Active site sequence representation of human kinases outperforms full sequence for affinity prediction

Jannis Born
IBM Research Europe
jab@zurich.ibm.com

Tien Huynh
IBM Research Yorktown

Astrid Stroobants
Department of Chemistry,
Imperial College London

Wendy D. Cornell
IBM Research Yorktown

Matteo Manica
IBM Research Europe
tte@zurich.ibm.com

Abstract

Focusing on the human kinome, we challenge a standard practice in proteochemometric, sequence-based affinity prediction models: instead of leveraging the full primary structure of proteins, each target is represented only by a sequence of 29 residues defining the ATP binding site. In kinase-ligand binding prediction, our results show that the reduced active site sequence representation is not only computationally more efficient but consistently yields significantly higher performance than the full primary structure. This trend persists across different models (a *k*-NN baseline and a multimodal deep neural network), datasets (BindingDB, IDG-DREAM), performance metrics (RMSE, Pearson correlation) and holds true when predicting affinity for both unseen ligands and kinases. For example, the RMSE on pIC50 can be reduced by 5% and 9% respectively for unseen kinases and kinase inhibitors. This trend is robust across kinases' families and classes of inhibitors with a few exceptions where the necessity of full sequence is explained by the drugs mechanism of action. Our interpretability analysis further demonstrates that, even without supervision, the full sequence model can learn to focus on the active site residues to a higher extent. Overall, this work challenges the assumption that full primary structure is indispensable for virtual screening of human kinases.

1 Introduction

Protein kinases are ubiquitous for cell life and have become a vital source of targets for drug discovery in the past 20 years [8, 9, 10]. Computational methods have supported our understanding of kinases and their inhibitors in many regards, e.g., compound protein interaction (CPI) prediction [27, 26, 28, 34] and drug response prediction [17, 20]. While early approaches to kinase affinity prediction were single-assay [27], or single-target models [26], proteochemometric approaches consider both chemical and protein information and can generalize to novel ligands and targets simultaneously [7, 13]. To avoid the need for costly protein structure information, most recent work relied only on primary structure information, like amino acid (AA) sequences, coupled with SMILES [42] strings to represent ligands [13, 18, 30, 15, 44, 6]. This model class has dominated the recent IDG-DREAM challenge [7]. Their winning model is highly similar to the BiMCA model [3] which is utilized herein.

Our contribution. In this work, we are first to systematically compare the impact of using active site residues and full sequence information to represent kinases for 1D proteochemometric modeling of drug-kinase binding affinity prediction (for overview see Figure 1). We perform all experiments on both representations and investigate two models. First, a simple, yet efficient and novel KNN regression model based on Levenshtein distance [21] of proteins and, secondly, a bimodal deep

Machine Learning for Structural Biology Workshop, NeurIPS 2021.

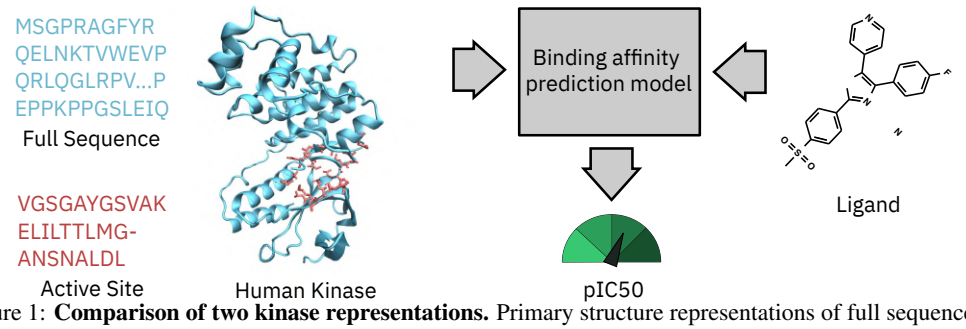


Figure 1: **Comparison of two kinase representations.** Primary structure representations of full sequence and active sites for human kinases are evaluated on binding affinity prediction to ligands.

neural network, the BiMCA model [3], which follows the current state-of-the-art and dispenses with traditional descriptors by relying solely on interpretable, textual inputs (SMILES and AA sequences).

2 Methods

Problem formulation. Let $\mathcal{P}$ denote the space of proteins, $\mathcal{M}$ the molecular space and $\mathcal{A}$ the affinity scores. We are then interested in learning a function $\Phi_A : \mathcal{P} \times \mathcal{M} \to \mathcal{A}$. The function Φ_A maps a protein-ligand tuple to an affinity score and is learned from the training data set $\mathcal{D} = \{p_i, m_i, a_i\}_{i=1}^{N}$ where $p_i \in \mathcal{P}, m_i \in \mathcal{M}$ and $a_i \in \mathcal{A}$ is the scalar binding strength, the pIC50.
K-Nearest Neighbor (KNN) regression. To address the presented problem, we first use a KNN model based on a joint space spanned by protein and ligand similarity. Kinases are represented by primary structure (either full sequence or only active site) and molecules by their ECFP4 fingerprint [32]. As a distance metric between samples we utilize a combination of the length-normalized Levenshtein distance for the primary structure and the Tanimoto similarity [36] of molecules, similar to the TITAN model for protein-protein interaction prediction [41]. More formally, let $\{p_j, m_j\}$ denote an unseen sample from the test dataset $\mathcal{D}_{Test} = \{p_i, m_i\}_{i=1}^{N_{Test}}$. With the goal of predicting $\hat{a}_j$ to approximate the unknown a_j, we first retrieve the subset of training data $\mathcal{D}_k$ containing the k nearest neighbors using the distance measure: $\mathbf{D}(p_i, m_i, p_j, m_j) = \frac{Lev(p_i, p_j)}{max(|p_i|, |p_J|)} + (1 - \mathcal{T}(m_i, m_j))$ Here, $|\cdot|$ denotes sequence length, $\mathcal{T}$ is the Tanimoto similarity measure and $Lev(\cdot, \cdot)$ is the Levenshtein distance [21]. Then, the prediction $\hat{a}_j$ is trivially computed by $\hat{a}_j = \frac{\sum_i^k a_i}{k}$ with $a_i \in \mathcal{D}_k$.
Bimodal multiscale convolutional attention (BiMCA) network. Alternatively, we utilize the BiMCA model to learn a function $\Phi_A : \mathcal{P} \times \mathcal{M} \to \mathcal{A}$ based on primary structure of proteins and SMILES sequences of molecules. As visualized in Figure A1, this model separately ingests an amino acid sequence and a SMILES sequence, converts the tokens into (learned) embedding vectors, performs 1D convolutions to aggregate local substructures, applies a contextual attention mechanism and then outputs a scalar pIC50 score. For details see appendix subsection A1.1.
Data. We extracted compound-protein interaction data from BindingDB [14] and curated a dataset consisting of $206,889$ protein-ligand pairs and their associated pIC50 score. The samples were distributed across $113,475$ ligands and 349 human kinases. The remaining data (i.e., all non-kinome samples) contained $485,461$ samples (2856 proteins, $331,169$ ligands) and was used in one configuration for pretraining the BiMCA model. For details on the data curation see subsection A1.2.
Human kinase sequence alignment. The binding site residues for each kinase were identified by applying the binding site definition of protein kinase A [34] to a structurally-validated multiple sequence alignment of 497 human protein kinase domains [29]. Sheridan's definition identified 29 residues representing the ATP binding site including but not limited to contributions from the Gly-rich-loop, gatekeeper, hinge, and DFG-in-out.

3 Results

Kinase data split. The kinase split (i.e., predicting affinity for unseen kinases) is the ideal setup to test the impact of the protein representation. It is more challenging than splitting on ligands because the shape of the binding pocket largely governs binding activity [40]. This task is highly relevant due to the *hidden ligand bias* [35], i.e., the observation that binding affinity predictions are mostly based

on ligand rather than interaction features [6]. The results of a 10-fold cross validation (CV) of all three models show a consistent and strong superiority of the active site models (cf. Figure 2).

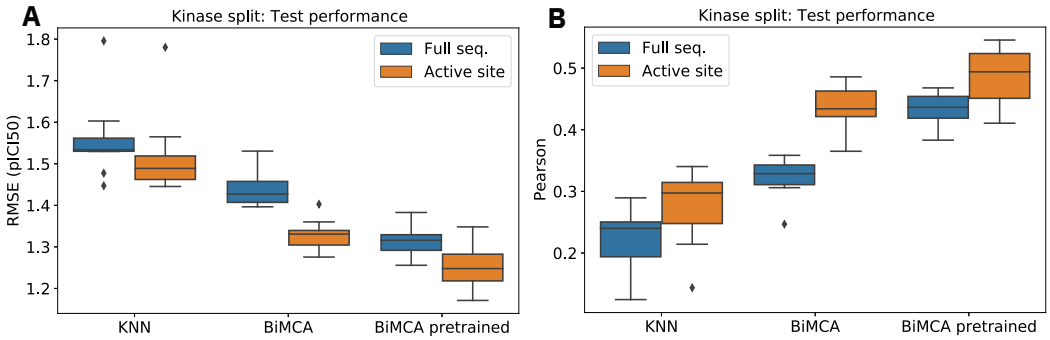


Figure 2: **Binding affinity prediction results on kinase split.** Subfigure **A)** and **B)** respectively show the RMSE and Pearson (PCC) on predicting pIC50 of the test samples (10-fold CV). On both metrics, the active site configurations significantly outperform the full sequence configuration, irrespective of the utilized model. For the exact numerical scores and validation data performance see Table A1 and Table A2.

For all three model configurations, the active site models significantly outperform the full sequence models ($p < 0.01$, Wilcoxon signed-rank test, $W+$). This is remarkable because the full sequence contains an order of magnitude more information (mean sequence length: 742 vs. 29 amino acids) and the active site BiMCA models only have 5% of the parameters of the full sequence model. In the BiMCA *pretrained* setting we exploited all non-kinase data from BindingDB to warm up the BiMCA model before finetuning on the human kinome. Notably, all pretrained BiMCA models outperform the regular ones, demonstrating that patterns of protein-ligand interactions benefit the development of kinase-inhibitor affinity prediction models. When comparing the performance for the eight different groups of conventional protein kinases (ePK, classification by [16], mapped with the catalogue from [24]), the superiority of the active site configuration is consistent across the kinase families with few exceptions (cf. Figure A4). Only for the TKL group in the KNN and the STE and CMGC group in the BiMCA, the full sequence model achieves better performance than the active site model. The TKL results do not resonate with the remaining findings on the KNN because many TKL kinases (e.g., all RAF kinases [11]) have multiple binding sites which are not captured in the active site sequence alone. To verify that the prediction performance does not hinge on the availability of similar kinases in the training data, we investigated the per-kinase performance as a function of the similarity to the nearest neighbor in the training data (cf. Figure A5). While all PCCs are positive, none of them exceed values of 0.11 suggesting that our models do not require data from similar kinases to work well. Last, in a comparison with previous work, we verified the generalization abilities of our model toward four held-out protein families (Estrogen Receptors, Ion Channels, RTKs and GPCRs). The results in Table A6 show that across all previous works, the BiMCA performed best. These experiments were solely done with the full-sequence models due to the lack of active site data for non-kinase proteins.

Ligand data split. This split (i.e., predicting affinity for unseen ligands) is the classical setting of kinase inhibitor discovery. The results on the test dataset of the 10-fold CV show that, like in the ligand split, all BiMCA active site models are superior to their full sequence analogs (8.2% and 4.7% RMSE improvement for BiMCA and pretrained BiMCA, cf Table 1). Again, for both models, these differences are statistically significant across the ten folds for both validation and test data as well as RMSE and pearson correlation as metrics ($p < 0.001$, $W+$). However, Table 1 indicates that the KNN model performed similarly well on both active sites and full sequences. This is because the protein information is of negligible performance for our KNN model *in a ligand split*. When retrieving the KNN, the first addend collapses to 0 for all samples of the same kinase, irrespective of whether active site or full sequence information is used. This dilutes differences between the representations and indeed, for 98.9% and 99.3% of the predicted samples, the nearest neighbor is a sample with the same kinase. To remedy this confound, we evaluated the KNN performance exclusively on the remaining samples. In alignment with the overall findings, the active site model is clearly superior for these samples (RMSE 1.35 vs. 1.59, Pearson's r 0.56 vs. 0.33 on the test data). In Figure 3 we investigate the performance of the model for different groups of kinase inhibitors; assessed by

Table 1: Results on test dataset (ligand split).

Model	Config	RMSE	PCC
KNN	Full seq.	**0.76**±0.00	**0.83**±0.01
	Active site	0.77±0.00	**0.83**±0.01
BiMCA	Full seq.	0.91±0.01	0.74±0.00
	Active site	**0.83**±0.01	**0.79**±0.00
BiMCA pretrain	Full seq.	0.86±0.01	0.77±0.01
	Active site	**0.80**±0.01	**0.83**±0.01

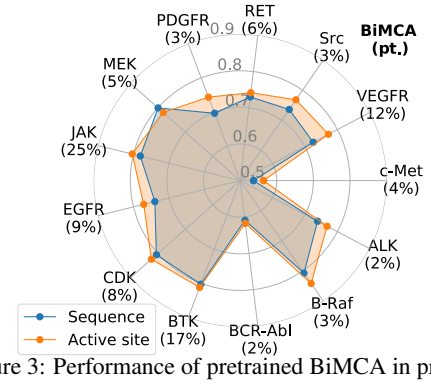


Figure 3: Performance of pretrained BiMCA in predicting affinity for unseen kinase inhibitors according to their primary protein target class.

the primary target for each kinase inhibitor, aggregated into thirteen groups of alleged mechanism of action based on an established classification [33]. With the exception of MEK inhibitors, the active site model performed better on all thirteen kinase inhibitor groups. Given that our sequence alignment only relied on ATP binding site residues [29], we hypothesize that the increased MEK (i.e., MAPK/ERK) inhibitor performance for full sequence models is due to the discovery of several ATP-noncompetitive MEK inhibitors that bind to a unique site near the ATP binding pocket [43]. In support of that, 94% of the 2909 MEK-inhibitor related samples making up this effect are indeed accounted for by eight kinases of the MAPK family. In other words, for most MEK inhibitors the binding pocket is not contained in the active site sequences. Equivalent to the kinase split, we verified that model performance does not hinge upon the availability of similar molecules during training and find only a very weak negative correlation between the per-ligand RMSE and the ECFP4-Tanimoto similarity to the nearest neighbor in training data (cf. Figure A6).

Validation on external dataset. To verify our hypothesis on an independent dataset, we utilized the IDG-DREAM challenge data [7]. The challenge focused on under-studied parts of the human kinome to catalogue the unexplored target space of kinase inhibitors and thus resembles a challenging dataset of 720 samples (for details on data processing see subsection A1.3). The results on this dataset are in alignment with our overall findings (cf. Table A5). The active site residues outperforms the full sequence information consistently in both models and the BiMCA yields better results than the KNN model. Notably, the active site BiMCA is the only model that achieves a satisfying performance in predicting activity in the under-studied and unseen kinases [7].

Model attention analysis. Given the surprising finding that providing *more* information on proteins hampers performance, we sought to examine whether the full sequence models had learned to recover tertiary structure information. For two exemplary kinases, MAPK11 and ABL1, we analyzed the attention scores of the BiMCA model (cf. Figure A8). This is an ante-hoc interpretability method that automatically assigns an attention (or relevance) score to each amino acid as well as SMILES token during prediction. For both kinases, the mean attention scores on the active site residues are significantly higher than on the remaining residues ($\alpha = 0.05\%$, *MWU*). In alignment with all previous attention-based models in CPI prediction, the BiMCA model also does not convincingly predict the active/interaction site when trained exclusively on binding affinity labels. On the positive side, however, it does exhibit a mild ability to focus on the relevant residues. In contrast to these subtle 3D effects in the full sequence model, the active site models convey the 3D information by design more prominently – which might contribute to their improved generalizability compared to full sequence models.

4 Discussion and conclusion

Here, we investigated proteochemometric modelling of CPI in light of different protein representations and report a superiority of active site residues to full protein sequences when predicting binding affinity for novel ligands as well as kinases. These findings are robust across two models (KNN and BiMCA) and datasets (BindingDB and IDG-DREAM). This is an important finding because the active site residues are a tiny subset of the full primary sequence which additionally codes for more distant determinants of binding dynamics. It seems that providing exclusively the active site residues

increases the signal-to-noise ratio in the sequences, consequently leading to better performance. Even without supervision on the importance of the residues, the full sequence model learns to focus to a significantly higher extent on the active site residues. While this might be interpreted as evidence that the model recovers elements of tertiary structure from the sequence information alone, we notice that this is a highly controversial and active topic in sequence-based CPI models [18, 12, 38, 22] and protein language models in general [39]. Another important finding is that the active site models even outperformed the full sequence models when both models were pretrained on full sequences, suggesting that proteochemoetric models benefit from pretraining on large-scale pan-protein data even if the final use case is limited to one family. Overall, our results suggest that "*more is less*" in sequence-based kinase affinity prediction models.

Availability

To facilitate reproduction of the results and ease comparison to other methods, the source code as well as the processed data (including the derived active sites) is publicly available from the following `GitHub` repository: `https://github.com/PaccMann/paccmann_kinase_binding_residues`.

References

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Massoud Amanlou, Jahan B Ghasemi, and Ali Masoudi-Nejad. Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. *Bioinformatics*, 36(17):4633–4642, 2020.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] Jannis Born, Matteo Manica, Joris Cadow, Greta Markert, Nil Adell Mill, Modestas Filipavicius, Nikita Janakarajan, Antonio Cardinale, Teodoro Laino, and María Rodríguez Martínez. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Machine Learning: Science and Technology*, 2(2):025024, 2021.

[4] Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and María Rodríguez Martínez. Paccmann^{RL}: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, 24(4):102269, 2021.

[5] Joris Cadow, Jannis Born, Matteo Manica, Ali Oskooei, and María Rodríguez Martínez. Paccmann: a web service for interpretable anticancer compound sensitivity prediction. *Nucleic acids research*, 48(W1):W502–W508, 2020.

[6] Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

[7] Anna Cichońska, Balaguru Ravikumar, Robert J Allaway, Fangping Wan, Sungjoon Park, Olexandr Isayev, Shuya Li, Michael Mason, Andrew Lamb, Ziaurrehman Tanoli, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature communications*, 12(1):1–18, 2021.

[8] Philip Cohen. Protein kinasesthe major drug targets of the twenty-first century? *Nature reviews Drug discovery*, 1(4):309–315, 2002.

[9] Philip Cohen and Dario R Alessi. Kinase drug discovery–whats next in the field? *ACS chemical biology*, 8(1):96–104, 2013.

[10] Philip Cohen, Darren Cross, and Pasi Janne. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery*, 2021.

[11] Andreas Fischer, Angela Baljuls, Joerg Reinders, Elena Nekhoroshkova, Claudia Sibilski, Renate Metz, Stefan Albert, Krishnaraj Rajalingam, Mirko Hekman, and Ulf R Rapp. Regulation of raf activity by 14-3-3 proteins: Raf kinases associate functionally with both homo-and heterodimeric forms of 14-3-3 proteins. *Journal of Biological Chemistry*, 284(5):3183–3194, 2009.

[12] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In *IJCAI*, volume 2018, pages 3371–3377, 2018.

[13] HA Gaspar, M Ahmed, T Edlich, B Fabian, Z Varszegi, M Segler, J Meyers, and M Fiscato. Proteochemometric models using multiple sequence alignments and a subword segmented masked language model. *ChemRxiv preprint (10.26434/chemrxiv.14604720.v1)*, 2021.

[14] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

[15] Adam Gonczarek, Jakub M Tomczak, Szymon Zaręba, Joanna Kaczmar, Piotr Dąbrowski, and Michał J Walczak. Interaction prediction in structure-based virtual screening using deep learning. *Computers in biology and medicine*, 100:253–258, 2018.

[16] Steven K Hanks and Tony Hunter. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1. *The FASEB journal*, 9(8):576–596, 1995.

[17] Liang-Chin Huang, Wayland Yeung, Ye Wang, Huimin Cheng, Aarya Venkat, Sheng Li, Ping Ma, Khaled Rasheed, and Natarajan Kannan. Quantitative structure–mutation–activity relationship tests (qsmart) model for protein kinase inhibitor response prediction. *BMC bioinformatics*, 21(1):1–22, 2020.

[18] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

[19] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2015.

[20] Krzysztof Koras, Ewa Kizling, Dilafruz Juraeva, Eike Staub, and Ewa Szczurek. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *bioRxiv*, 2021.

[21] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[22] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10(4):308–322, 2020.

[23] Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Saez-Rodriguez, and Mara Rodriguez Martinez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular pharmaceutics*, 16(12):4797–4806, 2019.

[24] Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[25] Greta Markert, Jannis Born, Matteo Manica, Gisbert Schneider, and Maria Rodriguez Martinez. Chemical representation learning for toxicity prediction. *PharML Workshop at ECML-PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases)*, 2020.

[26] Eric Martin and Prasenjit 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.

[27] Eric Martin, Prasenjit Mukherjee, David Sullivan, and Johanna Jansen. Profile-qsar: a novel meta-qsar method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. *Journal of chemical information and modeling*, 51(8):1942–1956, 2011.

[28] Eric J Martin, Valery R Polyakov, Li Tian, and Rolando C Perez. Profile-qsar 2.0: kinase virtual screening accuracy comparable to four-concentration ic50s for realistically novel compounds. *Journal of chemical information and modeling*, 57(8):2077–2088, 2017.

[29] Vivek Modi and Roland L Dunbrack. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Scientific reports*, 9(1):1–16, 2019.

[30] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[32] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[33] Robert Roskoski Jr. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological research*, 103:26–48, 2016.

[34] Robert P Sheridan, Kiyean Nam, Vladimir N Maiorov, Daniel R McMasters, and Wendy 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.

[35] Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *Journal of chemical information and modeling*, 59(3):947–961, 2019.

[36] Taffee T Tanimoto. Elementary mathematical theory of classification and prediction. 1958.

[37] Timothy F Truong Jr. *Interpretable deep learning framework for binding affinity prediction*. PhD thesis, Massachusetts Institute of Technology, 2020.

[38] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

[39] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[40] Huiwen Wang, Jiadi Qiu, Haoquan Liu, Ying Xu, Ya Jia, and Yunjie Zhao. Hkpocket: human kinase pocket database for drug design. *BMC bioinformatics*, 20(1):1–11, 2019.

[41] Anna Weber, Jannis Born, and Maria Rodriguez Martinez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement 1):i237–i244, 07 2021.

[42] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[43] Pui-Kei Wu and Jong-In Park. Mek1/2 inhibitors: molecular activity and resistance mechanisms. In *Seminars in oncology*, volume 42, pages 849–862. Elsevier, 2015.

[44] Lingling Zhao, Junjie Wang, Long Pang, Yang Liu, and Jun Zhang. Gansdta: Predicting drug-target binding affinity using gans. *Frontiers in genetics*, 10:1243, 2020.

A1 Appendices

A1.1 BiMCA model

As shown in Figure A1, the BiMCA is a bimodal neural network that consumes a SMILES sequence of the ligand and the AA sequence of the kinase.

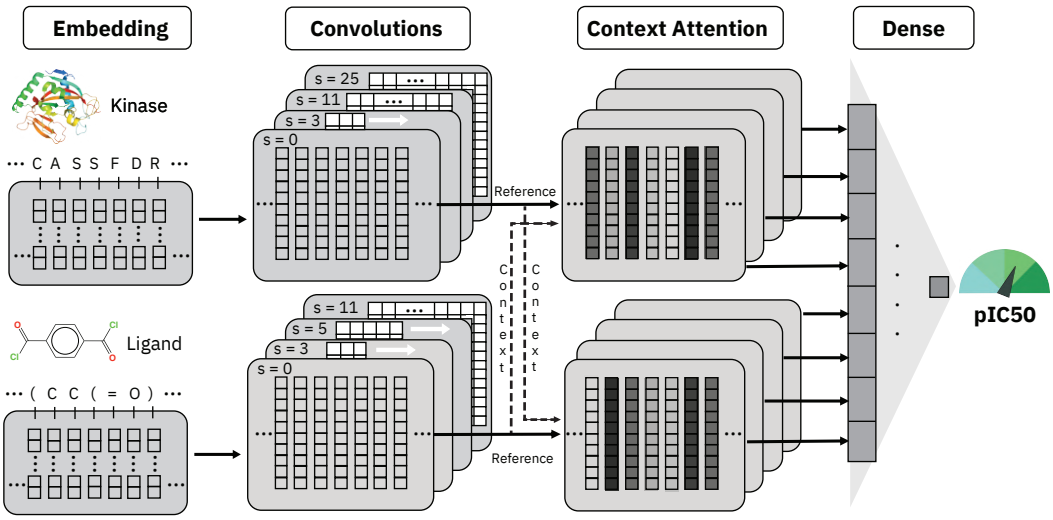


Figure A1: **The bimodal multiscale convolutional attention model (BiMCA).** Both, kinases and ligands are represented as text sequences of amino acids and SMILES respectively. The BiMCA uses learned embeddings and then applies convolutions of multiple kernel sizes c on the embedding matrices (hence the words "multiscale conolutional"). Afterwards, the context attention layers fuse information from both modalities and generate the attention scores over one input modality, using the other modality as context. Black arrows show the information flow through the network, white arrows the direction of the convolution sliding. Figure adjusted from [41].

The SMILES sequences of all ligands were padded to a length of 696 and the AA sequences representing the kinase sequences were padded to a length of 2536 in the full sequence case and 32 in the active site case. Both SMILES tokens and AA are represented by learned embedding vectors of dimensionality 32 and 8. We then used four parallel 1D convolutional layers ("multiscale") with kernel sizes of 3, 5, 11 for the ligands and 3, 11 and 25 on the proteins. Thereafter, a contextual attention mechanism combines both input streams and helps the model to focus on relevant substructures of proteins and ligands in light of the other modality. This mechanism is inspired by [2] and was proposed in our previous work [23, 3]. The model automatically assigns attention scores $\alpha_i \in [0, 1]$ to each amino acid and each SMILES token. For brevity, these attention scores are computed as:

$$\alpha_i = \frac{\exp(u_i)}{\sum_j^T \exp(u_j)} \quad \text{, where} \quad \vec{\mathbf{u}} = \tanh\left(\mathbf{X}_1\mathbf{W}_1 + \mathbf{W}_3(\mathbf{X}_2\mathbf{W}_2)\right)\vec{\mathbf{v}} \tag{1}$$

We call $\mathbf{X}_1 \in \mathbb{R}^{T_1 \times C}$ the *reference* input, where $T_1 \in \{T_M, T_P\}$ is the sequence length and C is the number of convolutional filters. Further, $\mathbf{X}_2 \in \mathbb{R}^{T_2 \times C}$ is the *context* input, where $T_2 \in \{T_M, T_P\}, T_1 \neq T_2$ is the sequence length in the other modality. $\mathbf{W}_1 \in \mathbb{R}^{C \times A}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times A}$, $\mathbf{W}_3 \in \mathbb{R}^{T_1 \times T_2}$ and $\vec{\mathbf{v}} \in \mathbb{R}^A$ are learnable parameters.

Because the context attention layer required $\mathcal{O}(nm)$ parameters where n and m are the sequence length of proteins and ligands respectively (for details see [3]), the full sequence model had substantially more parameters than the active site model. To partly counteract this effect, the number of filter kernels in the convolutional layers was 32 and 128 respectively for the full sequence model and the active site model on both modalities. The output of the attention layers was fed to a stack of dense layers with a single output node, interpreted as pIC50 affinity score for the provided protein-ligand pair. In total, the active site model only consisted of 651,891 parameters, less than 5% of the full sequence model (14,242,491). A dropout of 0.3 throughout convolutional and dense layers was

used. Flavors of this model have been used successfully for cancer drug sensitivity prediction [23, 5] and toxicity prediction [25]. All models were implemented in `PyTorch` [31] and used the `pytoda` package [4] for data handling and preprocessing. The BiMCA model optimized a MSE loss with Adam [19] and was trained for 50 epochs with a learning rate of 0.005, a batch size of 128 on a cluster equipped with `POWER8` processors and a single `NVIDIA Tesla P100`. The variant described here is identical to the binding affinity descriptor used in [3] for predicting antiviral activity of potential SARS-CoV-2 inhibitors.

A1.2 BindingDB data curation

We curated compound-protein interaction data from BindingDB [14]. From the 2,222,074 entries of the database as on 22.04.2021, ∼800,000 were retained after removing missing values and duplicates. Afterwards, samples with molecules whose SMILES strings were invalid or longer than 696 tokens, i.e. atoms and/or bonds, were removed. We chose IC50 as binding affinity metric, converted all values to pIC50 (i.e., the negative decimal logarithm of the half-maximal inhibitory concentration) and clipped all values to the interval [2, 11] (1mM to 0.01nM). Last, we filtered out all samples where the target proteins are not kinases. This resulted in $206,889$ samples distributed across $113,475$ ligands (mean pIC50 per ligand: 7.1 ± 1.2) and 349 human kinases (mean pIC50 per kinase: 6.2 ± 0.9). See Figure A2 for an overview of the dataset's statistics.

For example, a notable and strong bias in the dataset is that kinases screened against more ligands tend to have a higher average affinity ($r = 0.39$).

Non-kinase data. The remainder of the above data (i.e., all non-kinome samples) made up 485,461 samples distributed across 2856 proteins and 331,169 ligands. This data was used in one configuration for pretraining the BiMCA model. After 20 epochs of pretraining, this model achieved a RMSE of 0.86 ($r = 0.82$) on the non-kinase data.

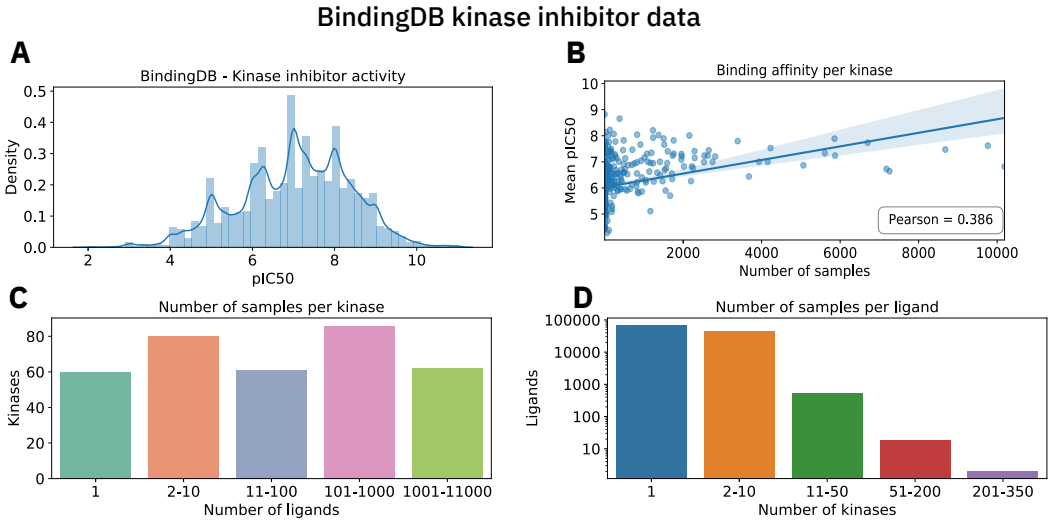


Figure A2: **Visualization of kinase inhibitor data in BindingDB [14]**. **A)** Distribution of pIC50 scores in database ($N = 206,989$). **B)** Kinases with more affinity samples tend to be more promiscuous. **C)** Histogram of number of data points for each kinase. **D)** Most ligands are screened on less than a dozen of kinases but some are screend against almost all 349 kinases.

A1.3 IDG-DREAM data processing

From the initial 825 samples, 720 remained after restriction to kinases with full sequence and active site information [29]. These samples were distributed across 276 kinases (32 unseen) and 93 ligands (all unseen). This data split is much more stringent than the ligand split because for many samples both ligands and kinases are unseen. Additional challenges posed by this dataset compared to BindingDB are 1) experimental differences in the dose-response assays (multi-dose assays with maximal concentration of 10μM that cause an incorrect lower limit for activity) and 2) the dose

response metric, given in logarithmic dissociation constant (pK_d) that differs from the pIC50 in BindingDB. For the KNN model we used all data available in BindingBD as training data whereas for the BiMCA we build an ensemble of the 10 models from the ligand split. Direct comparison with the results reported in the IDG-DREAM challenge is not possible due to the aforementioned differences to our training data.

A1.4 KNN

As KNN is a lazy learning method, the inference runtime scales with the dataset size ($N = 206,990$ samples) and one query thus requires computing almost half a million distances. Therefore, in practice we compute $\mathbf{D}$ not for all training samples but only for those samples $\{p_i, m_i\}$ where either 1) $p_i = p_j$, 2) $m_i = m_j$ or 3) p_i is one of the 10 most similar sequences to p_j in the training dataset. The KNN model was evaluated on all odd $k \leq 25$. For all results, we choose a value of $k = 13$ as this led to the lowest RMSE on the validation dataset on the ligand split (see Figure A7).

A1.5 Data splitting strategies

For proteochemometric models there are four different splitting strategies (see Figure A3). Here, we focus on two of these regimes, namely splitting affinity data based on ligands (while not controlling for proteins) as well as the reverse task.

Ligand split. Generalizing to new molecules is the classical task in drug discovery. First, we put aside the samples associated to 10% of the ligands. Then, we conducted a 10-fold cross-validation on the remainder of the data. All splits were stratified by the number of samples as well as the mean pIC50 per ligand.

Kinase split. With this setting, we wanted to assess the model's ability to predict binding affinities for unseen kinases. Like in the ligand split, we first put aside 10% of the kinases and then conducted a 10-fold cross-validation on the remainder. Again, all splits were stratified by the number of samples as well as the mean pIC50 per ligand.

Pretraining. The 485,461 non-kinase samples were split into train/test at a 90/10 ratio and this data was then used in one configuration of the BiMCA model for pretraining.

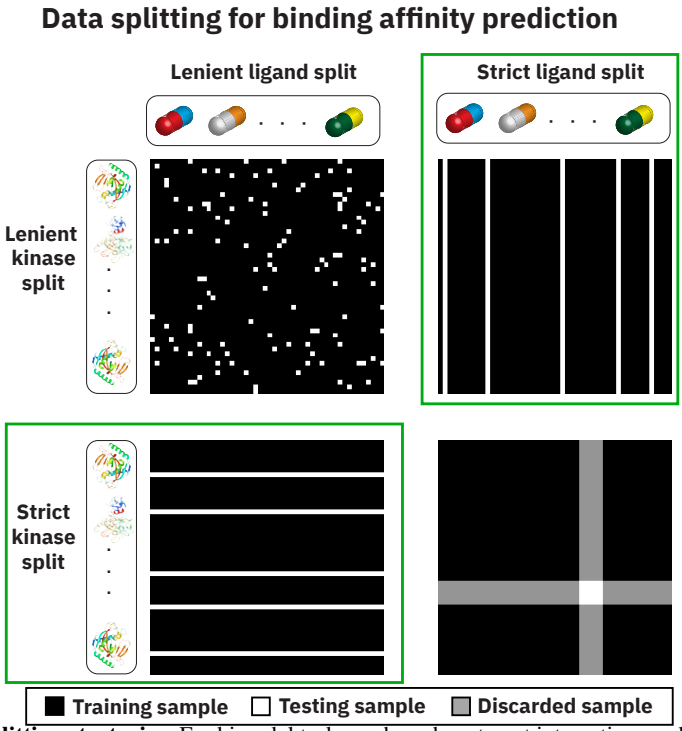


Figure A3: **Data splitting strategies.** For bimodal tasks such as drug-target interaction prediction, four splitting strategies are possible. In this work, a strict ligand split and a strict kinase split (colored in green) were explored.

A1.6 Complementary results for kinase split

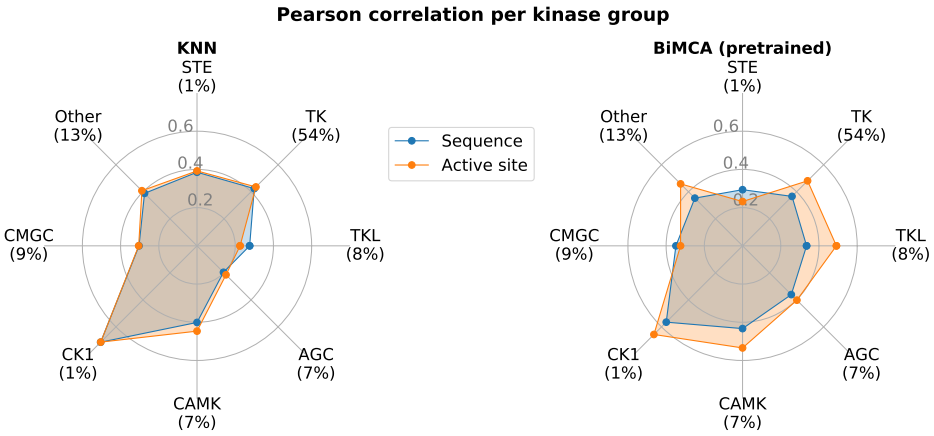


Figure A4: **Performance in predicting affinity for unseen kinases according to the kinase group.** For the KNN (left) and the pretrained BiMCA (right) the PCC of all samples of respective kinase group is shown. Kinases that could not be classified with the catalogue from [24] are grouped into *Other*.

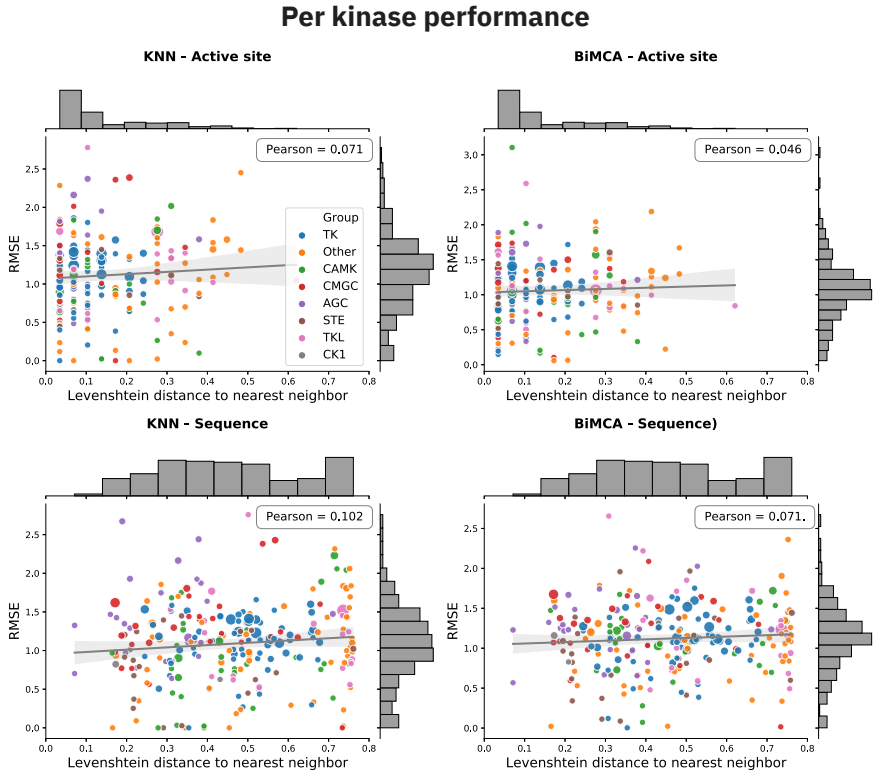


Figure A5: **Dependency of model performance on similarity to nearest neighbor in training data** In none of the four model configurations, a strong dependency/correlation between the performance on a specific kinase and the distance to the nearest neighbor in training data was found. Measures obtained considering results on validation data.

Table A1: RMSE (on pIC50) on validation and test data (kinase split).

Data	Config	KNN	BiMCA	BiMCA (pretrained)
Val.	Full seq.	1.34±0.16	1.38±0.08	1.30±0.13
	Active site	**1.32**±0.17	**1.28**±0.13	**1.21**±0.13
Test	Full seq.	1.56±0.09	1.44±0.04	1.32±0.04
	Active site	**1.52**±0.10	**1.33**±0.04	**1.25**±0.05

Table A2: Pearson correlation coefficient on validation and test data (kinase split).

Data	Config	KNN	BiMCA	BiMCA (pretrained)
Val.	Full seq.	0.41±0.09	0.32±0.05	0.39±0.08
	Active site	**0.42**±0.11	**0.46**±0.08	**0.49**±0.07
Test.	Full seq.	0.23±0.05	0.32±0.03	0.43±0.03
	Active site	**0.28**±0.06	**0.44**±0.04	**0.49**±0.05

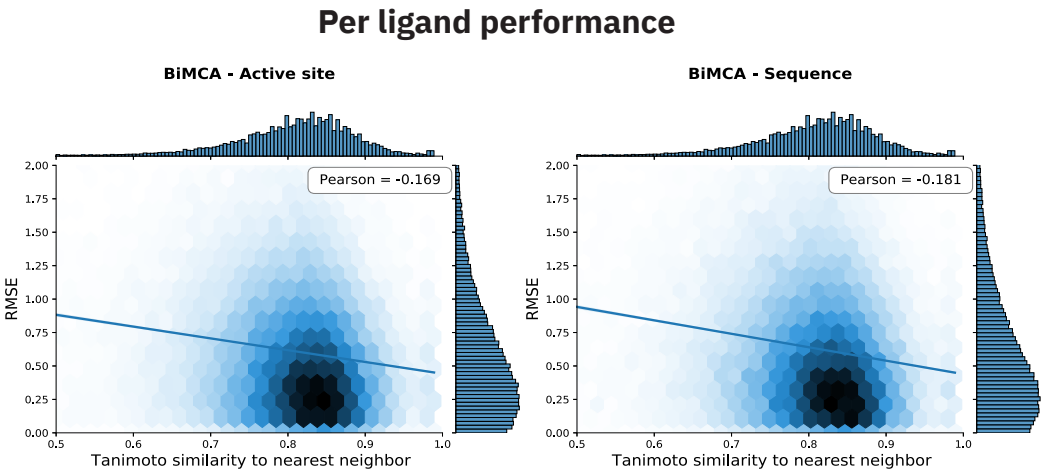


Figure A6: **Dependency of affinity prediction on similar ligands.** For each ligand, the performance is shown as a function to the Tanimoto similarity to the nearest training ligand. Measures computed on validation data.

Table A3: RMSE (on pIC50) on validation and test data (ligand split).

Data	Config	KNN	BiMCA	BiMCA (pretrained)
Val.	Full seq.	0.78±0.01	0.91±0.01	0.85±0.01
	Active site	**0.77**±0.01	**0.83**±0.01	**0.82**±0.01
Test	Full seq.	**0.76**±0.00	0.91±0.01	0.86±0.01
	Active site	0.77±0.00	**0.83**±0.01	**0.82**±0.01

Table A4: Pearson correlation coefficient on validation and test data (ligand split).

Data	Config	KNN	BiMCA	BiMCA (pretrained)
Val.	Full seq.	**0.83**±0.01	0.75±0.00	0.78±0.01
	Active site	**0.83**±0.01	**0.79**±0.00	**0.80**±0.01
Test.	Full seq.	**0.83**±0.01	0.74±0.00	0.77±0.01
	Active site	**0.83**±0.01	**0.79**±0.00	**0.80**±0.01

Table A5: Evaluation on IDG-DREAM dataset [7]. PCC values are reported.

Model	Config	All	Known kin.	Unknown kin.	Round 1	Round 2
KNN	Full seq.	0.224	0.242	0.032	0.132	0.32
	Active site	0.244	0.282	-0.141	0.145	0.344
BiMCA	Full seq.	0.16	0.169	0.064	0.102	0.185
	Active site	**0.32**	**0.327**	**0.238**	**0.179**	**0.412**

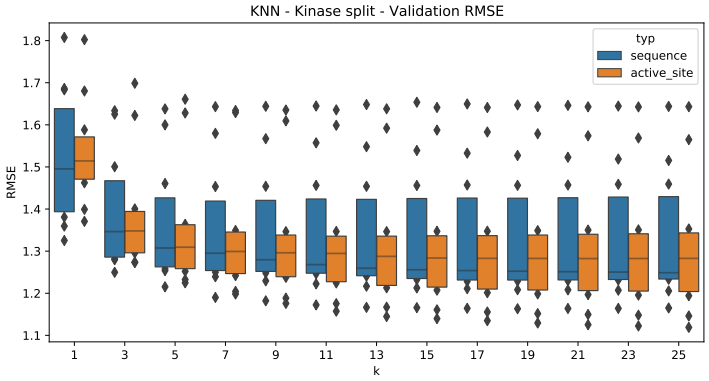
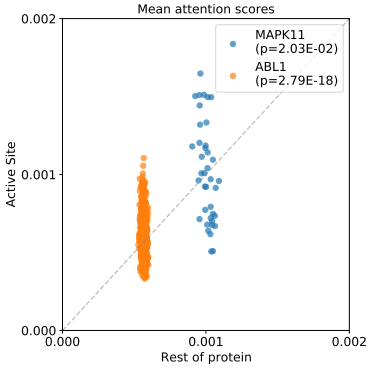


Figure A7: **Kinase split validation performance for different *k*.** Based on this plot, we fixed k to the lowest RMSE on this dataset ($k = 13$) and used the same k for all results throughout the paper.

A1.7 Complementary results for ligand split



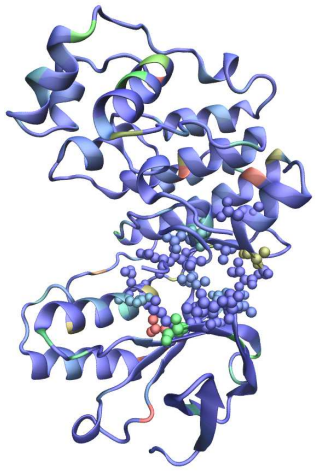


Figure A8: **Kinase attention scores.** *Left*: For each kinase-ligand pair of MAPK11 and ABL1, the mean attention scores on active site residues versus the remaining residues is shown. *Right*: Exemplary visualization of attention values overlayed on the MAPK11 structure highlighting atoms with high weight (blue means low, green medium and red high attention). Residues depicted as spheres belong to the active site.

Table A6: Generalization to new protein families based on fixed-split BindingDB dataset from DeepAffinity [18]. RMSE and Pearson correlation (PCC) for each model and protein family. Best performances are shown in bold. DeepAffinity models refer to unified RNN-CNN and RNN/GCNN-CNN models. All models below the single line are ours. The three last models above that line are ensembles which can hardly be directly compared to our models. Numbers from other works taken from their manuscripts since the split is fixed. DeepCDA did not report RMSE. The last columns report the average across the four tasks.

Model	ER		Ion Channel		RTK		GPCR		All	
	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
DeepAffinity SMILES [18]	1.53	0.16	1.34	0.17	1.24	0.39	1.40	0.24	1.38	0.24
DeepAffinity Graph [18]	1.68	0.05	1.43	0.10	1.74	0.01	1.63	0.04	1.62	0.05
DeepCDA [1]	-	0.10	-	0.31	-	0.42	-	0.28	-	0.28
Truong [37] (ECFP/Pfam)	1.74	0.19	1.32	0.27	1.27	0.43	1.49	0.22	1.46	0.28
DeepAffinity Ensemble [18]	1.46	0.30	1.30	0.18	1.23	0.42	1.36	0.30	1.34	0.30
MLP ensemble [37]	1.51	0.24	1.36	0.19	1.26	0.42	1.36	0.33	1.37	0.29
Transformer ensemble [37]	1.61	0.39	1.34	0.38	**1.14**	**0.47**	1.29	0.33	1.35	**0.39**
NN (k=1)	1.53	0.30	1.80	0.07	1.51	0.32	1.81	0.17	1.66	0.22
KNN (k=4)	1.36	0.30	1.52	0.11	1.31	0.37	1.50	0.20	1.42	0.25
KNN (k=13)	1.28	0.40	1.43	0.13	1.26	0.36	1.43	0.17	1.35	0.27
KNN (k=25)	**1.27**	**0.43**	1.41	0.13	1.25	0.34	1.42	0.15	1.33	0.26
BiMCA (full seq.)	1.35	0.32	**1.19**	**0.41**	1.38	0.40	**1.25**	**0.42**	**1.27**	**0.39**