Active site sequence representation of human kinases outperforms full sequence for affinity prediction and inhibitor generation: 3D effects in a 1D model

Jannis Born,^{*,†,‡} Tien Huynh,[¶] Astrid Stroobants,[§] Wendy D. Cornell,[¶] and Matteo Manica^{*,†}

[†]*IBM Research Europe, Zurich, Switzerland*

[‡]*Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*

[¶]*IBM Research, Yorktown Heights, New York, USA*

[§]*Department of Chemistry, Imperial College London, London, United Kingdom*

E-mail: jab@zurich.ibm.com; tte@zurich.ibm.com

Abstract

Recent advances in deep learning have enabled the development of large-scale multimodal models for virtual screening and de novo molecular design. The human kinome with its abundant sequence and inhibitor data presents an attractive opportunity to develop proteochemometric models that exploit the size and internal diversity of this family of targets. Here we challenge a standard practice in sequence-based affinity prediction models: instead of leveraging the full primary structure of proteins, each target is represented by a sequence of 29 residues defining the ATP binding site. In kinase-ligand binding affinity 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, datasets, performance metrics and holds true when predicting affinity for both unseen ligands and kinases. 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. We then investigate a de novo molecular design task and find that the active site provides benefits in the computational efficiency, but otherwise, both kinase representations yield similar optimized affinities (for both SMILES and SELFIES-based molecular generators). Our work challenges the assumption that full primary structure is indispensable for modelling human kinases. We hope that these results will inspire additional investigation into hybrid mechanistic-DL modeling approaches to support the identification and optimization of kinase inhibitors’ candidates.

Introduction

Protein kinases are ubiquitous for cell life and have become a vital source of targets for drug discovery after the FDA approval of the first kinase inhibitor, imatinib, 20 years ago¹⁻³. By 2021, about 60 kinase inhibitors have received market approval⁴ and helped to enrich our treatment options for cancer and, more recently, also neurodegenerative or viral diseases⁴. Despite this success, most research has focused on fractions of the kinome revealing therapeutic utility^{5,6}. The characteristics of the target family that led drug discovery researchers to avoid kinases for many years, i.e., binding site similarity and sheer number of family members³, make the family an ideal candidate for proteochemometric approaches which exploit this similarity systematically.

Proteochemometric modelling

Computational methods have supported our understanding of kinases and their inhibitors in many regards, including inhibitor selectivity⁷, identification of binding subpockets⁸ or promiscuity maps⁹, defining the kinome conformational space¹⁰ or, most typically, virtual screening such as compound protein interaction (CPI) prediction^{11–14} and drug response prediction^{15,16}. With the rise of deep learning (DL) in chemoinformatics and drug discovery¹⁷ proteochemometric models for bioactivity prediction can now be trained large-scale^{18,19}. Early approaches to kinase affinity prediction were single-assay¹¹, or single-target models¹². However, multi-target models enable to leverage larger data sources and bring along the benefit that cross-target information can be learned. These models are usually superior to single-task models²⁰, especially if training data is sparse²¹ or tasks are correlated²². However, they often do not incorporate protein descriptors and thus naturally lack the ability to predict activity for novel targets^{20,23}.

Instead, proteochemometric models that consider both chemical and protein information represent the most generic option. In contrast to multi-target models they can, in principle, generalize to novel ligands and targets simultaneously and typically rely only on ligand structure information (such as SMILES²⁴) and primary structure information for the target. Unlike single- or multi-task models which lack the inductive bias to learn the meaning of an interaction, the bimodal nature of proteochemometric models enables them to learn, for example, inter-molecular non-covalent interactions²⁵. With the growing availability of high-throughput screening data²⁶ the gap between traditional docking and DL methods shrank^{27,28}, partly thanks to hybrid approaches^{29,30}. A popular trend in predicting drug-target interactions exploits topological information by using graph neural networks or applying 3D convolutions to the binding site^{31–33} or by operating directly on secondary or tertiary structure^{34–37}. However, to avoid the need for structure information, most recent work relied exclusively on primary structure information^{25,38,39}. This research started in 2016 by means of binary fingerprints for proteins and ligands⁴⁰ but has advanced to more

interpretable inputs like amino acid (AA) or SMILES sequences^{19,38,39,41–43}.

The success of these models is maybe best exemplified in the recently conducted IDG-DREAM challenge about Drug-Kinase Binding Prediction¹⁸ where the top performing method out of 99 submissions in the second round was a multimodal deep neural network based on SMILES and AA sequences. This model is highly similar to our previously proposed affinity prediction model⁴⁴ which we further explore herein. To the best of our knowledge, no existing work has systematically compared the use of full protein sequences to active site in 1D proteochemometric models. Here, we aim to fill this gap.

De novo molecular design

Deep learning is facing an increasing adoption in the discovery of new molecules and materials^{45–47}. A prominent example is the landmark study about discovery of DDR1 kinase inhibitors⁴⁸. In that work, deep learning methods discovered potential DDR1 inhibitors of which two were found active *in-vitro* and one even *in-vivo*⁴⁸. The most common type of deep generative model for molecular design are variational autoencoders (VAE)⁴⁹ which can be seen as a global search method in the chemical space. VAEs have been coupled with Bayesian optimization (BO) principles by means of Gaussian processes (GP) to optimize chemocentric properties like drug-likeness^{50,51}. While GP optimization allows to maximize computationally costly functions efficiently, no previous work has, according to our knowledge, incorporated an additional modality (like proteins) into the evaluation function and thus optimized a more complex, biochemical property.

Our contribution

In this work, we systematically compare the impact of using active site residues and full sequence information to represent kinases for two common tasks related to kinase inhibitors: proteochemometric modeling of drug-kinase binding affinity prediction and de novo molecular design of kinase inhibitor candidates (for overview see Figure 1).

Figure 1: **Overview of the comparison between the considered kinase representations.** Primary structure representations of full sequence and active sites for human kinases are evaluated on two proteochemometric modeling tasks: binding affinity prediction to ligands, and generation of potentially novel ligands for the kinase of interest.

We utilize a bimodal deep neural network, the BiMCA model, which dispenses with traditional descriptors and relies solely on interpretable, textual inputs (SMILES and AA sequences). Moreover, we propose a simple, yet efficient and novel KNN regression model for protein-ligand affinity based on Levenshtein distance⁵² of proteins that yields competitive results. We further show how Gaussian processes (GP) can be used on deep molecular generative models to optimize the binding affinity of the generated molecules for specific kinases. This analysis is performed in light of comparing sequence and active-site based affinity predictors.

Methods

Protein ligand affinity prediction

Proteochemometric models that combine information from proteins and ligands are an established technique to approach a classic task in kinase inhibitor modeling: predicting binding affinities for pairs of proteins and ligands. To address this task, we utilize two methods; a bi-

modal neural network based on convolutional and attention layers from our previous work⁴⁴ and a novel KNN regression model based on protein and molecule similarity measures.

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 $\Phi_A : \mathcal{P} \times \mathcal{M} \rightarrow \mathcal{A}$. 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 regression problem, we first chose the k -nearest neighbor (KNN) algorithm from the realm of traditional machine learning methods due to its simplicity and ease of interpretation. Notably, the nearest neighbors are computed in a joint space spanned by protein and ligand similarity. KNNs were previously applied for affinity prediction with a bimodal similarity measure based on numerical descriptors⁵³, however, here we represent kinases by their primary structure (either full sequence or only active site) and molecules by their ECFP4 fingerprint⁵⁴ with a radius of 2 and 512 bits. As a distance metric between samples we utilize a combination of the length-normalized Levenshtein distance for the primary structure and the Tanimoto similarity⁵⁵ of molecules. This method is an adaptation of the KNN model proposed by Weber et al.⁵⁶ for protein-protein interaction prediction. 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

$$\mathbf{D}(p_i, m_i, p_j, m_j) = \frac{Lev(p_i, p_j)}{\max(|p_i|, |p_j|)} + (1 - \mathcal{T}(m_i, m_j)) \quad (1)$$

where $|\cdot|$ denotes sequence length, $\mathcal{T}$ is the Tanimoto similarity measure and $Lev(\cdot, \cdot)$ is the Levenshtein distance⁵²; a string-based distance measure that counts the number of single-AA changes required to transform one sequence into the other. Both the length-normalized Levenshtein distance and the Tanimoto similarity of Equation 1 are bound to $[0, 1]$ and the subtraction converts the similarity into a distance, such that $\mathbf{D}(\cdot, \cdot, \cdot, \cdot) \in [0, 1]$. Then, the prediction $\hat{a}_j$ is trivially computed by $\hat{a}_j = \frac{\sum_i^k a_i}{k}$ with $a_i \in \mathcal{D}_k$. 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.

Bimodal multiscale convolutional attention (BiMCA) network

As an alternative method, we utilize a bimodal neural network that learns a function $\Phi_A : \mathcal{P} \times \mathcal{M} \rightarrow \mathcal{A}$ based on primary structure of proteins and SMILES²⁴ sequences of molecules. As visualized in Figure 2, this model separately ingests an amino acid sequence and a SMILES sequence, converts the tokens into embedding vectors (which are learned during training) and performs 1D convolutions to aggregate local substructures.

Figure 2: The bimodal multiscale convolutional attention model (BiMCA).

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 Bahdanau et al.⁵⁷ and was proposed in our previous work^{44,58}. 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:

$$\alpha_i = \frac{\exp(u_i)}{\sum_j \exp(u_j)}, \text{ where } \vec{u} = \tanh(\mathbf{X}_1 \mathbf{W}_1 + \mathbf{W}_3(\mathbf{X}_2 \mathbf{W}_2)) \vec{v} \quad (2)$$

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 $\vec{v} \in \mathbb{R}^A$ are learnable parameters.

The output of this model is a scalar value interpreted as pIC50 affinity score for the provided protein-ligand pair. Flavors of this model have been used successfully for cancer drug sensitivity prediction^{58,59} and toxicity prediction⁶⁰. The variant described here is identical to the binding affinity descriptor used in Born et al.⁴⁴ for predicting antiviral activity of potential SARS-CoV-2 inhibitors.

Binding affinity data

Kinase data. We curated compound-protein interaction data from BindingDB⁶¹. From the 2,222,074 entries of the database as on 22.04.2021, $\sim 800,000$ 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 3

for an overview of the dataset’s statistics.

Figure 3: **Visualization of kinase inhibitor data in BindingDB⁶¹.** **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 screened against almost all 349 kinases.

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 (see below).

Human kinase sequence alignment

The binding site residues for each kinase were identified by applying the binding site definition of protein kinase A from Sheridan’s kinase selectivity study¹⁴ to a recently published structurally-validated multiple sequence alignment of 497 human protein kinase domains from Modi and Dunbrack⁶². Sheridan’s definition identified 29 residues representing the ATP binding site including but not limited to contributions from the Gly-rich-loop, gate-

keeper, hinge, and DFG-in-out. While the sequence alignment was done for 497 kinases, only 349 were included in the simulations because of lacking activity data in BindingDB.

Data splitting

For proteochemometric models there are four different splitting strategies (see Figure A1). 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.

Hyperparameters and model training

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 A2).

BiMCA. The SMILES sequences of all ligands were padded to a length of 696. 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 used four parallel 1D convolutional layers with kernel sizes of 3, 5, 11 for the ligands and 3, 11 and 25 on the proteins. The number of filter kernels was 32 respectively for the full sequence model and 128 for the active site model on both modalities. This was done to partly accommodate that the full sequence model had substantially more parameters than the active site model. This stemmed from the context attention layer which requires $\mathcal{O}(nm)$ parameters where n and m are the sequence length of proteins and ligands respectively (for details see Born et al.⁴⁴). 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. All models were implemented in PyTorch⁶³ and used the `pytoda` package⁶⁴ for data handling and preprocessing. The BiMCA model optimized a MSE loss with Adam⁶⁵ 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.

De novo molecular generation

In this task, the goal is to generate novel molecules with high predicted binding affinities against a target kinase of interest. To that end, we are building upon a molecular generative model developed in our previous work^{44,64}. This generative model is then explored in a novel scheme, using Bayesian optimization with Gaussian processes (GP) that optimizes the generative model to yield molecules with higher affinity. The GP approximates the predicted affinity of the protein (kinase) of interest and a molecule.

Generative model. Our generative model is implemented using a variational autoencoder (VAE) pretrained on ~ 1.5 million bioactive compounds from ChEMBL⁶⁶. The model consists of two layers of stack-augmented GRUs in both encoder and decoder. We trained two versions of this model, one using SMILES²⁴ sequences, the other one using SELFIES⁶⁷, a novel molecular representation that was devised for deep generative models. The SMILES

based model is identical to the one used in Born et al.⁶⁴, the SELFIES model was used in Born et al.⁴⁴. These models are trained to optimize the standard VAE objective⁶⁸:

$$\mathcal{L}_{VAE}(\theta, \phi) = \mathbb{E}_{q_{\theta}(\mathbf{z}|\mathbf{x})}[\log p_{\phi}(\mathbf{x}|\mathbf{z})] - D_{KL}(q_{\theta}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (3)$$

where D_{KL} denotes the Kullback-Leibler divergence. During training, each sample defines an encoding distribution $q_{\theta}(\mathbf{z}|\mathbf{x})$ and this distribution is constrained to be similar to a pre-defined prior distribution $p(\mathbf{z})$, in our case $q_{\theta}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\vec{0}, \mathbf{I})$, i.e., the latent code is modelled using a multivariate unit Gaussian following standard VAE formulation.⁶⁸ Upon training these models, we can sample from the latent distribution $p(\mathbf{z})$ and use the decoder network $p_{\phi}(\mathbf{x}|\mathbf{z})$ to produce a set of SMILES/SELFIES sequences.

Bayesian Optimization with Gaussian Processes. The rationale of performing Bayesian Optimization (BO) with a Gaussian Process (GP) is to facilitate the *efficient* exploration of the chemical space learned by the generative model with the objective of maximizing or minimizing an arbitrary function acting on the latent space points.⁵⁰ Herein, we adopt BO to maximize binding affinity for a kinase of interest. Given a protein p of interest as well as our molecular generator $p_{\phi}(\mathbf{x}|\mathbf{z})$ from above, the goal is to find the latent code $\hat{\mathbf{z}}$ that maximizes our affinity prediction $\Phi_A : \mathcal{P} \times M \rightarrow A$. Considering the sheer size of the latent space $\mathcal{Z}$ and especially the cost to evaluate the function Φ_A ¹ motivates us to formulate the problem in terms of Bayesian optimization: $\hat{\mathbf{z}} = \arg \max_{\mathbf{z} \in \mathcal{Z}} [\Phi_A(p, p_{\phi}(\mathbf{x}|\mathbf{z}))]$. BO adopts an iterative search with the objective to minimize the number of calls to Φ_A before we can guarantee that our chosen point in the latent space $\hat{\mathbf{z}}'$ yields an affinity a such that: $|a - a_{max}| < \varepsilon$. In BO, the function being subject to optimization is modelled with a prior specifying a probability distribution over functions, in our case a GP prior: $\Phi_A \sim GP[\hat{m}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')]]$ where $\hat{m}$ is a mean function and k is a kernel returning similarity between

¹either a costly forward pass through the BiMCA or an exhaustive computation of sample similarities with the KNN

two points. Therefore, during optimization, the affinity prediction function is assumed to follow a multivariate Gaussian. The BO algorithm implementation in this work relies on the negative expected improvement⁶⁹ as acquisition function, which trades off exploration and exploitation to determine the next query point. For each kinase, the optimization process was initiated from 40 initial (random) points in the latent space and the optimization was performed for 30 epochs with 50 calls per epoch with the `scikit-optimize`⁷⁰ package. After each epoch, 300 molecules were generated from the latent points (invalid SMILES were discarded).

Results

Protein ligand affinity prediction

Kinase data split

The kinase split is the ideal configuration to test the impact of the protein representation (active site vs. full sequence). When predicting affinity for an unseen kinase, the utilized representation is critical, both in the KNN model (based on the Levenshtein distance of kinases) and the BiMCA model (based on non-linear extrapolations). This task is significantly more challenging than splitting on ligands because the shape of the binding pocket largely governs the binding activity⁷¹. While this task is less commonly investigated than the ligand split, it is highly relevant since it was shown that binding affinity predictions are mostly based on ligand rather than interaction features³⁹, a phenomenon called *hidden ligand and bias*⁷². The results of a 10-fold cross validation of all three model types (KNN, BiMCA, BiMCA pretrained) can be found in Figure 4 and show a consistent and strong superiority of the active site models.

Figure 4: **Binding affinity prediction results on kinase split.** The left and right column show respectively the performance of all three models on the validation and test data. On both metrics, RMSE (upper row) and Pearson correlation coefficient (PCC, lower row) the active site configurations significantly outperform the full sequence configuration, irrespective of the utilized model. To see the exact numerical scores, please see Table A1 and Table A2.

On the validation data, the RMSE is reduced by 1.2%, 7.5% and 6.9% for the KNN, the BiMCA and the pretrained BiMCA respectively when comparing full sequence to active site models. 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 premodels only have 5% of the parameters of the full sequence model. Due to the heterogeneity of the data, the results are less consistent on the ten validation folds compared to the held-out test data. Now on the test data (see Figure 4B and D), the full sequence models achieve an average RMSE of 1.56, 1.44 and 1.31 compared to 1.52, 1.33 and 1.25 for the active site. For all three model configurations, the active site models significantly outperform the full sequence models ($p < 0.01$, Wilcoxon signed-rank test, $W+$). Moreover, the BiMCA models outperforms the KNN model by a large margin. 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. After 20 epochs of pretraining, this model achieved a RMSE of 0.86 ($r = 0.82$) on the non-kinase data. Notably, on the kinase split, all pretrained BiMCA models outperform the regular ones, demonstrating that inferring general patterns of protein-ligand interactions can massively benefit the development of proteochemometric models for kinase-inhibitor affinity prediction. Interestingly, the active site even outperformed the full sequence model although both models were pretrained on full protein sequences (no active site information was available for pretraining).

Kinase groups. We further investigated the performance for the eight different groups of conventional protein kinases (ePK) based on the classification by Hanks and Hunter⁷³. Using the catalogue from Manning et al.⁷⁴ that contains ~ 600 kinases, we mapped all kinases to their respective group. For all kinase groups, the PCC is shown in Figure 5 for the KNN and pretrained BiMCA model. With a few exceptions, the plot indicates the superiority of the active site configuration consistently across the kinase families.

Figure 5: **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 Manning et al.⁷⁴ are grouped into *Other*.

Notably, the most heavily screened kinase group is TK (Tyrosine kinases) which make up the majority of kinase-related samples in BindingDB (54%). TKs phosphorylate tyrosine residues and are thoroughly researched due to their significant role in cancer and the successful development of highly selective TK inhibitors such as imatinib, gefitinib or erlotinib. 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. Let us have a closer look at the TKL (Tyrosine Kinase-Like) group and try to explain why the results do not resonate with the remaining findings indicating superiority of the active site. The first observation is that for models based on sequence similarity (like the KNN), full protein structure is superior to active site alone. We presume this is because many TKL kinases (e.g., all RAF kinases⁷⁵) have multiple binding sites which are not captured in the active site sequence alone. The second observation is that the KNN overall performs poorly on TKL and that the performance gain for the BiMCA compared to the KNN is highest for the TKL group. We suspect that the KNN performs poorly on TKL as it is the most heterogeneous group of kinases⁷⁴. While the KNN predicts purely based on sequence similarity, the BiMCA can capture non-linear relations and accordingly, the performance gain for the BiMCA compared to the KNN is highest for the TKL group. Moreover, also in relation to the first observation, only the BiMCA can leverage information from distant samples that have more complex relations to the kinase of interest and thus the active site BiMCA configuration achieves the best performance in predicting affinity for unseen TKL kinases. Another remarkable finding is the good result for proteins from the CK1 (cell kinase 1) group. We suppose this to be due to the high intra-group and low-inter group similarity of CK1 kinases. CK1s are highly conserved sequences, very similar to each other but very distinct from other kinase groups⁷⁶ and form a distinct branch in the kinome tree⁷⁴ (note that samples were split by kinases, not by kinase groups).

Few others have investigated the performance of CPI models for unseen kinases. In a comparable work³⁸, a late fusion model ensemble achieved a PCC of 0.42 on tyrosine kinases.

Instead, our active site BiMCA model achieved a PCC of 0.48. However, this comparison is limited because 1) we did not leave out entire protein families, 2) we refrained from boosting performance with an ensemble and 3) the underlying set of used BindingDB samples might not be identical.

Similarity analysis. A reasonable suspicion in the kinase split is that the prediction performance hinges on the availability of similar kinases in the training data. For both protein representations and both, the KNN and the pretrained BiMCA model, we therefore investigated the per-kinase performance as a function of the similarity to the nearest neighbor in the training data (cf. Figure 6).

Figure 6: **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.

Overall, it can be safely excluded that our models require data from similar kinases to work well. While all PCCs are positive, none of them exceed values of 0.11. Critically, the active site models are more robust in that regard than the sequence models and indeed, the best model (BiMCA, active site) has the lowest dependence of all models. Moreover, the KNN has a stronger dependence on similar samples than the BiMCA.

Ligand data split

This split is the classical setting of kinase inhibitor discovery; based on some affinity data for a kinase of interest, the model should reveal the potential of a chemical to inhibit this kinase. While this task is easier than the kinase-split, we notice that it is substantially harder than a lenient split (where only pairs are left out and both, protein and ligand have been seen by the model). The results of the 10-fold CV as well as on the test dataset are shown in Table 1 (RMSE) and Table 2 (Pearson correlation). Like in the ligand split, all BiMCA

Table 1: 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 2: 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

models using active site information are superior to the ones using full primary structure (8.2% and 4.7% RMSE improvement for the BiMCA and pretrained BiMCA respectively). 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 ($p < 0.001$, $W+$). Based on the tables, it appears that the KNN model performed similarly well on

both active sites and full sequences. This observation is explained by the fact that the protein information is of negligible performance for our KNN model *in a ligand split*. When retrieving the $k = 13$ nearest neighbors according to Equation 1, the first addend (which measures protein similarity) will collapse to 0 for all samples of the same kinase. This occurs irrespective of whether active site or full sequence information is used and therefore dilutes differences between the representations. As the average number of samples per protein in the dataset is 593 (see histogram in Figure 3C), it is not surprising that for the active site and full sequence indeed in 98.9% and 99.3% of the predicted samples, the nearest neighbor is a sample with the same kinase. Moreover, the length-normalized Levenshtein distance is a more discriminative protein similarity measure in case of unequal sequence lengths. To remedy the described confound and compare the impact of the two representations for the KNN on the ligand split, we evaluated the performance exclusively on the remaining samples. In alignment with the overall findings of this manuscript, 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). Moreover, on this subset of samples the active site BiMCA model surpasses its KNN equivalent by a large margin (RMSE = 1.18, Pearson’s r = 0.64). This indicates that the KNN model strikes at interpolation, but falls behind the BiMCA in extrapolation; a hypothesis that is corroborated by an increased correlation of the prediction error with the distance to the nearest neighbor (KNN: r = 0.23, BiMCA: r = 0.18; active site models, validation data).

Kinase inhibitor classes. To investigate the performance of the model for different groups of kinase inhibitors we retrieved the primary target for each kinase inhibitor (as annotated in BindingDB) and grouped the ligands into thirteen groups of alleged mechanism of action based on the classification scheme by Roskoski Jr⁷⁷. From a total of around 372k validation samples, about a third could be automatically assigned to a kinase inhibitor class. Figure 7 shows the PCC of both models and both configurations for each kinase

inhibitor class.

Figure 7: **Performance in predicting affinity for unseen kinase inhibitors according to their primary protein target.** For the KNN (left) and the pretrained BiMCA (right) the PCC of all samples of respective kinase inhibitor class is shown.

In the right plot, we can see that, with the exception of MEK inhibitors, the active site model performed better on all thirteen kinase inhibitor groups. For the KNN model (left), the predictions were extremely correlated between sequence and active site model (see above) and therefore, there are only negligible differences between both representations. However, apparent across both models is that affinity prediction for MEK (i.e., MAPK/ERK) inhibitors is consistently higher if full sequence information is used. Since our sequence alignment only relied on ATP binding site residues⁶², we hypothesize that this is due to the successful discovery of several ATP-noncompetitive MEK inhibitors that bind to a unique site near the ATP binding pocket⁷⁸. 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 (MKNK2, MKNK1, MAPKAP2, MAPK3, MAP2K1, MAPK1, MAPK14, MAP3K5).

Similarity analysis. Generalization of proteochemometric models to distant manifolds of the chemical space is challenging, but critical to screen large virtual libraries. Equivalent

to the kinase split, it might be suspected that the model performance on the ligand split depends on whether similar molecules were available during training. However, both models exhibited 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 8).

Figure 8: **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.

Like in the kinase split, the active site model does not only outperform the sequence model but is also less dependent on the availability of similar samples during training. The KNN showed a slightly stronger negative correlation ($PCC = -0.23$, not shown).

Model attention analysis. It might be argued that learning on the active site alone (i.e., a subset of the full protein sequence) should in any case be avoided because it inevitably prevents the model from incorporating the entire protein information in its predictions. This, however, implicitly introduces an additional task for the model, namely to filter out irrelevant residues. Given that this task is performed without supervision and on a finite dataset, we cannot expect a sequence model to learn how to filter uninformative residues to a satisfying degree²⁵. A key advantage of the utilized BiMCA model lies in the contextual attention mechanism^{44,58} which enables us to evaluate its capacity to focus on relevant residues. 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 two exemplary kinases, MAPK11 and ABL1 we performed an interpretability analysis to assess whether the sequence model paid disproportionately high attention to the active site residues (see Figure 9). For both kinases, the mean attention scores on the active site residues are significantly

Figure 9: **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).

higher than on the remaining residues ($\alpha = 0.05\%$, *MWU*). While this is an encouraging finding that resonates with previous research³⁸ and emphasizes the model’s ability to focus on the relevant residues, Figure 9 (left) also illustrates that this trend was not consistent for all samples which may partly explain the inferior performance.

Validation on external test dataset

To evaluate the performance of our model on an independent test dataset, we utilized the data released in June 2021 as part of the IDG-DREAM challenge.¹⁸ The challenge focused on under-studied parts of the human kinome to catalogue the unexplored target space of kinase inhibitors. Thus, it resembles a particularly challenging dataset, encompassed by

825 samples (cf. Supplementary Data 1 by Cichońska et al.¹⁸). After restricting ourselves to kinases for which full sequence and active site information⁶² was available, 720 samples remained, 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 $10\mu\text{M}$ 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 pIC_{50} in BindingDB. For the KNN model we used all data available in BindingDB as training data whereas for the BiMCA we build an ensemble of the 10 models from the ligand split.

The results on this dataset are in alignment with our overall findings (cf. Table 3). The

Table 3: Evaluation on external dataset by Cichońska et al.¹⁸. 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

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 kinases from Cichońska et al.¹⁸ that were not included in BindingDB. Direct comparison with the results reported in the IDG-DREAM challenge is not possible due to the aforementioned differences to our training data.

De novo molecular generation

The goal of this task is to generate de novo molecules with high predicted binding affinities against a target kinase of interest. Our approach exploits BO using Gaussian processes, an established technique to optimize molecular generative models for chemocentric proper-

ties^{50,51}. We selected six kinases for further analysis; the four JAK family members, and ABL1 (the target the first FDA approved kinase inhibitor, of imatinib³) as well as MAPK11 (P38 β) a thoroughly studied target from the MAPK family and isoform of MAPK14/P38 α ⁷⁹. Starting from a random point in the learned chemical space (i.e., the latent space of our generative model), we applied GP-BO to optimize in different setups both the KNN and the BiMCA affinity predictions, either operating on the active site residues or on the full sequence. For the different configurations, we performed the optimization process on both the SMILES and the SELFIES generators.

A tangible difference was the increased runtime for the full sequence model. In general, utilizing the active site predictor can save significant computational resources because the average sequence length is smaller, less parameters are needed for the models and thus, training and, especially, inference speed are higher (cf. Table 4). Notably, the active sites

Table 4: Comparison of active site and full sequence models.

Model	Config	Sequence length	Model size	Training time	Inference time
KNN	Full seq.	742 $\pm$ 369	–	–	67ms (CPU)
	Active site	29 $\pm$ 0	–	–	50ms (CPU)
BiMCA	Full seq.	742 $\pm$ 369	14.2M	43h	18ms (GPU)
	Active site	29 $\pm$ 0	0.6M	41h	16ms (GPU)

consisted of only 29 residues making up 4% of the full sequences. Consequently, the BiMCA model size was reduced by a factor of 25 while still exhibiting superior performance in affinity prediction.

The results from the optimization of the molecular generative model toward high affinity to ABL1 and MAPK11 indicate that utilizing active site information can slightly accelerate the generation of generated molecules with high affinity (Figure 10).

Figure 10: **Bayesian optimization of latent space sampling with Gaussian processes.** For two selected targets, the (predicted) pIC50 distribution of generated molecules is shown. Dashed lines denote the mean pIC50 before the optimization. Molecules were either sampled from the SELFIES generator (top row) or SMILES generator (bottom row). Affinity was predicted with the KNN model.

The distributions indicate that the optimization led to higher average affinity scores compared to the baseline in all cases. While these results seem to indicate that active site sequences yield better results than full sequences when subject to affinity optimization with GP, the remaining results for the four JAK family members (see supplementary material, Figure A3) are inconclusive regarding a comparison of active site and sequence models (cf. Table 5). This is not surprising because the kinase representation does not directly impact the generative process but is only used in the affinity evaluation.

However, advantages of the lower runtime of the active site model (25% faster) include the faster convergence of the GP optimization as well as the increased number of sampled ligands with high predicted affinity in any time interval (cf. Table 6). In relation to the results of previous works on GP optimization of sampled molecules^{50,51}, we emphasize that optimizing

Table 5: **Results of GP optimization.** The average pIC50 across the six kinases and the molecules generated through the optimization process is shown.

Ligand repr.	Kinase repr.	Baseline	Optimized
SELFIES	Full seq.	6.55 $\pm$ 0.6	6.6 $\pm$ 0.5
	Active site	6.51 $\pm$ 0.6	6.59 $\pm$ 0.5
SMILES	Full seq.	6.51 $\pm$ 0.6	6.57 $\pm$ 0.6
	Active site	6.57 $\pm$ 0.6	6.6 $\pm$ 0.5

Table 6: **Runtime comparison in sampling effective ligands.** All ligands with a predicted IC50 < 100nM (i.e., pIC50 > 7) are considered effective.

Evaluation	Active site	Sequence
Time until 5% pIC50 improvement (min.)	14$\pm$8	21 $\pm$ 8
Number of effective ligands in 25mins	35$\pm$19	30 $\pm$ 16

target affinity is more challenging compared to chemocentric properties like drug-likeness or synthetic accessibility scores.

Qualitative evaluation of molecules

To assess the *de novo* kinase inhibitors proposed by the generative model in more depth, we show the ligand with the highest predicted affinity scores for each of the six targets and both kinase representations in the top row of Figure 11.

Figure 11: **Best generated molecules per target.** For each of the six targets, the molecule with the lowest predicted affinity is shown in the top two rows. The bottom row shows the distribution of measured pIC50 values (as reported in BindingDB) for the most similar kinase inhibitors of a subset of the generated molecules.

The molecules are versatile in their structure and have predicted affinities in the range of 2 – 50nM. We then selected for each kinase the top 5 most effective, yet aromatic molecules and retrieved the 10 most similar compounds measured for that target from BindingDB (based on Tanimoto similarity⁵⁵ of ECFP6 fingerprints⁵⁴). The bottom row of Figure 11 then displays the distribution of these measured affinity scores. A remarkable observation is that for the four JAK targets the BindingDB samples selected from the active-site molecules have a higher affinity than the molecules generated against the full sequence.

Discussion

In this paper, we have investigated two computational tasks related to understanding and advancing kinase inhibitors: proteochemometric modelling of protein ligand affinity prediction and *de novo* molecular design of kinase inhibitors. Both tasks were examined in light of two kinase representations, using either full primary sequence or only the active site residues. Regarding the affinity prediction, our results suggest a superiority of active site residues when

predicting affinity for novel ligands (ligand split) or kinases (kinase split). Moreover, these results are robust across two investigated models (a KNN regressor and a multimodal deep neural network) and were confirmed on a novel dataset from the IDG-DREAM challenge.

This is an important, and maybe surprising finding because the active site residues are a subset of the full primary sequence. Indeed, it codes for the active site structure as well as additional, more distant determinants of binding and dynamics. It seems that providing exclusively the active site residues increases the signal-to-noise ratio in the sequences, consequently leading to better performance. This hypothesis is partly corroborated by our attention analysis. Even without supervision on the importance of the residues, the sequence model learns implicitly to focus to a significantly higher extent on the active site residues (cf. Figure 9). An interpretation of this finding is that the BiMCA might implicitly learn some elements of tertiary structure (an area with much recent success for deep learning models⁸⁰) from the sequence information alone. Karimi et al.³⁸ have also found that CPI prediction models can implicitly learn to assign higher attention scores to binding sites compared to remaining residues. However, it was later shown by Li et al.²⁵ that the exact localization of binding sites requires explicit supervision and cannot be learned in a self-supervised manner by analyzing the attention scores on sequential or secondary structures as suggested in some case studies^{34,35,38}. Contrarily,⁸¹ demonstrated that in Transformer-based methods, which are now ubiquitous in protein language modeling, attention targets binding sites, captures the folding structure and learns amino acid representations that are consistent with the substitution matrix.

Another important finding is that the active site models even outperformed the full sequence models when both models were pretrained on full sequences. While the difference between both configurations is lower in this setting, this result suggests that proteochemoetric models benefit from pretraining on large-scale pan-protein data even if the final use case is limited to one family.

We then showcase an application of the aforementioned predictive models, trained on

both representations, to *de novo* molecular design via Bayesian optimization. Using two generative models based on SMILES and SELFIES respectively, we find that the active site representation, with its parsimonious description of the structure, speeds up the generation of binding ligands without sacrificing the performance in terms of affinity reported for the closest ligand reported in BindingDB to the generated molecules.

Outlook

Future research should validate our findings on the superiority of only using active site residues on more specific datasets, especially given the high discrepancy between random and "realistic" test sets of kinase inhibitors that can lead to performance gaps in practice¹³. However, these concerns are scant for proteochemometric models where each protein-ligand pair informs every prediction. Moreover, we emphasize that our results on affinity prediction were consistent across two methods (KNN and BiMCA), two splits (kinase and ligand) and two datasets (BindingDB and IDG-DREAM). This is especially relevant since BindingDB is 1) a very robust database with highly heterogeneous samples (composed of a multitude of assays and largely assembled from external sources) and 2) the largest public database for kinase screenings (with >200,000 samples⁶¹). Moreover, future work could explore hybrid approaches between full sequence and active site information. This could be achieved either by enriching the learning of the attention mask with pairwise non-covalent compound residue interaction in a semi-supervised setting based on the dataset by Li et al.²⁵ or simply provide a fixed attention mask to guarantee that the model predominantly (but not exclusively) focuses on the active site residues. If active site information is unavailable, potential protein-ligand binding residues could be extracted with existing computational methods⁸².

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

Declaration of interests

The authors declare no competing interest.

Funding

The authors received no external funding.

Author contributions

The manuscript has been initially drafted by J.B. with contributions from all other authors. The final version has been revisioned multiple times by J.B., M.M. and W.C. M.M. implemented the code to preprocess kinase data from BindingDB as well as the alignment data produced by W.C. and T.H. J.B. and M.M. implemented the code for model training and running the generation of small molecules using both AA sequence and active site representations. A.S. implemented the code for small molecule generation using Gaussian Processes that has been then adapted by M.M. J.B. run all the model trainings and computational experiments that are presented in the study. J.B. produced all the visualisations and analysis of the performance of the predictive and generative models. M.M. and T.H. worked at the production of the graphical abstract and the 3D protein visualizations. M.M. and W.C. conceived the study.

References

- (1) Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nature reviews Drug discovery* **2002**, *1*, 309–315.
- (2) Cohen, P.; Alessi, D. R. Kinase drug discovery—what’s next in the field? *ACS chemical biology* **2013**, *8*, 96–104.
- (3) Cohen, P.; Cross, D.; Janne, P. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery* **2021**,
- (4) Laufer, S.; Briner, K.; Bajorath, J.; Georg, G. I.; Wang, S. New Horizons in Drug Discovery-Understanding and Advancing Kinase Inhibitors: Special Issue and Call for Papers. 2020.
- (5) Fedorov, O.; Müller, S.; Knapp, S. The (un) targeted cancer kinome. *Nature chemical biology* **2010**, *6*, 166–169.
- (6) Elkins, J. M.; Fedele, V.; Szklarz, M.; Azeez, K. R. A.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X.-P.; Roth, B. L., et al. Comprehensive characterization of the published kinase inhibitor set. *Nature biotechnology* **2016**, *34*, 95–103.
- (7) Lo, Y.-C.; Liu, T.; Morrissey, K. M.; Kakiuchi-Kiyota, S.; Johnson, A. R.; Broccatelli, F.; Zhong, Y.; Joshi, A.; Altman, R. B. Computational analysis of kinase inhibitor selectivity using structural knowledge. *Bioinformatics* **2019**, *35*, 235–242.
- (8) Volkamer, A.; Eid, S.; Turk, S.; Rippmann, F.; Fulle, S. Identification and visualization of kinase-specific subpockets. *Journal of chemical information and modeling* **2016**, *56*, 335–346.
- (9) Blaschke, T.; Miljkovic, F.; Bajorath, J. Prediction of different classes of promiscuous and nonpromiscuous compounds using machine learning and nearest neighbor analysis. *ACS Omega* **2019**, *4*, 6883–6890.

- (10) Ung, P. M.-U.; Rahman, R.; Schlessinger, A. Redefining the protein kinase conformational space with machine learning. *Cell chemical biology* **2018**, *25*, 916–924.
- (11) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. 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* **2011**, *51*, 1942–1956.
- (12) Martin, E.; Mukherjee, P. Kinase-kernel models: accurate in silico screening of 4 million compounds across the entire human kinome. *Journal of chemical information and modeling* **2012**, *52*, 156–170.
- (13) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds. *Journal of chemical information and modeling* **2017**, *57*, 2077–2088.
- (14) Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. 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* **2009**, *49*, 1974–1985.
- (15) Huang, L.-C.; Yeung, W.; Wang, Y.; Cheng, H.; Venkat, A.; Li, S.; Ma, P.; Rasheed, K.; Kannan, N. Quantitative Structure–Mutation–Activity Relationship Tests (QSMART) model for protein kinase inhibitor response prediction. *BMC bioinformatics* **2020**, *21*, 1–22.
- (16) Koras, K.; Kizling, E.; Juraeva, D.; Staub, E.; Szczurek, E. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *bioRxiv* **2021**,
- (17) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today* **2018**,

- (18) Cichońska, A.; Ravikumar, B.; Allaway, R. J.; Wan, F.; Park, S.; Isayev, O.; Li, S.; Mason, M.; Lamb, A.; Tanoli, Z., et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature communications* **2021**, *12*, 1–18.
- (19) Gaspar, H.; Ahmed, M.; Edlich, T.; Fabian, B.; Varszegi, Z.; Segler, M.; Meyers, J.; Fiscato, M. Proteochemometric Models Using Multiple Sequence Alignments and a Subword Segmented Masked Language Model. *ChemRxiv preprint (10.26434/chemrxiv.14604720.v1)* **2021**,
- (20) Rodriguez-Perez, R.; Bajorath, J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. *Acs Omega* **2019**, *4*, 4367–4375.
- (21) Rodríguez-Pérez, R.; Bajorath, J. Prediction of compound profiling matrices, part II: relative performance of multitask deep learning and random forest classification on the basis of varying amounts of training data. *ACS omega* **2018**, *3*, 12033–12040.
- (22) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is multitask deep learning practical for pharma? *Journal of chemical information and modeling* **2017**, *57*, 2068–2076.
- (23) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science* **2018**, *9*, 5441–5451.
- (24) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.
- (25) Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems* **2020**, *10*, 308–322.

- (26) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W., et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery* **2011**, *10*, 188–195.
- (27) Zhang, H.; Liao, L.; Saravanan, K. M.; Yin, P.; Wei, Y. DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. *PeerJ* **2019**, *7*, e7362.
- (28) Bitencourt-Ferreira, G.; de Azevedo, W. F. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophysical chemistry* **2018**, *240*, 63–69.
- (29) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E. Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *Journal of Chemical Information and Modeling* **2021**, *61*, 1583–1592.
- (30) Bao, J.; He, X.; Zhang, J. Z. DeepBSP—a Machine Learning Method for Accurate Prediction of Protein–Ligand Docking Structures. *Journal of Chemical Information and Modeling* **2021**,
- (31) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling* **2017**, *57*, 942–957.
- (32) Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *Journal of chemical information and modeling* **2020**, *60*, 2791–2802.
- (33) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: protein–ligand

- absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018**, *58*, 287–296.
- (34) Gao, K. Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. Interpretable Drug Target Prediction Using Deep Neural Representation. *IJCAI*. 2018; pp 3371–3377.
- (35) Tsubaki, M.; Tomii, K.; Sese, J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **2019**, *35*, 309–318.
- (36) Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of chemical information and modeling* **2019**, *59*, 3981–3988.
- (37) Torng, W.; Altman, R. B. Graph convolutional neural networks for predicting drug–target interactions. *Journal of chemical information and modeling* **2019**, *59*, 4131–4149.
- (38) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329–3338.
- (39) Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **2020**, *36*, 4406–4414.
- (40) Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting compound-protein interaction prediction by deep learning. *Methods* **2016**, *110*, 64–72.
- (41) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829.

- (42) Gonczarek, A.; Tomczak, J. M.; Zareba, S.; Kaczmar, J.; Dąbrowski, P.; Walczak, M. J. Interaction prediction in structure-based virtual screening using deep learning. *Computers in biology and medicine* **2018**, *100*, 253–258.
- (43) Zhao, L.; Wang, J.; Pang, L.; Liu, Y.; Zhang, J. GANsDTA: Predicting drug-target binding affinity using GANs. *Frontiers in genetics* **2020**, *10*, 1243.
- (44) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakaraman, N.; Cardinale, A.; Laino, T.; Rodríguez Martínez, M. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Machine Learning: Science and Technology* **2021**, *2*, 025024.
- (45) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering* **2019**, *4*, 828–849.
- (46) Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future medicinal chemistry* **2019**, *11*, 567–597.
- (47) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The advent of generative chemistry. *ACS Medicinal Chemistry Letters* **2020**, *11*, 1496–1505.
- (48) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A., et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **2019**, *37*, 1038–1040.
- (49) Born, J.; Manica, M. Trends in Deep Learning for Property-driven Drug Design. *Current Medicinal Chemistry (in press)* **2021**,
- (50) Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M., et al.

- Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268–276.
- (51) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science* **2020**, *11*, 577–586.
- (52) Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. 1966; pp 707–710.
- (53) Nazarshodeh, E.; Sheikhpour, R.; Gharaghani, S.; Sarram, M. A novel proteochemometrics model for predicting the inhibition of nine carbonic anhydrase isoforms based on supervised Laplacian score and k-nearest neighbour regression. *SAR and QSAR in Environmental Research* **2018**, *29*, 419–437.
- (54) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.
- (55) Tanimoto, T. T. Elementary mathematical theory of classification and prediction. **1958**,
- (56) Weber, A.; Born, J.; Rodriguez Martinez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **2021**, *37*, i237–i244.
- (57) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015.
- (58) Manica, M.; Oskoei, A.; Born, J.; Subramanian, V.; Saez-Rodriguez, J.; Rodriguez Martinez, M. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular pharmaceutics* **2019**, *16*, 4797–4806.

- (59) Cadow, J.; Born, J.; Manica, M.; Oskooei, A.; Rodríguez Martínez, M. PaccMann: a web service for interpretable anticancer compound sensitivity prediction. *Nucleic acids research* **2020**, *48*, W502–W508.
- (60) Markert, G.; Born, J.; Manica, M.; Schneider, G.; Rodriguez Martinez, M. 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**,
- (61) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **2016**, *44*, D1045–D1053.
- (62) Modi, V.; Dunbrack, R. L. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Scientific reports* **2019**, *9*, 1–16.
- (63) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*, 8026–8037.
- (64) Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Markert, G.; Martínez, M. R. PaccMann^{RL}: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **2021**, *24*, 102269.
- (65) Kingma, D. P.; Ba, J. L. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.
- (66) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J., et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Research* **2013**, *42*, D1083–D1090.

- (67) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024.
- (68) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR. 2014.
- (69) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization* **1998**, *13*, 455–492.
- (70) Head, T.; Kumar, M.; Nahrstaedt, H.; Louppe, G.; Shcherbatyi, I. scikit-optimize/scikit-optimize. 2020; <https://doi.org/10.5281/zenodo.4014775>.
- (71) Wang, H.; Qiu, J.; Liu, H.; Xu, Y.; Jia, Y.; Zhao, Y. HKPocket: human kinase pocket database for drug design. *BMC bioinformatics* **2019**, *20*, 1–11.
- (72) Sieg, J.; Flachsenberg, F.; Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *Journal of chemical information and modeling* **2019**, *59*, 947–961.
- (73) Hanks, S. K.; Hunter, T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1. *The FASEB journal* **1995**, *9*, 576–596.
- (74) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1934.
- (75) Fischer, A.; Baljuls, A.; Reinders, J.; Nekhoroshkova, E.; Sibilski, C.; Metz, R.; Albert, S.; Rajalingam, K.; Hekman, M.; Rapp, U. R. 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* **2009**, *284*, 3183–3194.
- (76) Knippschild, U.; Krüger, M.; Richter, J.; Xu, P.; García-Reyes, B.; Peifer, C.;

- Halekotte, J.; Bakulev, V.; Bischof, J. The CK1 family: contribution to cellular stress response and its role in carcinogenesis. *Frontiers in oncology* **2014**, *4*, 96.
- (77) Roskoski Jr, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological research* **2016**, *103*, 26–48.
- (78) Wu, P.-K.; Park, J.-I. MEK1/2 inhibitors: molecular activity and resistance mechanisms. *Seminars in oncology*. 2015; pp 849–862.
- (79) Dhillon, A. S.; Hagan, S.; Rath, O.; Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **2007**, *26*, 3279–3290.
- (80) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.
- (81) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. BERTology Meets Biology: Interpreting Attention in Protein Language Models. 9th International Conference on Learning Representations, ICLR 2021. 2021.
- (82) Cui, Y.; Dong, Q.; Hong, D.; Wang, X. Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC bioinformatics* **2019**, *20*, 1–12.

Appendices

Data splitting strategies

Figure A1: **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.

Complementary results for affinity prediction

Figure A2: **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.re

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

Complementary results for de novo molecular generation

Figure A3: **Bayesian optimization of latent space sampling with Gaussian processes.** For all four JAK family members, the (predicted) pIC50 distribution of generated molecules is shown. Dashed lines denote the mean pIC50 before the optimization. In keeping with the remainder of this paper, orange indicates the active site and blue the full sequence models. Molecules were either sampled from the SELFIES generator (top row) or SMILES generator (bottom row).