v$ is a *Hilbert space*, d can even be infinite (see below). For view v the predictor function $f_v : \mathcal{X} \to \mathbb{R}$ is denoted with *(single) view predictor*. View predictors can be learned independently for each view utilising an appropriate regression algorithm like SVR or RLSR. As a special case we consider *concatenated predictors* f_v in Sect. 4 where the corresponding view v results from a concatenation of finite dimensional feature representations $\Phi_1, \ldots, \Phi_M$. Having different views on the data, an alternative is to learn M predictors $f_v : \mathcal{X} \to \mathbb{R}$ simultaneously that depend on each other, satisfying an optimisation criterion involving all views at once. Such a criterion could be the minimisation of the labelled error in line with co-regularisation which will be specified in the following subsection. The final predictor f will then be the average of the predictors f_v.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *kernel* if it is symmetric and positive semi-definite. Indeed, for every kernel k there is a feature mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) and $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_{\mathcal{H}}$ holds true for all $x_1, x_2 \in \mathcal{X}$ (*Mercer's theorem*). Thus, the function k is the corresponding reproducing kernel of $\mathcal{H}$, and for $x \in \mathcal{X}$ the mappings $\langle \Phi(x), \Phi(\cdot) \rangle = k(x, \cdot)$ are functions defined on $\mathcal{X}$. Choosing RKHSs $\mathcal{H}_v$ of multiple kernels k_v as candidate spaces for the predictors f_v, the *representer theorem* of Schölkopf et al. [10] allows for a parameterisation of the optimisation problems for co-regularisation presented below. A straightforward modification of the representer theorem's proof leads to a representation of the predictors f_v as finite kernel expansion

$$f_v(\cdot) = \sum_{i=1}^{n} \pi_{vi} k_v(x_i, \cdot) + \sum_{j=1}^{m} \pi_{v(j+n)} k_v(z_j, \cdot) \tag{1}$$

with linear coefficients $\pi_v \in \mathbb{R}^{n+m}$, centered at labelled and unlabelled instances $x_i \in X$ and $z_j \in Z$, respectively.

The kernel matrices $K_v = \{k_v(x_i, x_j)\}_{i,j=1}^{n+m}$ are the *Gram matrices* of the v-th view kernel k_v over labelled and unlabelled examples and have decompositions into an upper and a lower part $L_v \in \mathbb{R}^{n \times (n+m)}$ and $U_v \in \mathbb{R}^{m \times (n+m)}$, respectively. We

will consider the submatrices $k(Z,x) := (k(z_1,x),\ldots,k(z_m,x))^T$ and $k(Z,Z) := \{k(z_j,z_{j'})\}_{j,j'=1}^m$ of a Gram matrix with kernel k. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are RKHSs then their sum space $\mathcal{H}_\Sigma$ is defined as $\mathcal{H}_\Sigma := \{f : f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$. With $Y = (y_1,\ldots,y_n)^T \in \mathbb{R}^n$ we denote the vector of labels. We will abbreviate $v \in \{1,\ldots,M\}$ with $v \in [\![M]\!]$. And finally, we will utilise the squared loss $\ell_2(y,y') = \|y-y'\|^2$ and the ε-insensitive loss $\ell_\varepsilon(y,y') = \max\{0,|y-y'|-\varepsilon\}$, $y,y' \in \mathcal{Y}$.

3 The CoSVR Algorithm: Variants and Properties

3.1 Base CoSVR

In order to solve a regression task in the presence of multiple views $v = 1,\ldots,M$, the approach of *co-regularisation* is to jointly minimise two error terms involving M predictor functions $f_1,\ldots,f_M$. Firstly, every view predictor f_v is intended to have a small training error with respect to labelled examples. Secondly, the difference between pairwise view predictions over unlabelled examples should preferably be small. We introduce *co-regularised support vector regression* (CoSVR) as an ε-insensitive loss realisation of the co-regularisation principle.

Definition 1. *For $v \in \{1,\ldots,M\}$ let $\mathcal{H}_v$ be RKHSs. The co-regularised empirical risk minimisation*

$$\min_{f_v \in \mathcal{H}_v} \sum_{v=1}^{M} \left(\frac{\nu_v}{2}\|f_v\|^2 + \sum_{i=1}^{n} \ell^L(y_i, f_v(x_i)) \right) + \lambda \sum_{u,v=1}^{M} \sum_{j=1}^{m} \ell^U(f_u(z_j), f_v(z_j)), \tag{2}$$

where $\nu_v, \lambda \geq 0$ is called co-regularised support vector regression (CoSVR) if $\ell^L = \ell_{\varepsilon^L}$, $\varepsilon^L \geq 0$, and ℓ^U is an arbitrary loss function for regression. Furthermore, we define ε-CoSVR to be the special case where $\ell^U = \ell_{\varepsilon^U}$, $\varepsilon^U \geq 0$, as well as ℓ_2-CoSVR to satisfy $\ell^U = \ell_2$.

The minimum in (2) is taken over all f_v, $v = 1,\ldots,M$. For reasons of simplification we will abbreviate $\min_{f_1 \in \mathcal{H}_1,\ldots,f_M \in \mathcal{H}_M}$ with $\min_{f_v \in \mathcal{H}_v}$. Note that the loss function parameters ε^L and ε^U can have different values. The parameters ν_v and λ are trade-off parameters between empirical risk and co-regularisation term. The added norm terms $\|f_v\|$ prevent overfitting. We will also refer to the empirical risk term with loss function ℓ^L as *labelled error* and to the co-regularisation term with ℓ^U as *unlabelled error*. In the case of $\ell^L = \ell^U = \ell_2$, the optimisation in (2) is known as *co-regularised least squares regression* (CoRLSR). Brefeld et al. [4] found a closed form solution for CoRLSR as linear system of equations in $M(n+m)$ variables. In the following, we present a solution for ε-CoSVR and ℓ_2-CoSVR.

Lemma 1. *Let $\nu_v, \lambda, \varepsilon^L, \varepsilon^U \geq 0$. We use the notation introduced above. In particular, $\pi_v \in \mathbb{R}^{n+m}$ denote the kernel expansion coefficients of the single view predictors f_v*

from (1), whereas $\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n$ and $\gamma_{uv} \in \mathbb{R}^m$ are dual variables.
(i) The dual optimisation problem of ε-CoSVR equals

$$\max_{\alpha_v,\hat{\alpha}_v \in \mathbb{R}^n, \gamma_{uv} \in \mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha_v - \hat{\alpha}_v)^T Y \right.$$
$$\left. -(\alpha_v + \hat{\alpha}_v)^T \varepsilon^L \mathbf{1}_n - \sum_{u=1}^{M} \gamma_{uv}^T \varepsilon^U \mathbf{1}_m \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha_v, \hat{\alpha}_v \leq \mathbf{1}_n \\ \mathbf{0}_m \leq \gamma_{uv} \leq \lambda \mathbf{1}_m \end{array} \right\}_{v \in [\![M]\!],\ (u,v) \in [\![M]\!]^2},$$

where $\pi_v^T = \frac{1}{\nu_v}(\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\alpha_v - \hat{\alpha}_v \mid \sum_{u=1}^{M}(\gamma_{uv} - \gamma_{vu}))^T$.
(ii) The dual optimisation problem of ℓ_2-CoSVR is

$$\max_{\alpha_v,\hat{\alpha}_v \in \mathbb{R}^n, \gamma_{uv} \in \mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha_v - \hat{\alpha}_v)^T Y \right.$$
$$\left. -\varepsilon^L (\alpha_v + \hat{\alpha}_v)^T \mathbf{1}_n - \frac{1}{4\lambda} \sum_{u=1}^{M} \gamma_{uv}^T \gamma_{uv} \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha_v, \hat{\alpha}_v \leq \mathbf{1}_n \\ \gamma_{uv} = \frac{2\lambda}{\nu_u} U_u \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_u - \frac{2\lambda}{\nu_v} U_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v \end{array} \right\}_{v \in [\![M]\!],\ (u,v) \in [\![M]\!]^2},$$

where $\pi_v^T = \frac{1}{\nu_v}(\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\alpha_v - \hat{\alpha}_v \mid \sum_{u=1}^{M}(\gamma_{uv} - \gamma_{vu}))^T$.

Remark 1. The proofs of Lemma 1 as well as Lemmas 2 and 3 below use standard techniques from *Lagrangian dualisation* (e.g., [3]). They can be found in our CoSVR repository (see footnote 7).

We choose the concatenated vector representation $(\alpha \mid \gamma)_v^T \in \mathbb{R}^{n+m}$ in order to show the correspondence between the two problems ε-CoSVR and ℓ_2-CoSVR and further CoSVR variants below. Additionally, the similarities with and differences to the original SVR dual problem are obvious. We will refer to the optimisation in Lemma 1 as the base CoSVR algorithms.

3.2 Reduction of Variable Numbers

The dual problems in Lemma 1 are quadratic programs. Both depend on $2Mn + M^2m$ variables, where $m \gg n$. If the number of views M and the number of unlabelled examples m are large, the base CoSVR algorithm might cause problems with respect to runtime because of the large number of resulting variables. In order to reduce this number, we define modified versions of base CoSVR. We denote the variant with a modification in the labelled error with CoSVR^{mod} and in the unlabelled error with CoSVR$_{mod}$.

Modification of the Empirical Risk In base CoSVR the empirical risk is meant to be small for each single view predictor individually using examples and their corresponding labels. In the CoSVR^{mod} variant the average prediction, i.e., the final predictor, is applied to define the labelled error term.

Definition 2. *The co-regularised support vector regression problem with modified constraints for the labelled examples (CoSVR^{mod}) is defined as*

$$\min_{f_v \in \mathcal{H}_v} \sum_{v=1}^{M} \frac{\nu_v}{2} \|f_v\|^2 + \sum_{i=1}^{n} \ell_{\varepsilon^L}(y_i, f^{\text{avg}}(x_i)) + \lambda \sum_{u,v=1}^{M} \sum_{j=1}^{m} \ell^U \left(f_u(z_j), f_v(z_j)\right),$$

where $f^{\text{avg}} := \frac{1}{M}\sum_{v=1}^{M} f_v$ is the average of all single view predictors. We denote the case $\ell^U = \ell_{\varepsilon^U}$, $\varepsilon^U \geq 0$, with ε-CoSVR^{mod} and the case $\ell^U = \ell_2$ with ℓ_2-CoSVR^{mod}.

In the following lemma we present solutions for ε-CoSVR^{mod} and ℓ_2-CoSVR^{mod}.

Lemma 2. *Let $\nu_v, \lambda, \varepsilon^L, \varepsilon^U \geq 0$. We utilise dual variables $\alpha, \hat{\alpha} \in \mathbb{R}^n$ and $\gamma_{uv} \in \mathbb{R}^m$.*
(i) The ε-CoSVR^{mod} dual optimisation problem can be written as

$$\max_{\alpha,\hat{\alpha}\in\mathbb{R}^n,\ \gamma_{uv}\in\mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha - \hat{\alpha})^T Y - (\alpha + \hat{\alpha})^T \varepsilon^L \mathbf{1}_n - \sum_{u=1}^{M} \gamma_{uv}^T \varepsilon^U \mathbf{1}_m \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha, \hat{\alpha} \leq \mathbf{1}_n \\ \mathbf{0}_m \leq \gamma_{uv} \leq \lambda \mathbf{1}_m \end{array} \right\}_{v \in [\![M]\!]},$$

where $\pi_v^T = \frac{1}{\nu_v}(\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\frac{1}{M}(\alpha - \hat{\alpha}) \mid \sum_{u=1}^{M}(\gamma_{uv} - \gamma_{vu}))^T$.
(ii) The ℓ_2-CoSVR^{mod} dual optimisation problem equals

$$\max_{\alpha,\hat{\alpha}\in\mathbb{R}^n,\ \gamma_{uv}\in\mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha_v - \hat{\alpha}_v)^T Y - (\alpha_v + \hat{\alpha}_v)^T \varepsilon^L \mathbf{1}_n - \frac{1}{4\lambda} \sum_{u=1}^{M} \gamma_{uv}^T \gamma_{uv} \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha, \hat{\alpha} \leq \mathbf{1}_n \\ \gamma_{uv} = \frac{2\lambda}{\nu_u} U_u \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_u - \frac{2\lambda}{\nu_v} U_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v \end{array} \right\}_{v \in [\![M]\!]},$$

where $\pi_v^T = \frac{1}{\nu_v}(\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\frac{1}{M}(\alpha - \hat{\alpha}) \mid \sum_{u=1}^{M}(\gamma_{uv} - \gamma_{vu}))^T$.

We can also reduce the number of variables more effectively using modified constraints for the co-regularisation term. Whereas the CoSVR^{mod} algorithm is rather important from a theoretical perspective (see Sect. 3.3), the variant presented in the next section is very beneficial from a practical perspective if the number of views M is large.

Modification of the Co-Regularisation The unlabelled error term of base CoSVR bounds the pairwise distances of view predictions, whereas now in CoSVR$_{mod}$ only the disagreement between predictions of each view and the average prediction of the residual views will be taken into account.

Definition 3. *We consider RKHSs $\mathcal{H}_1, \ldots, \mathcal{H}_M$ as well as constants $\varepsilon^L, \varepsilon^U, \nu_v, \lambda \geq 0$. The co-regularised support vector regression problem with modified constraints for the unlabelled examples (CoSVR$_{mod}$) is defined as*

$$\min_{f_v \in \mathcal{H}_v} \sum_{v=1}^{M} \left(\frac{\nu_v}{2} \|f_v\|^2 + \sum_{i=1}^{n} \ell_{\varepsilon^L}(y_i, f_v(x_i)) \right) + \lambda \sum_{v=1}^{M} \sum_{j=1}^{m} \ell^U \left(f_v^{\text{avg}}(z_j),\ f_v(z_j) \right), \tag{3}$$

where now $f_v^{\text{avg}} := \frac{1}{M-1} \sum_{u=1}^{M,u\neq v} f_u$ is the average of view predictors besides view v. We denote the case $\ell^U = \ell_{\varepsilon^U}$, $\varepsilon^U \geq 0$, with ε-CoSVR$_{mod}$ and the case $\ell^U = \ell_2$ with ℓ_2-CoSVR$_{mod}$.

Again we present solutions for ε-CoSVR$_{mod}$ and ℓ_2-CoSVR$_{mod}$.

Lemma 3. *Let $\nu_v, \lambda, \varepsilon^L, \varepsilon^U \geq 0$. We utilise dual variables $\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n$ and $\gamma_v, \hat{\gamma}_v \in \mathbb{R}^m$, as well as $\gamma_v^{avg} := \frac{1}{M-1} \sum_{u=1}^{M,u\neq v} \gamma_u$ and $\hat{\gamma}_v^{avg} := \frac{1}{M-1} \sum_{u=1}^{M,u\neq v} \hat{\gamma}_u$ analogous to the residual view predictor average.*
(i) The ε-CoSVR$_{mod}$ dual optimisation problem can be written as

$$\max_{\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n,\ \gamma_v, \hat{\gamma}_v \in \mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha - \hat{\alpha})^T Y - (\alpha_v + \hat{\alpha}_v)^T \varepsilon^L \mathbf{1}_n - (\gamma_v + \hat{\gamma}_v) \varepsilon^U \mathbf{1}_m \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha_v, \hat{\alpha}_v \leq \mathbf{1}_n \\ \mathbf{0}_m \leq \gamma_v, \hat{\gamma}_v \leq \lambda \mathbf{1}_m \end{array} \right\}_{v \in [\![M]\!]},$$

where $\pi_v^T = \frac{1}{\nu_v} (\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\alpha_v - \hat{\alpha}_v \mid (\gamma_v - \gamma_v^{avg}) - (\hat{\gamma}_v - \hat{\gamma}_v^{avg}))^T$.
(ii) The ℓ_2-CoSVR$_{mod}$ dual optimisation problem equals

$$\max_{\alpha_v, \hat{\alpha}_v \in \mathbb{R}^n,\ \gamma_v \in \mathbb{R}^m} \sum_{v=1}^{M} \left(-\frac{1}{2\nu_v} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v^T K_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v + (\alpha_v - \hat{\alpha}_v)^T Y - (\alpha_v + \hat{\alpha}_v)^T \varepsilon^L \mathbf{1}_n - \frac{1}{4\lambda} \sum_{u=1}^{M} \gamma_v^T \gamma_v \right)$$
$$\text{s.t.} \left\{ \begin{array}{l} \mathbf{0}_n \leq \alpha_v, \hat{\alpha}_v \leq \mathbf{1}_n \\ \gamma_v = \frac{1}{M-1} \sum_{u=1}^{M,u\neq v} \frac{2\lambda}{\nu_u} U_u \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_u - \frac{2\lambda}{\nu_v} U_v \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_v \end{array} \right\}_{v \in [\![M]\!]},$$

where $\pi_v^T = \frac{1}{\nu_v} (\alpha \mid \gamma)_v^T$ and $(\alpha \mid \gamma)_v^T = (\alpha_v - \hat{\alpha}_v \mid \gamma_v - \gamma_v^{avg})^T$.

Remark 2. If we combine the modifications in the labelled and unlabelled error term we canonically obtain the variants ε-CoSVR$^{mod}_{mod}$ and ℓ_2-CoSVR$^{mod}_{mod}$.

In the base CoSVR versions the semi-supervision is realised with proximity constraints on pairs of view predictions. We show in the following lemma that the constraints of the closeness of one view prediction to the average of the residual predictions implies a closeness of every pair of predictions.

Lemma 4. *Up to constants, the unlabelled error bound of CoSVR$_{mod}$ is also an upper bound of the unlabelled error of base CoSVR.*

Proof. We consider the settings of Lemma 1(i) and Lemma 3(i). For part (ii) the proof is equivalent with $\varepsilon^U = 0$. In the case of $M = 2$, modified and base algorithm fall together which shows the claim. Now let $M > 2$. Because of the definition of the ε-insensitive loss we know that $|f_v(z_j) - f_v^{\text{avg}}(z_j)| \leq \varepsilon^U + c_{vj}$, where $c_{vj} \geq 0$ is the actual loss value for fixed v and j. We denote $c_j := \max_{v\in\{1,\dots,M\}}\{c_{1j}, \dots, c_{Mj}\}$ and, hence, $|f_v(z_j) - f_v^{\text{avg}}(z_j)| \leq \varepsilon^U + c_j$ for all $v \in \{1, \dots, M\}$. Now we conclude for $j \in \{1, \dots, m\}$ and $(u, v) \in \{1, \dots, M\}^2$

$$\begin{aligned}
&|f_u(z_j) - f_v(z_j)| \\
&\leq |f_u(z_j) - f_u^{\text{avg}}(z_j)| + |f_u^{\text{avg}}(z_j) - f_v^{\text{avg}}(z_j)| + |f_v^{\text{avg}}(z_j) - f_v(z_j)| \\
&\leq \varepsilon^U + c_j + \tfrac{1}{M-1}|f_v(z_j) - f_u(z_j)| + \varepsilon^U + c_j,
\end{aligned}$$

and therefore, $|f_u(z_j) - f_v(z_j)| \leq \frac{2(M-1)}{M-2}(\varepsilon^U + c_j)$. As a consequence we deduce from $\sum_{v=1}^M \sum_{j=1}^m \ell_{\varepsilon^U}(f_v^{\text{avg}}(z_j), f_v(z_j)) \leq M \sum_{j=1}^m c_j =: B$ that also the labelled error of CoSVR can be bounded $\sum_{u,v=1}^M \sum_{j=1}^m \ell_{\tilde{\varepsilon}}(f_u(z_j), f_v(z_j)) \leq \tilde{B}$ for $\tilde{\varepsilon} = \frac{2(M-1)}{M-2}\varepsilon^U$ and $\tilde{B} = \frac{2M(M-1)}{(M-2)}B$, which finishes the proof. □

3.3 Σ-CoSVR

Sindhwani and Rosenberg [11] showed that under certain conditions co-regularisation approaches of two views exhibit a very useful property. If $\ell^U = \ell_2$ and the labelled loss is calculated utilising an arbitrary loss function for the average predictor f^{avg}, the resulting multi-view approach is equivalent with a single-view approach of a fused kernel. We use the notion from Sect. 2.

Definition 4. *Let $\lambda, \nu_1, \nu_2, \varepsilon^L \geq 0$ be parameters and the Gram submatrices $k(Z, x)$ and $k(Z, Z)$ be defined as in Sect. 2. We consider a merged kernel k_Σ from two view kernels k_1 and k_2*

$$k_\Sigma(x, x') := k^{\oplus}(x, x') - k^{\ominus}(Z, x)^T \left(\tfrac{1}{\lambda} I_m + k^{\oplus}(Z, Z)\right)^{-1} k^{\ominus}(Z, x'), \tag{4}$$

for $x, x' \in \mathcal{X}$, where $k^{\oplus} := \frac{1}{\nu_1}k_1 + \frac{1}{\nu_2}k_2$ and $k^{\ominus} := \frac{1}{\nu_1}k_1 - \frac{1}{\nu_2}k_2$. We denote the SVR optimisation

$$\operatorname*{argmin}_{f\in\mathcal{H}_\Sigma} \|f\|^2 + \sum_{i=1}^n \ell_{\varepsilon^L}(y_i, \tfrac{1}{2}f(x_i)), \tag{5}$$

Σ-co-regularised support vector regression (Σ-CoSVR), where $\mathcal{H}_\Sigma$ is the RKHS of k_Σ.

Please notice that for each pair (x, x') the value of $k_\Sigma(x, x')$ is calculated in (4) with k_1 and k_2 including not only x and x' but also unlabelled examples $z_1, \ldots, z_m$. Hence, the optimisation problem in (5) is a standard SVR with additional information about unlabelled examples incorporated in the RKHS $\mathcal{H}_\Sigma$.

Lemma 5. *The algorithms ℓ_2-CoSVR^{mod} and Σ-CoSVR are equivalent and $\mathcal{H}_\Sigma$ is the sum space $\mathcal{H}_\Sigma = \{f : \mathcal{X} \to \mathbb{R} \mid f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$.*

Proof. The proof is an application of Theorem 2.2. of Sindhwani and Rosenberg [11] for the loss function V being equal to the ε-insensitive loss with $\varepsilon = \varepsilon^L$, the parameter of the labelled error of ℓ_2-CoSVR^{mod}. □

As Σ-CoSVR can be solved as a standard SVR algorithm we obtained a much faster co-regularisation approach. The information of the two views and the unlabelled examples are included in the candidate space $\mathcal{H}_\Sigma$ and associated kernel k_Σ.

3.4 Complexity

The CoSVR variants and CoRLSR mainly differ in the number of applied loss functions and the strictness of constraints. This results in different numbers of variables and constraints in total, as well as potentially non-zero variables (referred to as *sparsity*, compare Table 1). All presented problems are convex QPs with positive semi-definite

Table 1: Number of variables, constraints, and potential non-zero variables for different CoSVR versions and CoRLSR. The respective CoSVR^{mod} variant is included by cancelling the $\{M\}$-factor.

algorithm	**variables**	**constraints**	**sparsity**
ε-CoSVR	$2\{M\}n + M^2m$	$4\{M\}n + 2M^2m$	$\{M\}n + \frac{1}{2}(M^2 - M)m$
ℓ_2-CoSVR	$2\{M\}n + M^2m$	$4\{M\}n + M^2m$	$\{M\}n + M^2m$
ε-CoSVR$_{mod}$	$2\{M\}n + 2Mm$	$4\{M\}n + 4Mm$	$\{M\}n + Mm$
ℓ_2-CoSVR$_{mod}$	$2\{M\}n + Mm$	$4\{M\}n + Mm$	$\{M\}n + Mm$
Σ-CoSVR	$2n$	$4n$	n
CoRLSR	$Mn + Mm$	0	$Mn + Mm$

matrices in the quadratic terms. As the number m of unlabelled instances in real-world problems is much greater than n, the runtime of a QP-solver is dominated by the respective second summand in the constraints column of Table 1. Because of the ε-insensitive loss the number of actual non-zero variables in the learned model will be even smaller for the CoSVR variants than the numbers reported in the sparsity column of Table 1. In particular, for the modified variants this will allow for a more efficient model storage compared to CoRLSR. Indeed, according to the *Karush-Kuhn-Tucker conditions* (e.g., [3]), only for active inequality constraints the corresponding dual γ-variables can be non-zero. In this sense the respective unlabelled $z_j \in Z$ are *unlabelled support vectors*. This consideration is also valid for the α-variables and support vectors $x_i \in X$ as we

use the ε-insensitive loss for the labelled error in all CoSVR versions. And finally, in the two-view case with $M = 2$ the modified version with respect to the unlabelled error term and the base version coincide.

3.5 A Rademacher Bound for CoSVR

Similarly to the result of Rosenberg and Bartlett [9] we want to prove a bound on the *empirical Rademacher complexity* $\hat{\mathcal{R}}_n$ of CoSVR in the case of $M = 2$. Note that, despite the proof holding for the special case of $M = 2$, the CoSVR method in general is applicable to arbitrary numbers of views. The empirical Rademacher complexity is a data-dependent measure for the capacity of a function class $\mathcal{H}$ to fit random data and is defined as

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}^\sigma \left[\sup_{f \in \mathcal{H}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| : \{x_1, \ldots, x_n\} = X \right].$$

The random data are represented via *Rademacher random variables* $\sigma = (\sigma_1, \ldots, \sigma_n)^T$. We consider ε-CoSVR and ℓ_2-CoSVR and define bounded versions $\mathcal{H}_\Sigma^\varepsilon$ and $\mathcal{H}_\Sigma^2$ of the sum space $\mathcal{H}_\Sigma$ from Sect. 2 for the corresponding versions. Obviously, a pair $(\pi_1, \pi_2) \in \mathbb{R}^{(n+m)\times(n+m)}$ of kernel expansion coefficients (see (1)) represents an element of $\mathcal{H}_\Sigma$. For ε-CoSVR and ℓ_2-CoSVR we set

$$\mathcal{H}_\Sigma^\varepsilon := \{(\pi_1, \pi_2) \in \mathcal{H}_\Sigma : -\mu 1_{n+m} \leq \pi_1, \pi_2 \leq \mu 1_{n+m}\}, \quad \text{and} \tag{6}$$

$$\mathcal{H}_\Sigma^2 := \{(\pi_1, \pi_2) \in \mathcal{H}_\Sigma : \nu_1 \pi_1^T K_1 \pi_1 + \nu_2 \pi_2^T K_2 \pi_2 + \lambda (U_1\pi_1 - U_2\pi_2)^T (U_1\pi_1 - U_2\pi_2) \leq 1\}, \tag{7}$$

respectively. In (6) μ is an appropriate constant according to Lemma 1 and 2. The definition in (7) follows the reasoning of Rosenberg and Bartlett [9]. Now we derive a bound on the empirical Rademacher complexity of $\mathcal{H}_\Sigma^\varepsilon$ and $\mathcal{H}_\Sigma^2$, respectively. We point out that the subsequent proof is also valid for the modified versions with respect to the empirical risk. For two views the base and modified versions with respect to the co-regularisation fall together anyway. For reasons of simplicity, in the following lemma and proof we omit ^{mod} and $_{mod}$ for the CoSVR variants. Furthermore, we will apply the infinity vector norm $\|v\|_\infty$ and row sum matrix norm $\|L\|_\infty$ (consult, e.g., Werner [15]).

Lemma 6. *Let $\mathcal{H}_\Sigma^\varepsilon$ and $\mathcal{H}_\Sigma^2$ be the function spaces in (6) and (7) and, without loss of generality, let $\mathcal{Y} = [-1, 1]$.*
(i) The empirical Rademacher complexity of ε-CoSVR can be bounded via

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^\varepsilon) \leq \frac{2s}{n} \mu (\|L_1\|_\infty + \|L_2\|_\infty),$$

where μ is a constant dependent on the regularisation parameters and s is the number of potentially non-zero variables in the kernel expansion vector $\pi \in \mathcal{H}_\Sigma^\varepsilon$.
(ii) The empirical Rademacher complexity of ℓ_2-CoSVR has a bound

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^2) \leq \frac{2}{n} \sqrt{tr_n(K_\Sigma)},$$

where $tr_n(K_\Sigma) := \sum_{i=1}^n k_\Sigma(x_i, x_i)$ *with the sum kernel* k_Σ *from (4).*

Our proof applies Theorem 2 and 3 of Rosenberg and Bartlett [9].

Proof. At first, using Theorem 2 of Rosenberg and Bartlett [9], we investigate the general usefulness of the empirical Rademacher complexity $\hat{\mathcal{R}}_n$ of $\mathcal{H}_\Sigma^{loss}$ in the CoSVR scenario. The function space $\mathcal{H}_\Sigma^{loss}$ can be either $\mathcal{H}_\Sigma^{\varepsilon}$ or $\mathcal{H}_\Sigma^{2}$. Theorem 2 requires two preconditions. First, we notice that the ε-insensitive loss function utilising the average predictor $\ell^L(y, f(x)) = \max\{0, |y-(f_1(x)+f_2(x))/2|-\varepsilon^L\}$ maps into $[0, 1]$ because of the boundedness of $\mathcal{Y}$. Second, it is easy to show that ℓ^L is *Lipschitz continuous*, i.e. $|\ell^L(y, y')-\ell^L(y, y'')|/|y'-y''| \leq C$, for some constant $C > 0$. With similar arguments one can show that the ε-insensitive loss function of base CoSVR is Lipschitz continuous as well. According to Theorem 2 of Rosenberg and Bartlett [9], the expected loss $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ \ell^L(f(X), Y)$ can then be bounded by means of the empirical risk and the empirical Rademacher complexity

$$\mathbb{E}_{\mathcal{D}}\ \ell^L(f(X), Y) \leq \frac{1}{n}\sum_{i=1}^{n} l^L(f(x_i), y_i) + 2C\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^{loss}) + \frac{2+3\sqrt{\ln(2/\delta)/2}}{\sqrt{n}}$$

for every $f \in \mathcal{H}_\Sigma^{loss}$ with probability at least $1-\delta$. Now we continue with the cases (i) and (ii) separately.

(i) We can reformulate the empirical Rademacher complexity

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^{\varepsilon}) = \frac{2}{n}\mathbb{E}^\sigma\left[\sup_{(\pi_1\ |\ \pi_2)^T\in\mathcal{K}} \left|\sigma^T(L_1\pi_1 + L_2\pi_2)\right|\right],$$

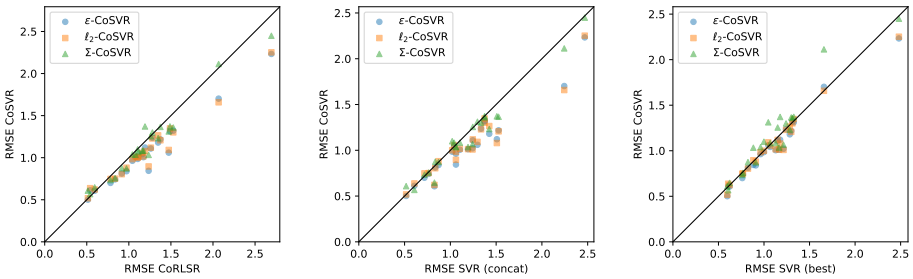
where $\mathcal{K} := \{(\pi_1\ |\ \pi_2)^T \in \mathbb{R}^{2(n+m)} : -\mu\mathbf{1}_{n+m} \leq \pi_1, \pi_2 \leq \mu\mathbf{1}_{n+m}\}$. The kernel expansion π of ε-CoSVR optimisation is bounded because of the box constraints in the respective dual problems. Therefore, π lies in the ℓ_1-ball of dimension s scaled with $s\mu$, i.e., $\pi \in s\mu \cdot B_1$. The dimension s is the sparsity of π, and thus, the number of expansion variables π_{vj} different from zero. From the dual optimisation problem we know that $s \ll 2(n+m)$. It is a fact that $\sup_{\pi\in s\mu\cdot B_1}|\langle v, \pi\rangle| = s\mu\|v\|_\infty$ (see Theorems II.2.3 and II.2.4 in Werner [15]). Let $L \in \mathbb{R}^{n\times 2(n+m)}$ be the concatenated matrix $L = (L_1\ |\ L_2)$, where L_1 and L_2 are the upper parts of the Gram matrices K_1 and K_2 according to Sect. 2. From the definition we see that $v = \sigma^T L$ and, hence,

$$\begin{aligned} s\mu\|v\|_\infty = s\mu\|\sigma^T L\|_\infty &\leq s\mu\|\sigma\|_\infty\|L\|_\infty \leq s\mu\|L\|_\infty \\ &= s\mu \max_{i=1,\ldots,n}\sum_{j=1}^{n+m}\sum_{v=1,2}|k_v(x_i, x_j)|. \end{aligned}$$

Finally, we obtain the desired upper bound for the empirical Rademacher complexity of ε-CoSVR

$$\hat{\mathcal{R}}_n(\mathcal{H}_\Sigma^{\varepsilon}) \leq \frac{2}{n}\mathbb{E}^\sigma s\mu\|L\|_\infty \leq \frac{2s}{n}\mu(\|L_1\|_\infty + \|L_2\|_\infty).$$

(ii) Having the Lipschitz continuity of the ε-insensitive loss ℓ^L, the claim is a direct consequence of Theorem 3 in the work of Rosenberg and Bartlett [9], which finishes the proof. □



(a) (b) (c)

Fig. 1: Comparison of ε-CoSVR, ℓ_2-CoSVR, and Σ-CoSVR with the baselines CoRLSR, SVR(concat), and SVR(best) on 24 datasets using the fingerprints GpiDAPH3 and ECFP4 in terms of RMSEs. Each point represents the RMSEs of the two methods compared on one dataset.

4 Empirical Evaluation

In this section we evaluate the performance of the CoSVR variants for predicting the affinity values of small compounds against target proteins.

Our experiments are performed on 24 datasets consisting of ligands and their affinity to one particular human protein per dataset, gathered from BindingDB. Every ligand is a single molecule in the sense of a connected graph and all ligands are available in the standard molecular fingerprint formats ECFP4, GpiDAPH3, and Maccs. All three formats are binary and high-dimensional. An implementation of the proposed methods and baselines, together with the datasets and experiment descriptions are available as open source[7].

We compare the CoSVR variants ε-CoSVR, ℓ_2-CoSVR, and Σ-CoSVR against CoRLSR, as well as SVR with a single-view (SVR([fingerprint name])) in terms of root mean squared error (RMSE) using the linear kernel. We take the two-view setting in our experiments as we want to include Σ-CoSVR results in the evaluation. Another natural baseline is to apply SVR to a new view that is created by concatenating the features of all views (SVR(concat)). We also compare the CoSVR variants against an oracle that chooses the best SVR for each view and each dataset (SVR(best)) by taking the result with the best performance in hindsight.

We consider affinity prediction as semi-supervised learning with many unlabelled data instances. Therefore, we split each labelled dataset into a labelled (30% of the examples) and an unlabelled part (the remaining 70%). For the co-regularised algorithms, both the labelled and unlabelled part are employed for training, i.e., in addition to labelled examples they have access to the entire set of unlabelled instances without labels. Of course, the SVR baselines only consider the labelled examples for training. For all algorithms the unlabelled part is used for testing. The RMSE is measured using 5-fold

[7] CoSVR open source repository, https://bitbucket.org/Michael_Kamp/cosvr

Table 2: Comparing RMSEs using Wilcoxon signed-rank test (hypothesis test on whether CoSVR has significantly smaller RMSEs than the baselines).

	ε-CoSVR		ℓ_2-CoSVR		Σ-CoSVR	
baseline	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**
CoRLSR	8.0	$< \mathbf{0.00005}$	13.0	$< \mathbf{0.00009}$	70.0	$< \mathbf{0.02226}$
SVR(GpiDAPH3)	1.0	$< \mathbf{0.00002}$	1.0	$< \mathbf{0.00002}$	1.0	$< \mathbf{0.00002}$
SVR(ECFP4)	22.5	$< \mathbf{0.00027}$	44.0	$< \mathbf{0.00738}$	94.0	< 0.1096
SVR(concat)	3.0	$< \mathbf{0.00003}$	24.0	$< \mathbf{0.00032}$	79.5	$< \mathbf{0.04397}$
SVR(best)	27.0	$< \mathbf{0.00044}$	56.0	$< \mathbf{0.02208}$	88.0	< 0.07649

cross-validation. The parameters for each approach on each dataset are optimised using grid search with 5-fold cross-validation on a sample of the training set.

In Fig. 1 we present the results of the CoSVR variants compared to CoRLSR (a), SVR(concat) (b), and SVR(best) (c) for all datasets using the fingerprints GpiDAPH3 and ECFP4. Fig. 1 (a), (b) indicate that all CoSVR variants outperform CoRLSR and SVR(concat) on the majority of datasets. Fig. 1 (c) indicates that SVR(best) performs better than the other baselines but is still outperformed by ε-CoSVR and ℓ_2-CoSVR. Σ-CoSVR performs similar to SVR(best).

The indications in Fig. 1 are substantiated by a *Wilcoxon signed-rank test* on the results (presented in Table 2). In this table, we report the test statistics (Z and p-value). Results in which a CoSVR variant statistically significantly outperforms the baselines (for a significance level $p < 0.05$) are marked in bold. The test confirms that all CoSVR variants perform statistically significantly better than CoRLSR and SVR(concat). Moreover, ε-CoSVR and ℓ_2-CoSVR statistically significantly outperform an SVR trained on each individual view as well as taking the best single-view SVR in hindsight. Although Σ-CoSVR performs slightly better than SVR(best), the advantage is not statistically significant.

In Table 3 we report the average RMSEs of all CoSVR variants, CoRLSR and the single-view baselines for all combinations of the fingerprints Maccs, GpiDAPH3, and ECFP4. In terms of average RMSE, ε-CoSVR and ℓ_2-CoSVR outperform the other approaches for the view combination Maccs and GpiDAPH3, as well as GpiDAPH3 and ECFP4. For the views Maccs and ECFP4, these CoSVR variants have lower average RMSE than CoRLSR and the single-view SVRs. However, for this view combination, the SVR(best) baseline outperforms CoSVR. Note that SVR(best) is only a hypothetical baseline, since the best view varies between datasets and is thus unknown in advance. The Σ-CoSVR performs on average similar to CoRLSR and the SVR(concat) baseline and slightly worse than SVR(best). To avoid confusion about the different performances of Σ-CoSVR and ℓ_2-CoSVR, we want to point out that Σ-CoSVR equals ℓ_2-CoSVR^{mod} (see Lemma 5) and not ℓ_2-CoSVR (equivalent with ℓ_2-CoSVR$_{mod}$ for $M = 2$) which we use for our experiments.

The advantage in learning performance of ε-CoSVR and ℓ_2-CoSVR comes along with the cost of a higher runtime as shown in Fig. 4. In concordance with the theory, Σ-CoSVR equalises the runtime disadvantage with a runtime similar to the single-view methods.

In conclusion, co-regularised support vector regression techniques are able to exploit the information from unlabelled examples with multiple sparse views in the practical setting of ligand affinity prediction. They perform better than the state-of-the-art single-view approaches [12], as well as a concatenation of features from multiple views. In particular, ε-CoSVR and ℓ_2-CoSVR outperform the multi-view approach CoRLSR [4] and SVR on all view combinations. ℓ_2-CoSVR outperforms SVR(concat) on all, ε-CoSVR on 2 out of 3 view combinations. Moreover, both variants outperform SVR(best) on 2 out of 3 view combinations.

Table 3: Average RMSEs for all combinations of the fingerprints Maccs, GpiDAPH3, and ECFP4

Method	**View Combinations**		
	Maccs, ECFP4	**Maccs, GpiDAPH3**	**GpiDAPH3, ECFP4**
ε-CoSVR	1.035	**1.016**	**1.049**
ℓ_2-CoSVR	1.007	1.019	1.062
Σ-CoSVR	1.116	1.114	1.151
CoRLSR	1.06	1.073	1.199
SVR(view1)	1.04	1.041	1.355
SVR(view2)	1.094	1.37	1.106
SVR(concat)	1.011	1.12	1.194
SVR(best)	**0.966**	1.027	1.104

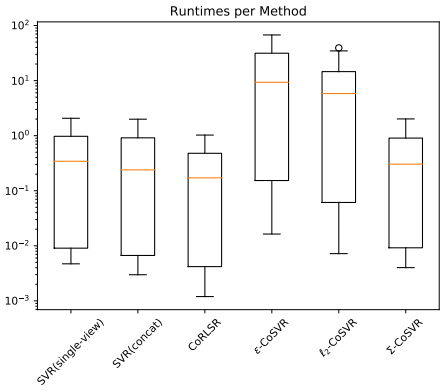


Fig. 2: Runtimes of the CoSVR variants, CoRLSR, and single-view SVRs on 24 ligand datasets and all view combinations (runtime in log-scale).

5 Conclusion

We proposed CoSVR as a semi-supervised multi-view regression method that copes with the practical challenges of few labelled data instances and multiple adequate views on data. Additionally, we presented CoSVR variants with considerably reduced numbers of variables and a version with substantially decreased runtime. Furthermore, we proved upper bounds on the Rademacher complexity for CoSVR. In the experimental part, we applied CoSVR successfully to the problem of ligand affinity prediction. The variants ε-CoSVR and ℓ_2-CoSVR empirically outperformed the state-of-the-art approaches in ligand-based virtual screening. However, this performance came at the cost of solving a more complex optimisation problem resulting in a higher runtime than single-view approaches. The variant Σ-CoSVR still outperformed most state-of-the-art approaches with the runtime of a single-view approach.

Bibliography

[1] Bender, A., Jenkins, J.L., Scheiber, J., Sukuru, S.C.K., Glick, M., Davies, J.W.: How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. J. Chem. Inf. Model. (2009)

[2] Blum, A., Mitchell, T.: Combining Labeled and Unlabeled Data with Co-Training. In: Proceedings of the 11th Annual Conference on Learning Theory (1998)

[3] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)

[4] Brefeld, U., Gärtner, T., Scheffer, T., Wrobel, S.: Efficient Co-Regularised Least Squares Regression. In: Proceedings of the 23rd International Conference on Machine Learning (2006)

[5] Farquhar, J.D.R., Meng, H., Szedmak, S., Hardoon, D., Shawe-Taylor, J.: Two view learning: SVM-2K, Theory and Practice. In: Advances in Neural Information Processing Systems 18 (2006)

[6] Geppert, H., Humrich, J., Stumpfe, D., Gärtner, T., Bajorath, J.: Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. J. Chem. Inf. Model. (2009)

[7] Myint, K.Z., Wang, L., Tong, Q., Xie, X.Q.: Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. Mol. Pharmaceutics (2012)

[8] Nisius, B., Bajorath, J.: Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of Hits. Chem. Biol. Drug Des. (2010)

[9] Rosenberg, D.S., Bartlett, P.L.: The Rademacher Complexity of Co-Regularized Kernel Classes. In: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (2007)

[10] Schölkopf, B., Herbrich, R., Smola, A.J., Williamson, R.: A Generalized Representer Theorem. In: Proceedings of the Annual Conference on Computational Learning Theory (2001)

[11] Sindhwani, V., Rosenberg, D.S.: An RKHS for Multi-View Learning and Manifold Co-Regularization. In: Proceedings of the 25th International Conference on Machine Learning (2008)

[12] Sugaya, N.: Ligand Efficiency-Based Support Vector Regression Models for Predicting Bioactivities of Ligands to Drug Target Proteins. J. Chem. Inf. Model. (2014)

[13] Ullrich, K., Mack, J., Welke, P.: Ligand Affinity Prediction with Multi-Pattern Kernels. In: Proceedings of Discovery Science (2016)

[14] Wang, X., Ma, L.and Wang, X.: Apply semi-supervised support vector regression for remote sensing water quality retrieving. IEEE International Geoscience and Remote Sensing Symposium (2010)

[15] Werner, D.: Funktionalanalysis. Springer (1995)

[16] Xu, C., Tao, D., Xu, C.: A Survey on Multi-view Learning. arXiv (2013)