Comparing active site sequence representations for kinase-ligand affinity prediction

Abstract

This document is an addendum to our recent publication "*Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction and Inhibitor Generation: 3D Effects in a 1D Model*" [Born et al., 2021]. All experiments in our original paper were based on an active site definition from Sheridan et al. [2009] that comprises 29 residues. In this report, we investigate whether an alternative definition of the kinase active site, comprising 16 residues following the analysis in Martin and Mukherjee [2012] can further improve the predictive power for kinase-ligand affinity prediction in proteochemometric models. We additionally investigate a "combined" active site definition with 36 residues contained in either of the two definitions.

Our results show that, for predicting affinity for unseen kinases, there is no clear evidence in favor of any active site definition compared to the other two. However, for predicting affinity for unseen ligands, there is significant evidence that the combined kinase representation is superior to the Sheridan as well as the Martin representation. In sum, our results corroborate the finding that superior performance in kinase-ligand affinity prediction can be achieved when restricting to a subset of residues rather than considering the full protein sequence.

1 Experimental setup

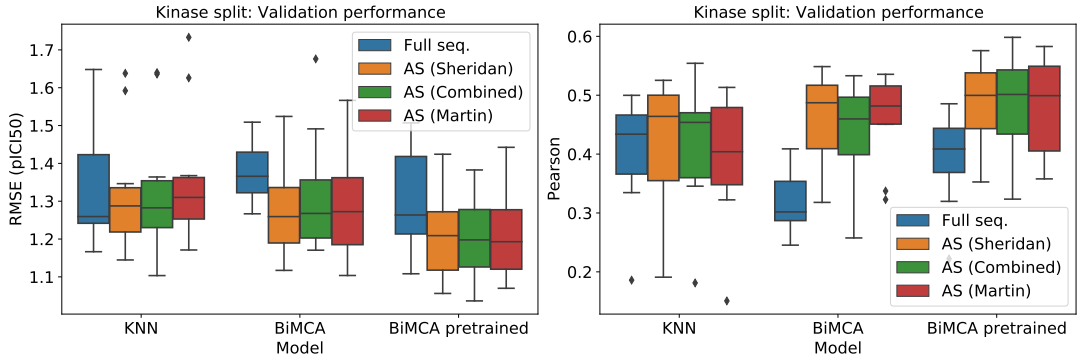
The experimental setup for the binding affinity prediction task is identical to the one described in Born et al. [2021]. The active site representation in that paper relies on 29 residues defined originally in Sheridan et al. [2009]. For a table with the PKA numbering of the 29 residues and for details of the global sequence alignment process, please see Sheridan et al. [2009, Table 1]. The predictive power of this active site definition is now compared to the active site definition by Martin and Mukherjee [2012] which involves 16 residues. These 16 residues were identified empirically from a starting set of 46 residues based on how frequently they were picked by a large set of kinase-kernel models. 10 of these 16 residues are overlapping with the definition of Sheridan et al. [2009]. To explore the potential benefit of combining both definitions, we also investigate a joint active site definition, comprising a total of 36 residues.

For both the ligand and the kinase split, we use the exact same 10-fold cross validation as reported in the paper Born et al. [2021]. In both cases, 10% of the samples were held out for independent testing, prior to the cross validation. For all experiments, the The "full sequence" configuration (blue) and "AS (Sheridan)" configuration are identical to the ones reported in the paper Born et al. [2021].

2 Kinase split

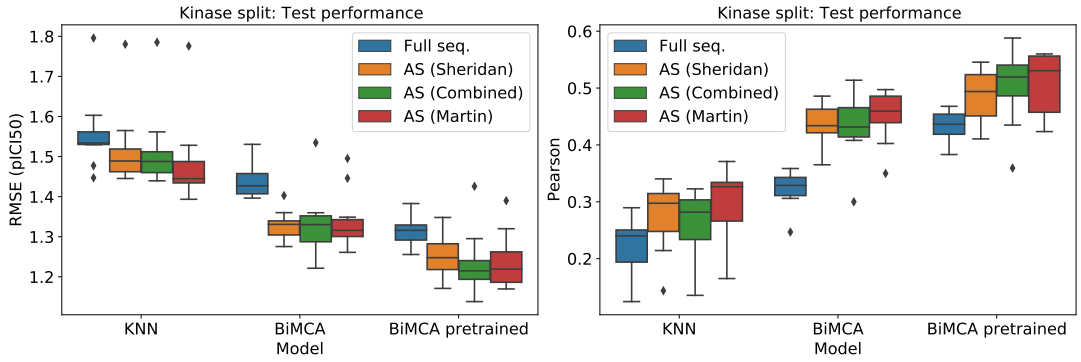
2.1 Results

This split tests the ability of the model to predict the binding affinity for a protein kinase unseen during training. The results on the kinase split for the KNN and the BiMCA on the validation and test data are shown in Figure 1 and Figure 2 respectively. On the validation data, no clear trend is visible



(a) RMSE in binding affinity prediction (b) Pearson's r in binding affinity prediction

Figure 1: Validation performance on kinase split



(a) RMSE in binding affinity prediction (b) Pearson's r in binding affinity prediction

Figure 2: Test performance on kinase split

when comparing the three active site configurations. Statistically, the Sheridan representation is superior to the Martin representation for the KNN ($p < 0.05$, $W+$) and it is superior to the Combined representation for the BiMCA. But these trends do not persist across all three model types.

Instead, on the test data performance shown in Figure 2, the Martin representation consistently obtained the highest Pearson correlation, irrespective of the model. However, this finding does not corroborate when using the RMSE as response metric. Moreover, this trend about the superioty of the Martin representation is only statistically significant for the KNN model, where it outperforms both other AS representations in all metrics ($p < 0.05$, $W+$).

2.2 Discussion

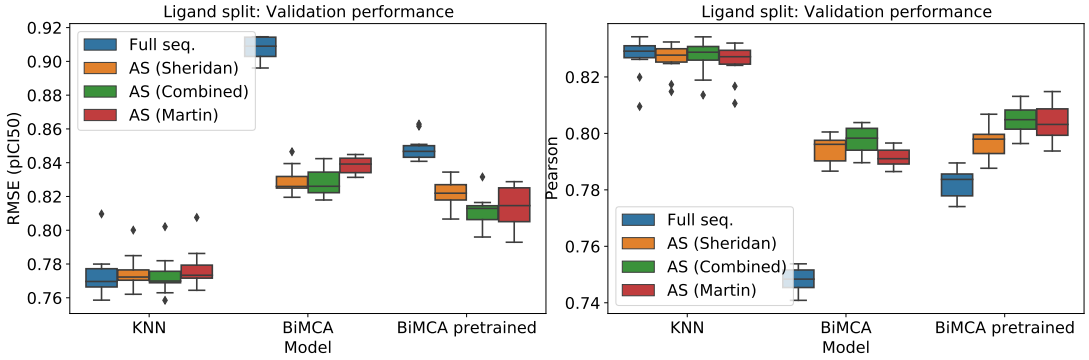
In general, splitting the dataset by kinases induces a strong heterogeneity across each fold/chunk of data. Therefore, care has to be taken in drawing conclusions from the results on the test data because the findings might be due to the composition of kinases in the test set. Since the validation performance averages the performance across the 10 folds from the cross-validation, it is more reliably.

Due to the heterogeneity however, the performance variance is higher and thus, less conclusions can be drawn. Here, an analysis that groups performance by individual kinases or families of kinases could help to further understand the impact of the representation. For example, it is likely that some residues are of critical importance in some kinase families but can be safely ignored in others.

3 Ligand split

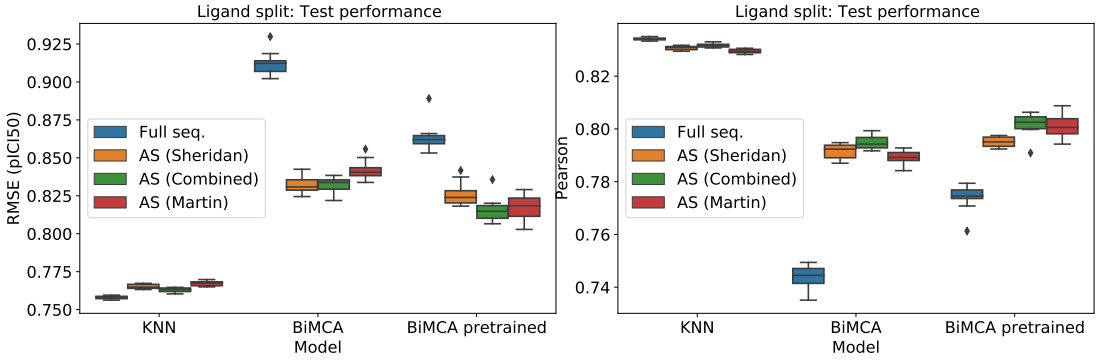
3.1 Results

In the ligand split, kinases are shared across train and validation/test dataset whereas the ligands are not. Therefore, the results of the KNN model can be disregarded for this task because its prediction are based purely on sequence similarity. Note that this implies that the chosen kinase representation does not impact the KNN's prediction in $> 99\%$ of the cases. The results on the ligand split on the validation and test data are shown in Figure 3 and Figure 4 respectively. Like for the ligand split,



(a) RMSE in binding affinity prediction (b) Pearson's r in binding affinity prediction

Figure 3: Validation performance on ligand split



(a) RMSE in binding affinity prediction (b) Pearson's r in binding affinity prediction

Figure 4: Test performance on ligand split

the results indicate that all three active site configurations are clearly superior to the full sequence representation.

Differences *between* the active site representations are less pronounced, however several findings can be extracted:

1. The Combined representation yields consistently the best results for both models (BiMCA and BiMCA pretrained), both metrics and for both the validation and the test data (cf. Table 1).

2. Statistical significance tests with the Wilcoxon signed-rank test show that the Combined representation yields significantly better results than the Sheridan representation for both models.
3. This also applies to a comparison between the Combined and the Martin representation, but only for the BiMCA model and not the pretrained BiMCA model.

Table 1: Results on validation and test data (ligand split). For each model and data partition we mark the better representation in bold.

		RMSE		**Pearson**	
Data	Config	BiMCA	BiMCA (pre.)	BiMCA	BiMCA (pre.)
Val.	AS (Sheridan)	$0.829_{\pm 0.01}$	$0.821_{\pm 0.01}$	$0.794_{\pm 0.01}$	$0.797_{\pm 0.01}$
	AS (Martin)	$0.839_{\pm 0.01}$	$0.813_{\pm 0.01}$	$0.791_{\pm 0.01}$	$\mathbf{0.804}_{\pm 0.01}$
	AS (Combined)	$\mathbf{0.828}_{\pm 0.01}$	$\mathbf{0.811}_{\pm 0.01}$	$\mathbf{0.797}_{\pm 0.01}$	$\mathbf{0.804}_{\pm 0.01}$
Test	AS (Sheridan)	$\mathbf{0.832}_{\pm 0.01}$	$0.826_{\pm 0.01}$	$0.792_{\pm 0.01}$	$0.795_{\pm 0.01}$
	AS (Martin)	$0.842_{\pm 0.01}$	$0.818_{\pm 0.01}$	$0.789_{\pm 0.01}$	$0.801_{\pm 0.01}$
	AS (Combined)	$\mathbf{0.832}_{\pm 0.01}$	$\mathbf{0.816}_{\pm 0.01}$	$\mathbf{0.795}_{\pm 0.01}$	$\mathbf{0.802}_{\pm 0.01}$

3.2 Discussion

We observe significant evidence that the Combined kinase representation is superior to the Sheridan and the Martin representation. This might seem unsurprising given that it simply includes information from more residues than the individual representations, however, keep in mind that this in stark contrast to the main message of our analysis, i.e., incorporating *less* residues yields better results. Even after this work, it certainly remains unclear what the ideal set of residues for kinase-ligand affinity prediction is. However, for the classic drug discovery setting, where the goal is to predict affinity for an unseen ligand on a known target, it is likely to exist a sweat spot somewhere between the full sequence and the active site definition from Sheridan et al. [2009]. Our findings suggest that taking together the residues identified by Sheridan et al. [2009] and Martin and Mukherjee [2012] moves us closer to this sweat spot.

References

J. Born, T. Huynh, A. Stroobants, W. D. Cornell, and M. Manica. Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3d effects in a 1d model. *Journal of Chemical Information and Modeling*, 2021.

E. Martin and P. Mukherjee. Kinase-kernel models: accurate in silico screening of 4 million compounds across the entire human kinome. *Journal of chemical information and modeling*, 52 (1):156–170, 2012.

R. P. Sheridan, K. Nam, V. N. Maiorov, D. R. McMasters, and W. D. Cornell. Qsar models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *Journal of chemical information and modeling*, 49(8):1974–1985, 2009.