JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

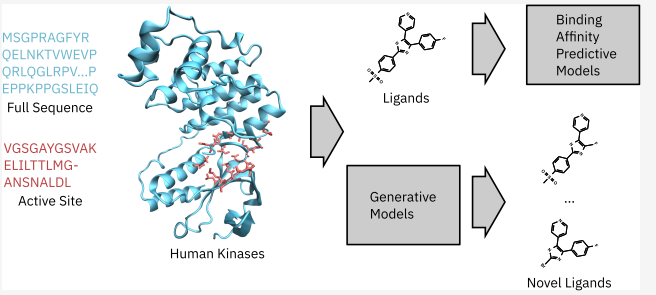
Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction and Inhibitor Generation: 3D Effects in a 1D Model

Jannis Born,* Tien Huynh, Astrid Stroobants, Wendy D. Cornell, and Matteo Manica*

Cite This: J. Chem. Inf. Model. 2022, 62, 240−257 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 discontiguous 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, data sets, and performance metrics and holds true when predicting pIC_{50} for both unseen ligands and kinases. Our interpretability analysis reveals a potential explanation for the superiority of the active site models: whereas only mild statistical effects about the extraction of three-dimensional (3D) interaction sites take place in the full sequence models, the active site models are equipped with an implicit but strong inductive bias about the 3D structure stemming from the discontiguity of the active sites. Moreover, in direct comparisons, our models perform similarly or better than previous state-of-the-art approaches in affinity prediction. 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 the full primary structure is indispensable for modeling human kinases.



INTRODUCTION

Protein kinases are ubiquitous for cell life and have become a vital source of targets for drug discovery after the Food and Drug Administration (FDA) approval of the first kinase inhibitor, imatinib, 20 years ago.[1−3] By 2021, about 60 kinase inhibitors have received market approval[4] and helped us to enrich our treatment options for cancer and, more recently, also neurodegenerative or viral diseases.[4] Despite this success, most research has focused on narrow 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 (e.g., binding site similarity and sheer size[3]) make the family an ideal candidate to exploit this similarity systematically with proteochemometric approaches.

Proteochemometric Modeling. Proteochemometric modeling is concerned with using machine learning methods on a combination of protein and ligand descriptors.[7] It differs from standard quantitative structure−activity relationship (QSAR) approaches, which typically assume a single (or multiple) target(s), but only featurize the ligands and not the proteins. Computational methods have supported our understanding of kinases and their inhibitors in many regards, including inhibitor selectivity,[8] identification of binding subpockets[9] or promiscuity maps,[10] and defining the kinome conformational space[11] or, most typically, virtual screening such as compound−protein interaction (CPI) prediction[12−15] and drug response prediction.[16,17] With the advent of deep learning (DL) in chemoinformatics and drug discovery,[18] proteochemometric models for bioactivity prediction can now be trained large scale.[19,20] Early approaches to kinase affinity prediction were single-assay[12] or single-target models,[13] i.e., models that only considered ligand descriptors and were trained exclusively on data from one protein target or assay type. However, multitarget models[a] enable us to leverage larger data sources and bring along the benefit that cross-target

Received: July 23, 2021
Published: December 14, 2021

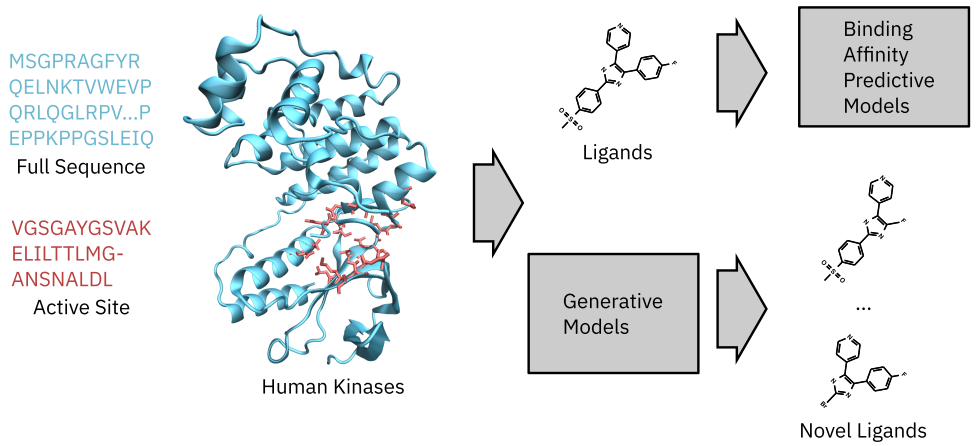


Figure 1. Overview of the comparison between the considered kinase representations. Primary structure representations of full sequence and active sites for human kinases are evaluated on two proteochemometric modeling tasks: compound−kinase interaction prediction and generation of potentially novel ligands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

information can be learned.[21] These models are usually superior to single-task models,[22] especially if training data are sparse[23] or tasks are correlated.[24] However, they often do not incorporate protein descriptors and thus naturally lack the ability to predict activity for novel targets.[22,25]

Instead, proteochemometric models that consider both chemical and protein information represent the most generic option. In contrast to multitarget models, they can, in principle, generalize to novel ligands and targets simultaneously and typically rely only on ligand structure information (such as SMILES[26]) and primary structure information for the target. Unlike single- or multitask 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, intermolecular noncovalent interactions.[27] With the growing availability of high-throughput screening data,[28] the gap between traditional docking and DL methods shrank,[29,30] partly thanks to hybrid approaches.[31,32] A popular trend in predicting drug−target interactions exploits topological information using graph neural networks or applying three-dimensional (3D) convolutions to the binding site[33−35] or by operating directly on the secondary or tertiary structure.[36−39] However, to avoid the need for structure information, the most recent work relied exclusively on primary structure information.[27,40,41] This line of research was started in 2016 by means of binary fingerprints for proteins and ligands[42] but has advanced to more interpretable inputs like amino acid (AA) or SMILES sequences.[20,40,41,43−45]

The success of these models is may be best exemplified in the recently conducted IDG-DREAM challenge about drug−kinase binding prediction,[19] 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 model,[46] which we further explore herein. This so-called bimodal multiscale convolutional attention (BiMCA) model for CPI prediction was successfully utilized to learn generic drug−target interactions from BindingDB and to steer a molecular generative model toward generating molecules that potentially bind to SARS-CoV-2-related protein targets.[46] Moreover, a predecessor of this model called PaccMann set a new state-of-the-art in drug sensitivity prediction on the GDSC data set[47] and has been deployed in a public webservice[48] that is used widely in the community.[49,50] To the best of our knowledge, no existing work has systematically compared one-dimensional (1D) proteochemometric models that use full protein sequences to those that use active site sequences alone. 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.[51−53] A prominent example is the landmark study about the discovery of DDR1 kinase inhibitors.[54] In that work, deep learning methods discovered potential DDR1 inhibitors of which two were found active in vitro and one even in vivo.[54] The most common type of deep generative model for molecular design is variational autoencoders (VAEs),[55] 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 (GPs) to optimize chemocentric properties such as drug-likeness.[56,57] While GP optimization allows us to maximize computationally costly functions efficiently, no previous work has, according to the best of 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 interaction prediction (assessed with experimentally measured pIC_{50}) and de novo molecular design of kinase inhibitor candidates (for an overview, see Figure 1).

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 *K*-nearest neighbor (KNN) regression model for CPI based on Levenshtein distance[58] of proteins that yields competitive results. These two models were selected in light of the goal of comparing active site and full sequence representations for modeling kinase inhibitors. Since hypercomplex models with countless nonlinearities could dilute the effects present in the data itself, we opted for one simplistic and ante hoc interpretable model (the KNN) and a state-of-the-art deep neural network (the BiMCA). We further show how Gaussian processes (GPs) can be used on deep molecular

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

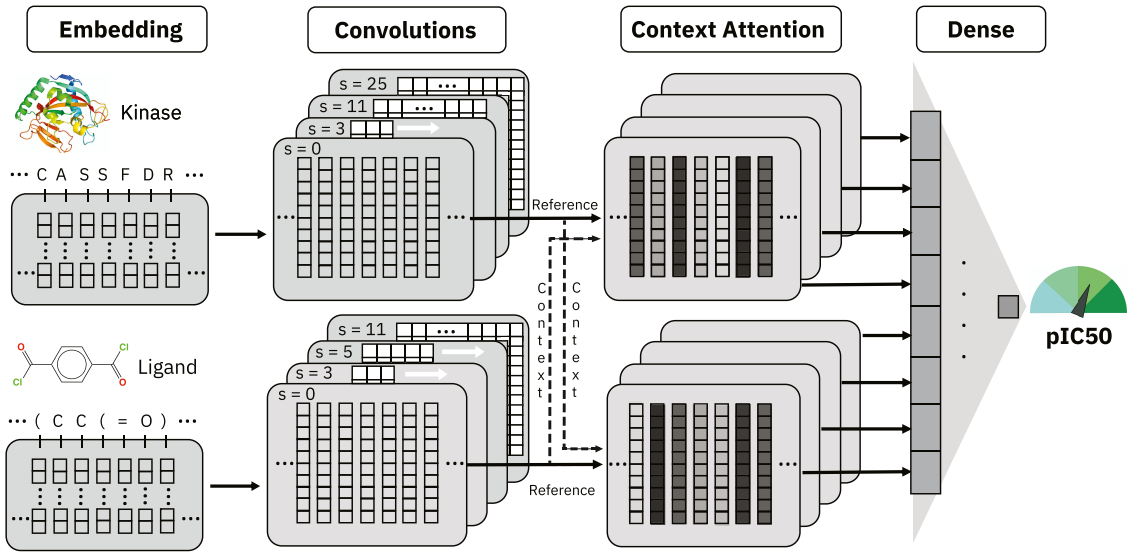


Figure 2. Bimodal multiscale convolutional attention model (BiMCA). Both kinases and ligands are represented as text sequences of amino acids and SMILES, respectively. The BiMCA uses learned embeddings and then applies convolutions of multiple kernel sizes c on the embedding matrices (hence the words "multiscale convolutional"). Afterward, the context attention (CA) 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, and white arrows show the direction of the convolution sliding. Figure adjusted from Weber et al.[62]

generative models to optimize the predicted binding affinity of the generated molecules for specific kinases. This analysis is performed in light of comparing sequence- and active site-based affinity predictors and validated against randomly sampled molecules from a deep molecular generative model.

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 bimodal neural network based on convolutional and attention layers from our previous work[46] and a novel KNN regression model based on protein and molecule similarity measures.

Problem Formulation. Let $\mathcal{P}$ denote the space of proteins, $\mathcal{M}$ denote the molecular space, and $\mathcal{A}$ denote 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 pIC_{50}.

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 similarities. KNNs were previously applied for affinity prediction with a bimodal similarity measure based on numerical descriptors;[59] however, here we represent kinases by their primary structure (either full sequence or only active site) and molecules by their ECFP4 fingerprint[60] 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 similarity[61] of molecules. This method is an adaptation of the KNN model proposed by Weber et al.[62] for protein−protein interaction prediction.

More formally, let $\{p_j, m_j\}$ denote an unseen sample from the test data set $\mathcal{D}_{\text{Test}} = \{p_i, m_i\}_{i=1}^{N_{\text{Test}}}$. 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

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

where $|\cdot|$ denotes sequence length, $\mathcal{T}$ is the Tanimoto similarity measure, and $\text{Lev}(\cdot,\cdot)$ is the Levenshtein distance,[58] 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 eq 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 data set size (N = 206 989 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 data set.

Bimodal Multiscale Convolutional Attention (BiMCA) Network. As an alternative method, we utilize a bimodal neural network that learns a function $\Phi_A: \mathcal{P} \times \mathcal{M} \to \mathcal{A}$ based on the primary structure of proteins and SMILES[26] sequences of molecules. As visualized in Figure 2, this model separately ingests an amino acid sequence and a SMILES sequence and first converts the tokens into embedding vectors, which are learned during training. In an ablation study, we compare the impact of learned embeddings to one-hot encodings and the

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

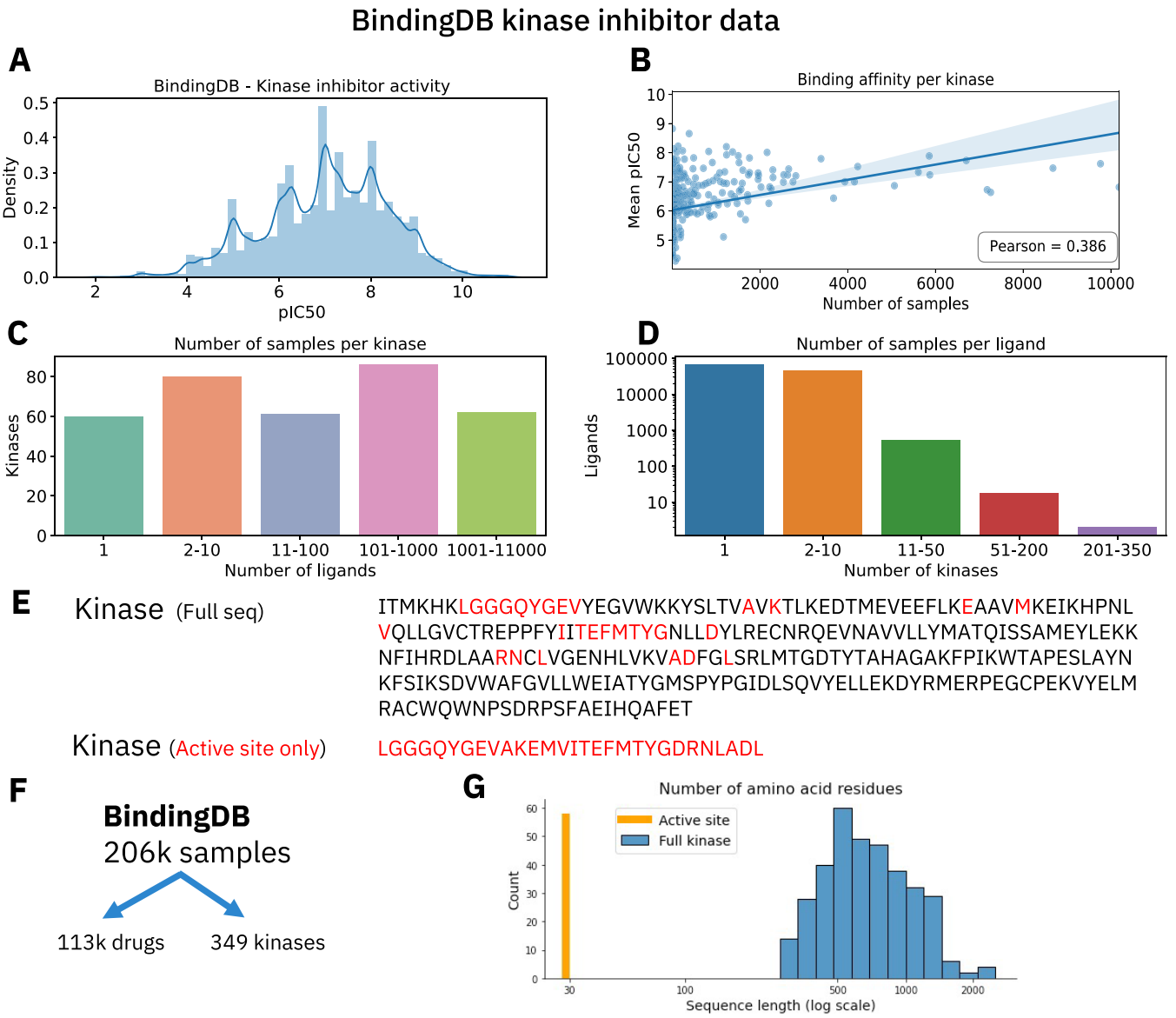


Figure 3. Visualization of kinase inhibitor data in BindingDB.[66] (A) Distribution of pIC_{50} scores in database ($N = 206\,989$). (B) Kinases with more samples tend to be more promiscuous. (C) Histogram of a number of data points for each kinase. (D) Most ligands are screened on less than a dozen of kinases but some are screened against almost all 349 kinases. (E) Full primary structure and active site sequence of the ABL1 kinase. (F) Division of 206k BindingDB samples into unique kinases and inhibitors. (G) Distribution of the sequence length of full sequences and active sites (log-scale).

(standardized) BLOSUM62 matrix,[63] a representation that encodes each AA based on the evolutionary similarity to all other AAs.

It then performs convolutions to aggregate local substructures. Specifically, we use three parallel 1D convolutional layers with kernel sizes of $s \times e$, where e is the embedding size (32 and 8 for SMILES and AAs, respectively) and s is the kernel size (3, 5, and 11 for the ligands and 3, 11, and 25 for 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 Bahdanau et al.[64] and was proposed in our previous work.[46,47] In detail, we employ eight context attention (CA) layers, four in the protein stream and four in the ligand stream (cf. Figure 2). Within both streams, the inputs are identical but either the kinase or the ligand is used as a reference over which the attention scores are computed. Each of those layers consumes the convolved SMILES and AA sequences and performs a series of linear and nonlinear operations before assigning the attention scores $\alpha_i \in [0, 1]$. The output of the largest convolution kernels of the kinase stream (size 25) is coupled with the output of the largest kernels of the ligand stream (size 11) and so on. In addition to the three convolutional layers, we employ a skip connection that omits the convolutions ($s = 0$ in Figure 2). The attention scores are computed as

$$\alpha_i = \frac{\exp(u_i)}{\sum_j^T \exp(u_j)}, \quad \text{where}$$

$$\vec{\mathbf{u}} = \tanh(\mathbf{X}_1\mathbf{W}_1 + \mathbf{W}_3(\mathbf{X}_2\mathbf{W}_2))\vec{\mathbf{v}} \tag{2}$$

We call $\mathbf{X}_1 \in \mathbb{R}^{T_i \times C}$ the reference input, where $T_1 \in \{T_M, T_P\}$ is the sequence length and C is the number of convolutional

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

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. For a visualization of the context attention layer, see Figure S1. The output of this model is a scalar value interpreted as pIC_{50} score for the provided protein−ligand pair. Flavors of this model have been used successfully for cancer drug sensitivity prediction[47,48] and toxicity prediction.[65] The variant described here is identical to the CPI model used in Born et al.[46] for predicting the antiviral activity of potential SARS-CoV-2 inhibitors.

Binding Data. Kinase Data. We curated compound−protein interaction data from BindingDB.[66] BindingDB is a large but heterogeneous source of protein−ligand binding affinity data that is continuously updated. The data was curated from public sources (e.g., publications), and the reported binding values were thus obtained under highly diverging experimental procedures. From the 2 222 074 entries of the database as of 22.04.2021, ∼800 000 were retained after removing missing values and duplicates. Afterward, samples with molecules whose SMILES strings were invalid or longer than 696 tokens, i.e., atoms and/or bonds, were removed. Following the previous work,[27,40,43] we chose half-maximal inhibitory concentration (IC_{50}) as the binding affinity metric, converted all values to pIC_{50} (i.e., the negative decimal logarithm of the half-maximal inhibitory concentration), and clipped all values to the interval [2, 11] (1 mM to 0.01 nM). While we emphasize that the experimentally obtained half-maximal inhibitory concentration (IC_{50}) does not necessarily directly reflect the ligand's binding affinity strength, the K_d and IC_{50} values are highly correlated in BindingDB. Lastly, we filtered out all samples where the target proteins are not kinases. This resulted in 206 989 samples distributed across 113 475 ligands (mean pIC_{50} per ligand: 7.1 ± 1.2) and 349 human kinases (mean pIC_{50} per kinase: 6.2 ± 0.9). See Figure 3 for an overview of the data set's statistics.

For example, a notable and strong bias in the data set is that kinases screened against more ligands tend to have a higher-average pIC_{50} ($r = 0.39$).

Nonkinase Data. The remainder of the above data (i.e., all nonkinome 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 study[15] to a recently published structurally validated multiple sequence alignment (MSA) of 497 human protein kinase domains from Modi and Dunbrack.[67] 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. Using the 29 residues in the PKA sequence as a reference, the amino acids at those positions in each other kinase were then extracted from the kinase family alignment by Modi and Dunbrack.[67] The resulting active site sequences are more than an order of magnitude shorter than the original full sequences (cf. Figure 3G) and are usually discontiguous in the original sequence (cf. Figure 3E). 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 S2). Here, we focus on two of these regimes, namely, splitting samples based on ligands (while not controlling for proteins) as well as the reverse task.

Ligand Split. Building a predictive model that can generalize to new molecules is the classical task in drug discovery. First, we put aside the samples associated with 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 pIC_{50} 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 pIC_{50} per ligand.

Pretraining. The 485 461 nonkinase 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.

DeepAffinity Data Sets. To facilitate direct comparison with previous work, we conducted additional experiments on the BindingDB data sets, as processed and split by DeepAffinity.[40] This data set comes with 263 583 training samples, 113 168 test samples (lenient split; cf. Figure S2), and four held-out sets to test the generalization to entirely unseen protein families, namely, nuclear estrogen receptors (ER; 34 318 samples), ion channels (14 599 samples), receptor tyrosine kinases (RTK; 34 318 samples), and G-protein-coupled receptors (GPCRs; 60 238 samples). In the absence of active site information, these experiments were conducted only using full protein sequences. In one additional setting, we compared active site and full sequence representation for generalization to the RTK family. The surrogate training set with available active sites contained 50 668 instead of 263 584 samples, and from the RTK test data set, we retrieved active sites for 92% of all samples (active sites for 56/127 kinases could not be found because the data set contained nonhuman RTKs, s.t., the sequence alignment from Modi and Dunbrack[67] could not be applied).

Hyperparameters and Model Training. The KNN model was evaluated on all $k \leq 25$. For all results, we choose a value of $k = 13$ as this led to the lowest root-mean-square error (RMSE) on the validation data set on the ligand split (see Figure S3).

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 dimensionalities 32 and 8. The number of filter kernels was 32 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 $O(nm)$ parameters, where n and m are the sequence length of proteins and ligands, respectively (for details, see Born et al.[46]). 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 `PyTorch`[68] and used the `pytoda` package[69] for data handling and preprocessing. The

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

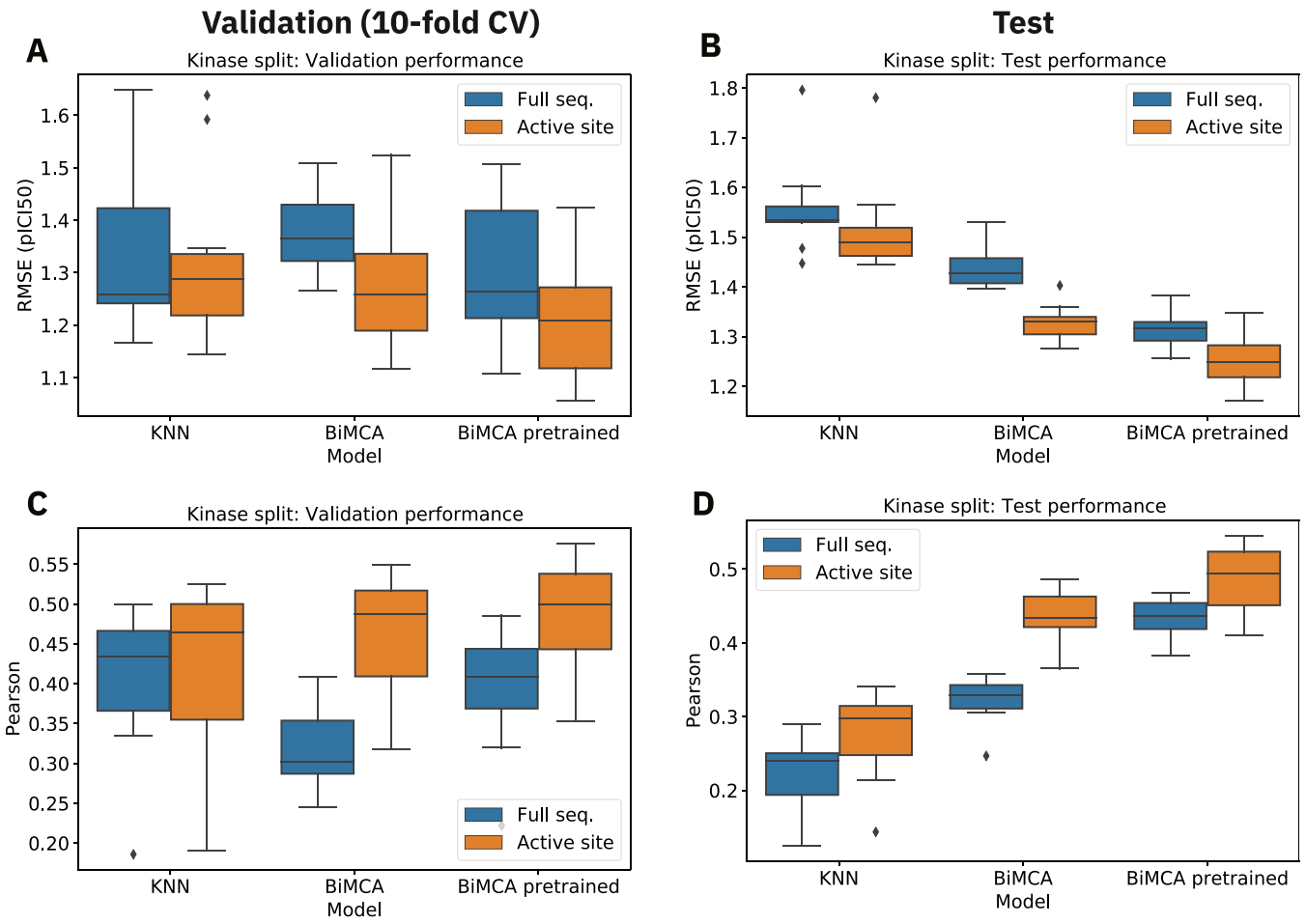


Figure 4. pIC_{50} prediction results on kinase split. The left and right columns show, respectively, the performance of all three models on the validation and test data. On both metrics, RMSE (upper row) and Pearson correlation coefficient (PCC, lower row), the active site configurations significantly outperform the full sequence configuration, irrespective of the utilized model. To see the exact numerical scores, please see Tables S1 and S2.

BiMCA model optimized an MSE loss with Adam[70] and was trained for 50 epochs with a learning rate of 0.005 and 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.[46,69] This generative model is then explored in a novel scheme, using Bayesian optimization with Gaussian processes (GPs) that optimizes the generative model to yield molecules with higher pIC_{50}. The GP approximates the predicted pIC_{50} 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 ChEMBL.[71] The model consists of two layers of stack-augmented gated recurrent units (GRUs) in both encoder and decoder. We trained two versions of this model, one using SMILES[26] sequences and the other one using SELFIES,[72] a novel molecular representation that was devised for deep generative models. The SMILES-based model is identical to that used in Born et al.;[69] the SELFIES model was used in Born et al.[46] These models are trained to optimize the standard VAE objective[73]

$$L_{\mathrm{VAE}}(\theta, \phi) = E_{q_\theta(\mathbf{z}|\mathbf{x})}[\log p_\phi(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \quad (3)$$

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 predefined 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 modeled using a multivariate unit Gaussian following standard VAE formulation.[73] 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.[56] Herein, we adopt BO to maximize pIC_{50} 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 Φ_A: $\mathcal{P} \times M \to A$. Considering the sheer size of the latent space $\mathcal{Z}$ and especially the cost to evaluate the function Φ_A[b] motivates us to formulate the problem in terms of Bayesian optimization: $\hat{z} = \mathrm{argmax}_{\mathbf{z}\in\mathcal{Z}}[\Phi_A(p, p_\phi(\mathbf{x}|\mathbf{z}))]$. BO

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

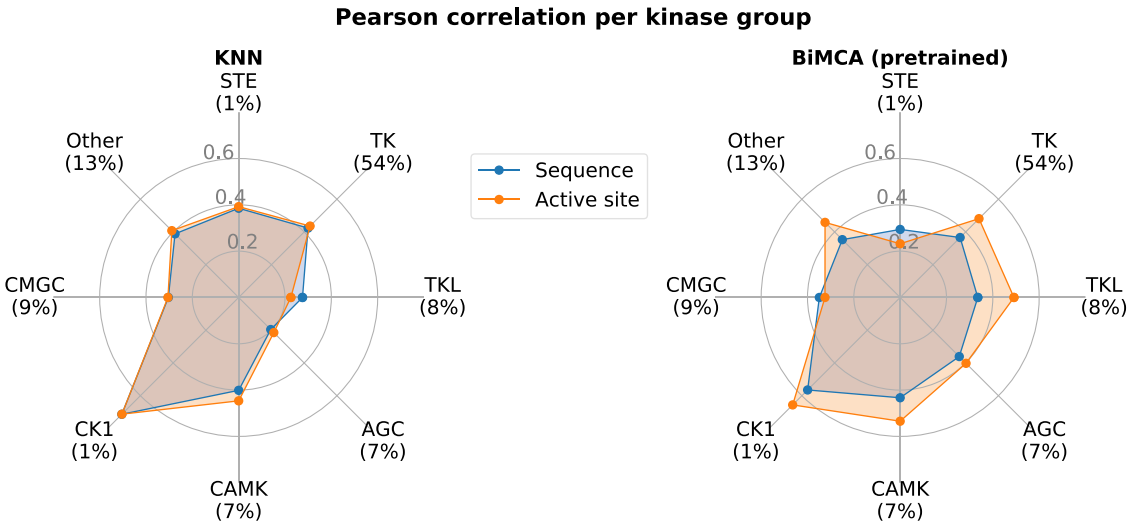


Figure 5. Performance in predicting pIC_{50} for unseen kinases according to the kinase group. For the KNN (left) and the pretrained BiMCA (right), the PCC of all samples of the respective kinase group is shown. Kinases that could not be classified with the catalogue from Manning et al.[79] are grouped into other.

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 modeled, with a prior specifying a probability distribution over functions, in our case, a GP prior: $\Phi_A \sim \mathrm{GP}[\hat{m}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')]$, where $\hat{m}$ is the mean function and k is the kernel returning similarity between 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 improvement[74] 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`[75] package. After each epoch, 300 molecules were generated from the latent points (invalid SMILES were discarded).

RESULTS ON PROTEIN−LIGAND INTERACTION 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 in both the KNN model (based on the Levenshtein distance of kinases) and the BiMCA model (based on nonlinear extrapolations). This task is significantly more challenging than splitting on ligands because the shape of the binding pocket largely governs the binding activity.[76] 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 features,[41] a phenomenon called hidden ligand bias.[77] 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.

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 the 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; cf. Figure 3G) and the active site models 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 10 validation folds compared to the held-out test data. Now, on the test data (see Figure 4B,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 outperform the KNN model by a large margin. In the BiMCA pretrained setting, we exploited all nonkinase data from BindingDB to warm up the BiMCA model before fine-tuning on the human kinome. After 20 epochs of pretraining, this model achieved an RMSE of 0.86 ($r = 0.82$) on the nonkinase 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). In an ablation study on the embeddings, we compared the effect of our learned embeddings with one-hot encodings and the BLOSUM62 substitution matrix (cf. Table S3 and Figure S4). The general trend regarding the superiority of the active site models strongly manifests, irrespective of the embedding type. All embedding types yield similar results, but overall, the learned embeddings perform best.

Kinase Groups. We further investigated the performance for the eight different groups of conventional protein kinases (ePKs) based on the classification by Hanks and Hunter.[78] Using the catalogue from Manning et al.[79] 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 models. With a few exceptions, the plot

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

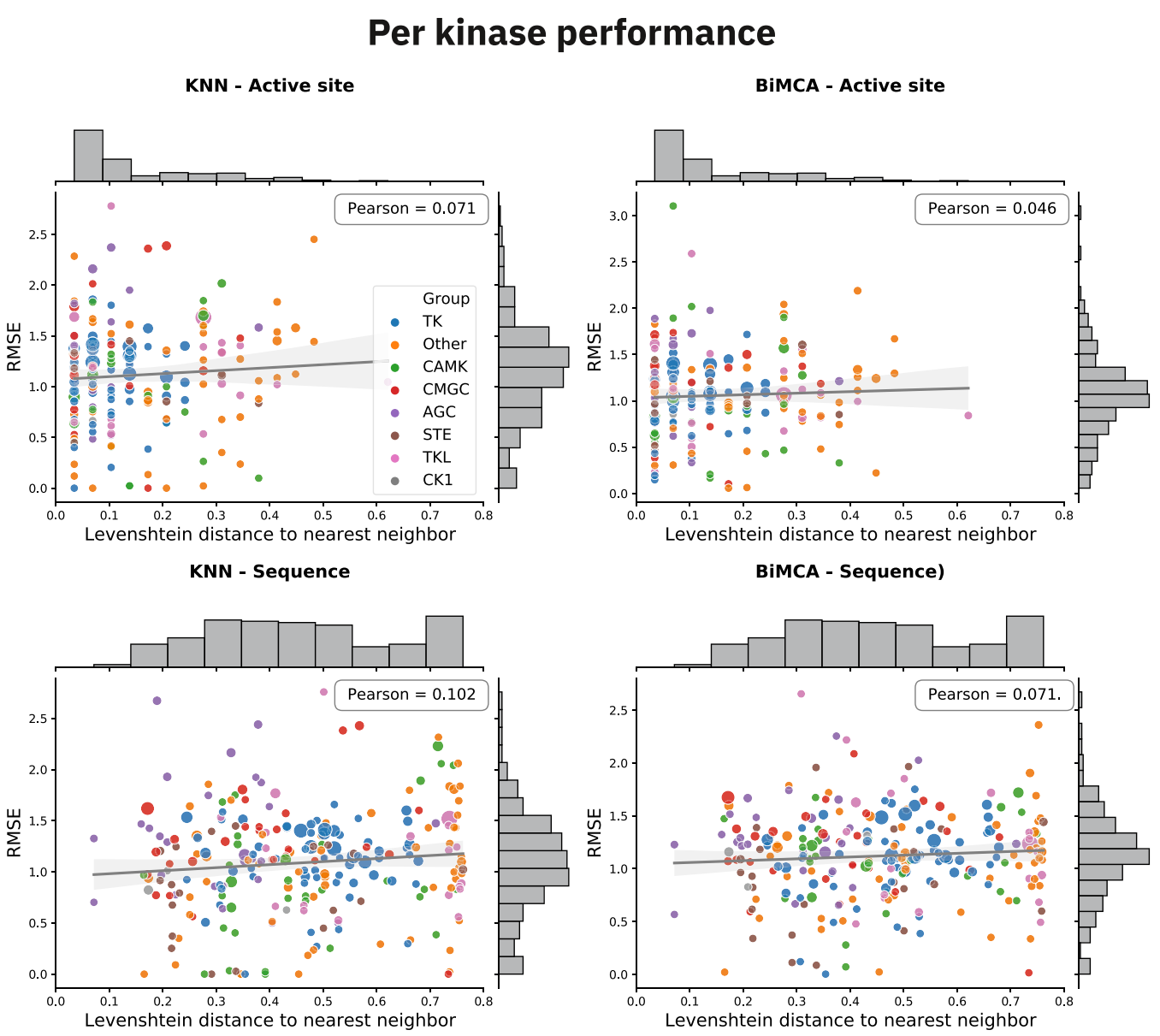


Figure 6. Dependency of model performance on similarity to the nearest neighbor in training data. In none of the four model configurations, a strong dependency/correlation between the performance on a specific kinase and the distance to the nearest neighbor in training data was found. Measures are obtained considering results on validation data.

indicates the superiority of the active site configuration consistently across the kinase families.

Notably, the most heavily screened kinase group is tyrosine kinases (TK), 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 tyrosine kinase-like (TKL) 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 structures are superior to active sites alone. We presume that this is because many TKL kinases (e.g., all RAF kinases[80]) 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 that for 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 kinases.[79] While the KNN predicts purely based on sequence similarity, the BiMCA can capture nonlinear relations and accordingly the performance gain for the BiMCA compared to that for 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 CPI for unseen TKL kinases. Another remarkable finding is the good result for proteins from the cell kinase 1 (CK1) group. We suppose this to be due to the high-intra-group and low-inter-group similarities of CK1 kinases. CK1s are highly conserved sequences, very similar to

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

each other but very distinct from other kinase groups[81] and form a distinct branch in the kinome tree[79] (note that samples were split by kinases, not by kinase groups).

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 the KNN as well as the pretrained BiMCA models, we therefore investigated the per-kinase performance as a function of the similarity to the nearest neighbor in the training data (cf. Figure 6).

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 model, 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 data set, are shown in Tables 1

Table 1. RMSE (on pIC_{50}) on Validation and Test Data (Ligand Split)[a]

data	configuration	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

[a]For each model and data partition, we mark the better representation in bold.

Table 2. Pearson Correlation Coefficient on Validation and Test Data (Ligand Split)[a]

data	configuration	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

[a]The same legend as Table 1.

(RMSE) and 2 (Pearson correlation). Like in the kinase split, all BiMCA models using active site information are superior to those using full primary structures (8.2 and 4.7% RMSE improvement for the BiMCA and pretrained BiMCA, respectively). For both models, these differences are statistically significant across the 10-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 eq 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 data set is 593 (see the 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). Similar to the kinase split, the ablation studies on the embedding type (cf. Table S3 and Figure S4) strengthen the general finding regarding the superiority of the active site models. The three embedding types (learned, one-hot encodings, BLOSUM62) all yielded very similar performances.

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 13 groups of the alleged mechanism of action based on the classification scheme by Roskoski.[82] 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.

In the right plot, we can see that, with the exception of MEK inhibitors, the active site model performed better on all 13 kinase inhibitor groups. For the KNN model (left), the predictions were extremely correlated between the sequence and active site model (see above) and therefore there are only negligible differences between both representations. However, apparent across both models is that prediction performance 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 residues,[67] 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 pocket.[83] 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,

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

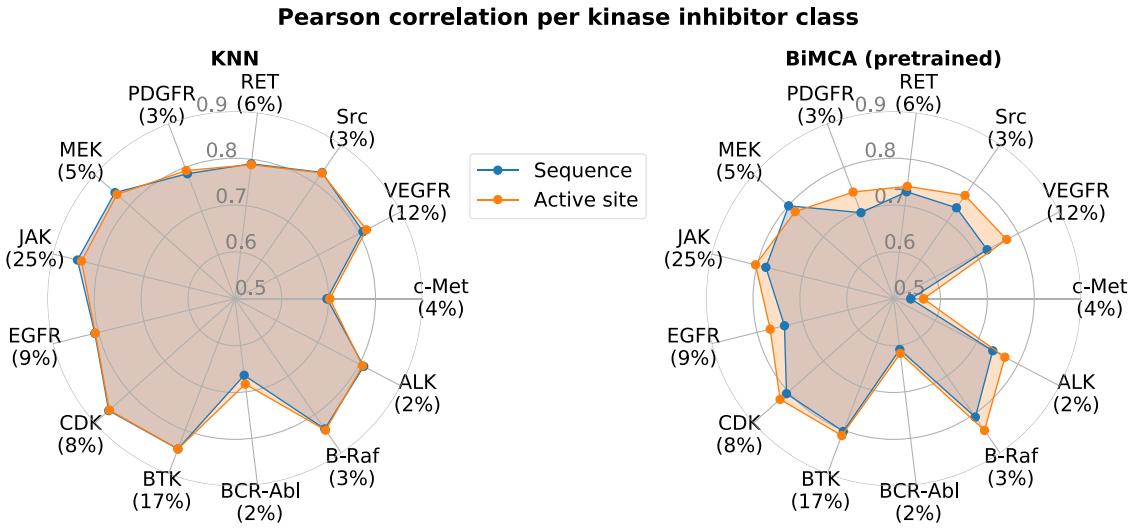


Figure 7. Performance in predicting pIC_{50} for unseen kinase inhibitors according to their primary protein target. For the KNN (left) and the pretrained BiMCA (right), the PCC of all samples of the respective kinase inhibitor class is shown.

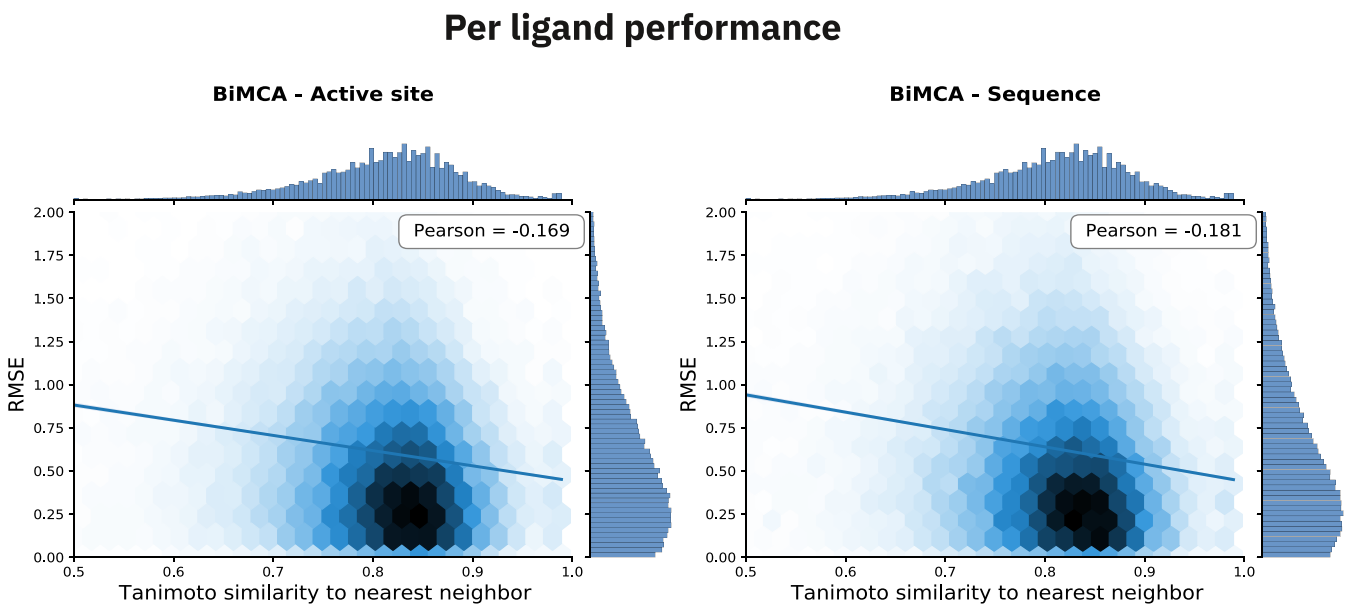


Figure 8. Dependency of pIC_{50} prediction on similar ligands. For each ligand, the performance (RMSE) is plotted against the Tanimoto similarity to the nearest training ligand. The color gradient shows the density of the molecules, and the line shows the correlation between both axes. Measures computed on validation data.

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).

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. Theoretically, the full sequence BiMCA models are strictly more expressive than their active site counterparts since they use a superset of residues and can exploit information from the entire protein. We suspect that, in practice, they perform worse for two reasons. First, unlike the active site models, they are implicitly posed with the additional task to filter out irrelevant residues, and second, the active site sequences carry implicit information about the 3D structure (since they are discontiguous in the full sequence). To investigate both hypotheses, we asked whether the model learned to capture 3D information from the full sequence alone. To that end, we analyzed the attention weights of the BiMCA's context attention mechanism and sought to assess whether the full sequence models automatically learned to focus on the active site residues. Several recent publications on CPI prediction that omitted 3D structure and completely relied on 1D text sequences incorporate neural attention mechanisms in their models.[27,36,37,40] 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. Many of the above works provided visualizations of protein attention and showed examples where the attention mechanism captured protein interaction sites.[36,37,40] However, these analyses were of qualitative nature and it was later demonstrated in a rigorous quantitative

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

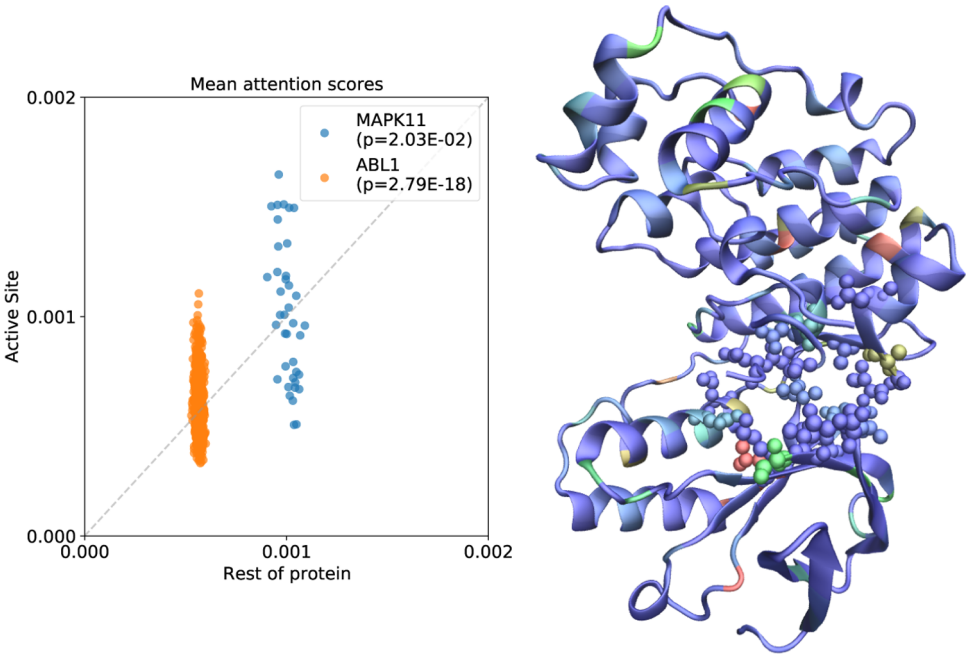


Figure 9. Kinase attention scores. Left: for each kinase−ligand pair of MAPK11 and ABL1, the mean attention scores on active site residues versus the remaining residues are shown. Right: exemplary visualization of attention values overlayed on the MAPK11 structure highlighting atoms with high weights (blue means low, green means medium, and red means high attention). Residues depicted as spheres belong to the active site.

evaluation that none of those models systematically highlights interaction sites.[27] Instead, Li et al.[27] showed that explicit supervision is required to excel at predicting pairwise noncovalent interactions.

For two exemplary kinases, MAPK11 and ABL1, we performed a statistical interpretability analysis to assess whether the sequence model paid disproportionally high attention to the active site residues (see Figure 9). Each protein−ligand pair (N = 39 for MAPK11 and N = 749 for ABL1) is shown as one point in Figure 9. Following the analysis by Li et al.[27] on predicting interaction sites, we assessed the model's ability to highlight the active site by two metrics: first, AUC, i.e., area under the receiver operating characteristic (ROC) curve between the binary labels and the per-residue attention scores. Second, the enrichment score, which is a precision-based metric derived from the binarized attention values that accounts for sequence length and randomly expected hits. Both metrics were computed per sample and averaged across all samples of a protein (for details on the metrics, see Li et al.[27]). Third, we evaluated statistical significance with a one-sided Mann−Whitney-U test. The AUC scores for MAPK11 and ABL1 are 0.518 and 0.516, respectively (AUC of random classifier: 0.5), and the average enrichment scores are 1.16 and 3.48 (random classifier: 1). In the comparison by Li et al.,[27] all investigated unsupervised attention-based methods[36,37,40] achieved AUCs around 0.5 and enrichment scores around 1, whereas explicit supervision on the interaction site yielded an AUC of 0.76 and an enrichment score of 10.7. Exact comparison, however, is not possible because our notion of active sites differs from their interaction site and the analysis was performed on slightly different subsets of BindingDB. Lastly, we find that for both kinases, the mean attention scores on the active site residues are significantly higher than on the remaining residues (MWU).

These results are insightful to understand why the active site models performed better. 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. While this trend is not consistent across all samples (cf. Figure 9), all three quantitative scores (AUC, enrichment, and MWU test) suggest that the BiMCA performs significantly above the chance level in extracting active sites. 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. This finding is in line with a predecessor of the BiMCA model where the attention mechanism was found to identify genes that significantly enrich apoptotic processes and capture structural differences between molecules in drug sensitivity prediction.[47]

Validation on External Test Data Set. To evaluate the performance of our model on an independent test data set, we utilized the data released in June 2021 as part of the IDG-DREAM challenge.[19] The challenge focused on understudied parts of the human kinome to catalogue the unexplored target space of kinase inhibitors. Thus, it resembles a particularly challenging data set, encompassed by 825 data points (cf. Supporting Data 1 by Cichońska et al.[19]). After restricting ourselves to kinases for which full sequence and active site information[67] was available, 720 samples remained, distributed across 276 kinases (32 unseen) and 93 ligands (all unseen). Evaluating on this data set is much more challenging than both the ligand and the kinase split because for many samples both ligands and kinases are unseen. Additional challenges posed by this data set compared to BindingDB are (1) experimental differences in the dose−response assays (multidose assays with a maximal concentration of 10 μM 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

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

Table 3. Evaluation on External Data Set by Cichońska et al.[19,a]

model	configuration	all	known kin.	unknown kin.	round 1	round 2
KNN	full sequence	0.224	0.242	0.032	0.132	0.32
	active site	0.244	0.282	-0.141	0.145	0.344
BiMCA	full sequence	0.16	0.169	0.064	0.102	0.185
	active site	**0.32**	**0.327**	**0.238**	**0.179**	**0.412**

[a]PCC values are reported.

data available in BindingBD as training data, whereas for the BiMCA, we built an ensemble of the 10 models from the ligand split.

The results on this data set are in alignment with our overall findings (cf. Table 3). The active site residue representation outperforms the full sequence representation 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 understudied kinases from Cichońska et al.[19] 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 in our training data.

Comparison to Related Work. In this subsection, we aim to compare the performance of the BiMCA and KNN models with previous proteochemometric approaches for protein−ligand affinity prediction. We largely rely on full sequence models because of unavailability of active site sequences. The results of our BiMCA and KNN models on the BindingDB split as provided by Karimi et al.[40] are shown in Table 4. It is curious to note that the simplistic KNN model achieved the best results on this split and outperformed all state-of-the-art deep neural networks such as MONN,[27] DeepAffinity,[40] or DeepDTA.[43] This is a surprising finding casting doubt on the necessity to develop complex architectures to predict IC_{50} of protein−ligand pairs, in particular, if the data split is lenient across proteins and ligands (cf. Figure S2 for visualization of the four splitting strategies). Furthermore, while this lenient split enables exact comparison to previous work, we emphasize that the authors of DeepDTA, DeepAffinity, and MONN all build model ensembles consisting of up to 30 individual models. These models achieved better results than the individual models shown in Table 4 (up to an RMSE of 0.658 and a PCC of 0.895[27]), but to ensure a fair comparison with our single models (it is widely known that model ensembles improve performance[85,86]), we omitted all ensembles from the results in Table 4. The performance of the BiMCA model on the lenient split falls behind the listed state-of-the-art models on the lenient split (cf. Table 4). However, the same BiMCA model excels at the generalization to novel protein families (cf. Table 5) and outperforms all previous models we found in the literature on three out of four protein families (estrogen receptors, ion channels, and GPCRs). Interestingly, only in the generalization task to receptor tyrosine kinases (RTKS) is the BiMCA outperformed by previous work. Unlike the data sets we curated for the kinase and ligand splits, the RTK data included both human and nonhuman kinases. In a final experiment, we removed all kinases without active site information from the RTK data set and then tested the generalization performance of active site and full sequence models when trained on a subset of the training data from Karimi et al.[40] The performances in Table 6 are results from single runs and are thus inconclusive regarding a superiority of active site or full sequences. One notable difference to the kinase and ligand split is that an entire protein family was held out. The results indicate that the KNN performed better at interpolation (mean prediction, RMSE), whereas the BiMCA better captured the fluctuations.

Table 4. Performance Evaluation of Different Prediction Approaches on Fixed-Split BindingDB Data Set from Karimi et al.[40,a]

model	RMSE	PCC
DeepDTA[43]	0.782	0.848
DeepAffinity[40]	0.780	0.840
DeepCDA[84]	0.808	0.844
MONN[27]	0.764	0.858
NN (k = 1)	0.862	0.83
KNN (k = 4)	**0.728**	**0.871**
KNN (k = 13)	0.783	0.848
BiMCA (full seq.)	0.892	0.786

[a]All models were trained and evaluated on the same samples. Models in the second half of the table are ours.

RESULTS ON 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 properties.[56,57] We selected six kinases for further analysis: the four JAK family members, as well as ABL1 (the target of imatinib, the first FDA-approved kinase inhibitor[3]) and MAPK11 (P38β), a thoroughly studied target from the MAPK family and isoform of MAPK14/P38α.[88] 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 predictions, operating either 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 7). Notably, the active sites 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 pIC_{50} prediction. Since the BiMCA operates on batches (size 128), its inference wall time is still below the KNN even when used on CPU. Notably, the KNN has to compute distances exhaustively for each sample and thus inference speed depends

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

Table 5. Generalization to New Protein Families Based on Fixed-Split BindingDB Data Set from DeepAffinity[40,a]

Model	ER		Ion Channel		RTK		GPCR		Mean	
	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
DeepAffinity SMILES[40]	1.53	0.16	1.34	0.17	1.24	0.39	1.40	0.24	1.38	0.24
DeepAffinity Graph[40]	1.68	0.05	1.43	0.10	1.74	0.01	1.63	0.04	1.62	0.05
DeepCDA[84]	-	0.10	-	0.31	-	0.42	-	0.28	-	0.28
Truong Jr[87] (ECFP/Pfam)	1.74	0.19	1.32	0.27	1.27	0.43	1.49	0.22	1.46	0.28
DeepAffinity Ensemble[40]	1.46	0.30	1.30	0.18	1.23	0.42	1.36	0.30	1.34	0.30
MLP ensemble[87]	1.51	0.24	1.36	0.19	1.26	0.42	1.36	0.33	1.37	0.29
Transformer ensemble[87]	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**

[a]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. Models below the dashed line and above the regular line are ensembles that can hardly be directly compared to our models. Numbers from other works are taken from their manuscripts since the split is fixed. DeepCDA did not report RMSE. The last two columns report the mean across the four data sets.

Table 6. Performance of Active Site and Full Sequence Representation for Generalization toward RTKs from Data Set by Karimi et al.[40,a]

model	configuration	RMSE	PCC
KNN	full sequence	**1.25**	**0.325**
	active site	1.27	0.299
BiMCA	full sequence	1.32	0.358
	active site	**1.28**	**0.359**

[a]These models were trained and evaluated on a subset of the data by Karimi et al.[40] for which active site information was available.

Table 7. Comparison of Active Site and Full Sequence Models[a]

model	configuration	no. of AAs	model size (M)	inference time (ms)
KNN	full sequence	742 ± 369		59 ± 19
	active site	29 ± 0		29 ± 8
BiMCA	full sequence	742 ± 369	14.2	18 ± 1
	active site	29 ± 0	0.6	6 ± 1

[a]Inference time measured on a 2.7 GHz Quad-Core `Intel Core i7`. The BiMCA was trained on a single `NVIDIA Tesla P100`.

on training data size. Moreover, the KNN inference speed linearly depends on the protein sequence length and increases from ∼30 to ∼ 100 ms for the shortest (29) to the longest (2527) sequences in the data set. Since the BiMCA uses zero-padding, inference time is practically independent of the underlying sequence length.

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).

The distributions indicate that the optimization led to higher-average pIC_{50} 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 pIC_{50} optimization with GP, the remaining results for the four JAK family members (see the Figure S5, Supporting Information) are inconclusive regarding a comparison of active site and sequence models (cf. Table 8). This is not surprising because the kinase representation does not directly impact the generative process but is only used in the pIC_{50} 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 pIC_{50} in any time interval (cf. Table 9). In relation to the results of previous works on GP optimization of sampled molecules,[56,57] we emphasize that optimizing target affinity is more challenging compared to many chemocentric properties such as drug-likeness or molecular weight.

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 pIC_{50} scores for each of the six targets and both kinase representations in the top row of Figure 11.

The molecules are rich in their structure and have predicted affinities in the range of 2−50 nM. We then selected for each kinase the top 5 most effective, yet aromatic molecules and retrieved the 10 most similar compounds experimentally measured for that target from BindingDB (based on Tanimoto similarity[61] of ECFP6 fingerprints[60]). The bottom row of Figure 11 then displays the distribution of these measured pIC_{50} scores. A remarkable observation is that for the four JAK targets the BindingDB samples selected from the active site molecules have a lower predicted IC_{50} 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 modeling of protein−ligand interaction 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 CPI prediction, our results suggest a superiority of active site residues when predicting affinity (measured by pIC_{50}) 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 data set from the IDG-DREAM challenge.

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

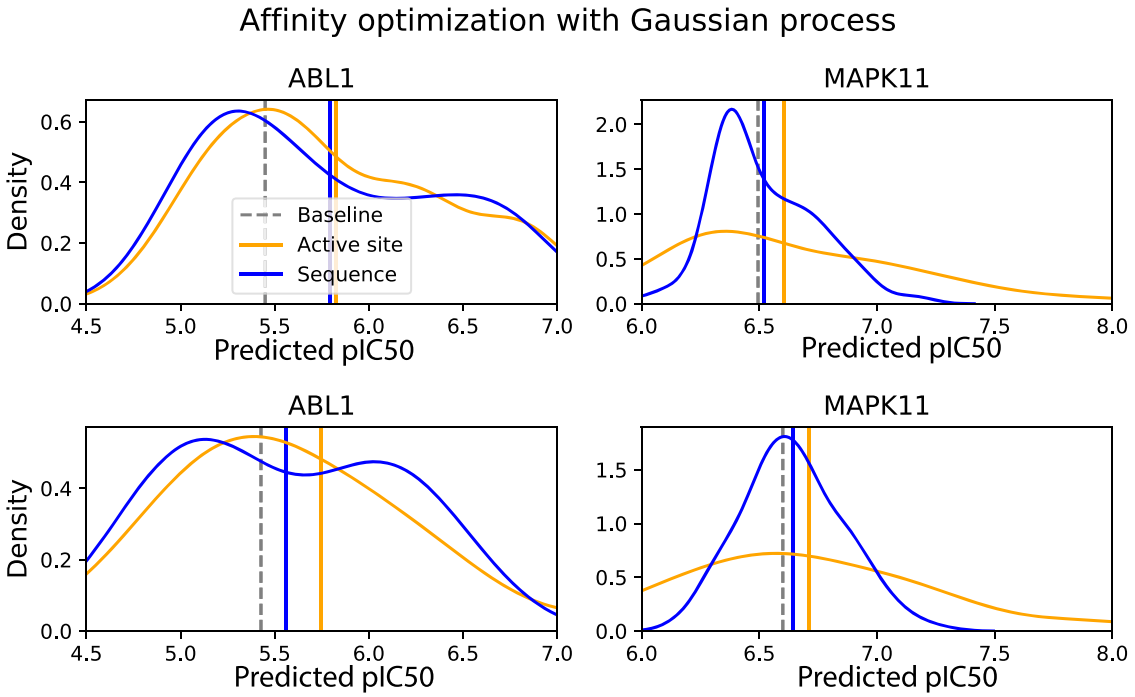


Figure 10. Bayesian optimization of latent space sampling with Gaussian processes. For two selected targets, the (predicted) pIC_{50} distribution of generated molecules is shown. Dashed lines denote the mean pIC_{50} before the optimization. Molecules were sampled from either the SELFIES generator (top row) or SMILES generator (bottom row). pIC_{50} was predicted with the KNN model.

Table 8. Results of GP Optimization[a]

ligand repr.	kinase repr.	baseline	optimized
SELFIES	full sequence	6.55 ± 0.6	6.60 ± 0.5
	active site	6.51 ± 0.6	6.59 ± 0.5
SMILES	full sequence	6.51 ± 0.6	6.57 ± 0.6
	active site	6.57 ± 0.6	6.60 ± 0.5

[a]The average pIC_{50} across the six kinases and the molecules generated through the optimization process is shown.

Table 9. Runtime Comparison in Sampling Effective Ligands[a]

evaluation	active site	sequence
time until 5% pIC_{50} improvement (min)	**14 ± 8**	21 ± 8
number of effective ligands in 25 min	**35 ± 19**	30 ± 16

[a]All ligands with a predicted IC_{50} < 100 nM (i.e., pIC_{50} > 7) are considered effective.

This is an important and maybe surprising finding because the active site residues are a subset of the full primary sequence. Indeed, the full sequence codes for the active site structure as well as additional, more distant determinants of binding and dynamics. It seems that exclusively providing the active site residues increases the signal-to-noise ratio in the sequences and implicitly conveys information about the 3D structure, consequently leading to better performance. This hypothesis is partly corroborated by our attention analysis. Without supervision on the importance of the residues, the sequence model exhibits mild and statistically significant but overall very limited abilities to focus on the active site residues (cf. Figure 9). In contrast to these sparse elements of tertiary structures in the full sequence models, the active site models are equipped with an implicit but strong inductive bias about the 3D structure stemming from the discontiguity of the active site residues in the full sequence. Indeed, Li et al.[27] have shown 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.[36,37,40] Contrarily, Vig et al.[89] 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 proteochemometric models benefit from pretraining on large-scale pan-protein data even if the final use case is limited to one family. Lastly, we compared both models (KNN and BiMCA) to the state-of-the-art on the same data sets from Karimi et al.[40] and found that our models obtain similar and partly superior performance to all existing approaches.

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 pIC_{50} reported for BindingDB's closest ligand to the generated molecules. While our molecular generation experiments focused on single-property optimization, the Gaussian process framework can be readily applied to simultaneous multiobjective optimization. Indeed, this composition problem is much more challenging than single-property optimization and most works on GPs for molecular

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

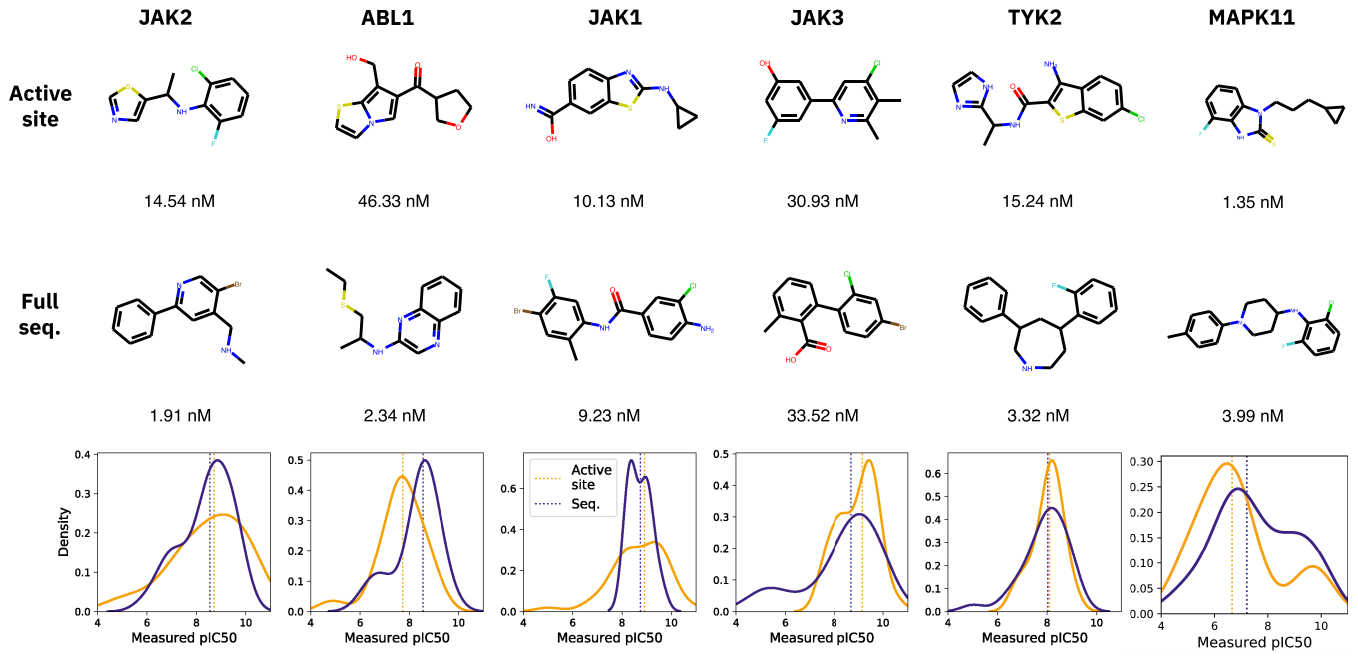


Figure 11. Best-generated molecules per target. For each of the six targets, the molecule with the lowest predicted IC_{50} is shown in the top two rows. Predictions were obtained with a model from the ligand split trained either on active sites or on full sequences. The bottom row shows the distribution of measured pIC_{50} values (as reported in BindingDB) for the most similar kinase inhibitors of a subset of the generated molecules. The dotted vertical lines denote the medians.

optimization focused on single properties.[56,57] However, it was shown that multiobjective optimization can be efficiently addressed with particle swarm optimization.[90]

Outlook. Further research is needed to validate our findings on the superiority of only using active site residues on more specific data sets, especially given the high discrepancy between random and "realistic" test sets of kinase inhibitors that can lead to performance gaps in practice.[14] 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 data sets (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 samples[66]). Moreover, future work could explore hybrid approaches between full sequence and active site information to leverage 3D information in 1D models. This could be achieved either by enriching the training process with pairwise noncovalent compound−protein residue interactions in a semisupervised setting based on the data set by Li et al.[27] or simply by providing 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 methods.[91]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00889.

Additional details about the data splitting and complementary results for affinity prediction and de novo molecular generation (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jannis Born − *IBM Research Europe, 8804 Rüschlikon, Switzerland; Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland;* orcid.org/0000-0001-8307-5670; Email: jab@zurich.ibm.com

Matteo Manica − *IBM Research Europe, 8804 Rüschlikon, Switzerland;* orcid.org/0000-0002-8872-0269; Email: tte@zurich.ibm.com

Authors

Tien Huynh − *IBM Research, Yorktown Heights, New York 10598, United States*

Astrid Stroobants − *Department of Chemistry, Imperial College London, SW7 2AZ London, United Kingdom;* orcid.org/0000-0002-8567-8480

Wendy D. Cornell − *IBM Research, Yorktown Heights, New York 10598, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00889

Author Contributions

The manuscript has been initially drafted by J.B. with contributions from all other authors. The final version has been revised multiple times by J.B., M.M., and W.D.C. T.H. and W.D.C. produced the alignment data, and J.B. 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

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257

then adapted by M.M. J.B. executed the experiments presented in the study. J.B. produced all of the visualizations 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.D.C. conceived the study. The authors thank the anonymous reviewers for their constructive feedback. The final version has been revised.

Notes

The authors declare no competing financial interest.

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.

■ ADDITIONAL NOTES

[a]That is, combining multiple single-target models into one model with multilabel classification.

[b]Either a costly forward pass through the BiMCA or an exhaustive computation of sample similarities with the KNN.

■ REFERENCES

(1) Cohen, P. Protein kinases — the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.

(2) Cohen, P.; Alessi, D. R. Kinase drug discovery−what's next in the field? *ACS Chem. Biol.* **2013**, *8*, 96−104.

(3) Cohen, P.; Cross, D.; Jänne, P. A. Kinase drug discovery 20 years after imatinib: Progress and future directions. *Nat. Rev. Drug Discovery* **2021**, *20*, 551−569.

(4) Yu, M.; Tadesse, S.; Wang, S. US FDA-Approved Small-Molecule Kinase Inhibitors for Cancer Therapy. *Burger's Medicinal Chemistry and Drug Discovery*; John Wiley & Sons, Inc., 2021.

(5) Fedorov, O.; Müller, S.; Knapp, S. The (un)targeted cancer kinome. *Nat. Chem. Biol.* **2010**, *6*, 166−169.

(6) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X.-P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2016**, *34*, 95−103.

(7) Bongers, B. J.; IJzerman, A. P.; Van Westen, G. J. Proteochemometrics − recent developments in bioactivity and selectivity modeling. *Drug Discovery Today: Technol.* **2019**, *32*−*33*, 89−98.

(8) 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.

(9) Volkamer, A.; Eid, S.; Turk, S.; Rippmann, F.; Fulle, S. Identification and Visualization of Kinase-Specific Subpockets. *J. Chem. Inf. Model.* **2016**, *56*, 335−346.

(10) Blaschke, T.; Miljković, 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.

(11) Ung, P. M.-U.; Rahman, R.; Schlessinger, A. Redefining the Protein Kinase Conformational Space with Machine Learning. *Cell Chem. Biol.* **2018**, *25*, 916.e2−924.e2.

(12) 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. *J. Chem. Inf. Model.* **2011**, *51*, 1942−1956.

(13) Martin, E.; Mukherjee, P. Kinase-Kernel Models: Accurate In silico Screening of 4 Million Compounds Across the Entire Human Kinome. *J. Chem. Inf. Model.* **2012**, *52*, 156−170.

(14) 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. *J. Chem. Inf. Model.* **2017**, *57*, 2077−2088.

(15) 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. *J. Chem. Inf. Model.* **2009**, *49*, 1974−1985.

(16) 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 Bioinf.* **2020**, *21*, No. 520.

(17) 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. *Sci. Rep.* **2021**, *11*, No. 15993.

(18) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(19) Cichońska, A.; Ravikumar, B.; Allaway, R. J.; Wan, F.; Park, S.; Isayev, O.; Li, S.; Mason, M.; Lamb, A.; Tanoli, Z.; Jeon, M.; Kim, S.; Popova, M.; Capuzzi, S.; Zeng, J.; Dang, K.; Koytiger, G.; Kang, J.; Wells, C. I.; Willson, T. M.; Tan, M.; Huang, C.-H.; Shih, E. S. C.; Chen, T.-M.; Wu, C.-H.; Fang, W.-Q.; Chen, J.-Y.; Hwang, M.-J.; Wang, X.; Ben Guebila, M.; Shamsaei, B.; Singh, S.; Nguyen, T.; Karimi, M.; Wu, D.; Wang, Z.; Shen, Y.; öztürk, H.; Ozkirimli, E.; Özgür, A.; Lim, H.; Xie, L.; Kanev, G. K.; Kooistra, A. J.; Westerman, B. A.; Terzopoulos, P.; Ntagiantas, K.; Fotis, C.; Alexopoulos, L.; Boeckaerts, D.; Stock, M.; De Baets, B.; Briers, Y.; Luo, Y.; Hu, H.; Peng, J.; Dogan, T.; Rifaioglu, A. S.; Atas, H.; Atalay, R. C.; Atalay, V.; Martin, M. J.; Lee, J.; Yun, S.; Kim, B.; Chang, B.; Turu, G.; Misák, Á.; Szalai, B.; Hunyady, L.; Lienhard, M.; Prasse, P.; Bachmann, I.; Ganzlin, J.; Barel, G.; Herwig, R.; Oršolić, D.; Lučić, B.; Stepanić, V.; Šmuc, T.; Oprea, T. I.; Schlessinger, A.; Drewry, D. H.; Stolovitzky, G.; Wennerberg, K.; Guinney, J.; Aittokallio, T.; IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat. Commun.* **2021**, *12*, No. 3307.

(20) 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* **2021**, DOI: 10.26434/chemrxiv.14604720.v1.

(21) Fare, C.; Turcani, L.; Pyzer-Knapp, E. O. Powerful, transferable representations for molecules through intelligent task selection in deep multitask networks. *Phys. Chem. Chem. Phys.* **2020**, *22*, 13041−13048.

(22) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4*, 4367−4375.

(23) 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.

(24) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.

(25) 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. *Chem. Sci.* **2018**, *9*, 5441−5451.

(26) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(27) 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 Syst.* **2020**, *10*, 308.e11−322.e11.
(28) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.
(29) 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*, No. e7362.
(30) Bitencourt-Ferreira, G.; de Azevedo, W. F. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophys. Chem.* **2018**, *240*, 63−69.
(31) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. F. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E. Improved Protein−Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* **2021**, *61*, 1583−1592.
(32) Bao, J.; He, X.; Zhang, J. Z. H. DeepBSP—a Machine Learning Method for Accurate Prediction of Protein−Ligand Docking Structures. *J. Chem. Inf. Model.* **2021**, *61*, 2231−2240.
(33) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.
(34) 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. *J. Chem. Inf. Model.* **2020**, *60*, 2791−2802.
(35) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: Protein−ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.
(36) Gao, K. Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. In *Interpretable Drug Target Prediction Using Deep Neural Representation*, Interpretable Drug Target Prediction Using Deep Neural Representation, 2018.
(37) 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.
(38) 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. *J. Chem. Inf. Model.* **2019**, *59*, 3981−3988.
(39) Torng, W.; Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* **2019**, *59*, 4131−4149.
(40) 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.
(41) 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.
(42) Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting compound-protein interaction prediction by deep learning. *Methods* **2016**, *110*, 64−72.
(43) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug−target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821−i829.
(44) Gonczarek, A.; Tomczak, J. M.; Zaręba, S.; Kaczmar, J.; Dąbrowski, P.; Walczak, M. J. Interaction prediction in structure-based virtual screening using deep learning. *Comput. Biol. Med.* **2018**, *100*, 253−258.
(45) Zhao, L.; Wang, J.; Pang, L.; Liu, Y.; Zhang, J. GANsDTA: Predicting Drug-Target Binding Affinity Using GANs. *Front. Genet.* **2020**, *10*, No. 1243.
(46) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakarajan, 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. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 025024.
(47) Manica, M.; Oskooei, A.; Born, J.; Subramanian, V.; Sáez-Rodríguez, J.; Rodríguez Martínez, M. Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. *Mol. Pharmaceutics* **2019**, *16*, 4797−4806.
(48) Cadow, J.; Born, J.; Manica, M.; Oskooei, A.; Rodríguez Martínez, M. PaccMann: A web service for interpretable anticancer compound sensitivity prediction. *Nucleic Acids Res.* **2020**, *48*, W502−W508.
(49) Thirunavukkarasu, M. K.; Shin, W.-H.; Karuppasamy, R. Exploring safe and potent bioactives for the treatment of non-small cell lung cancer. *3 Biotech* **2021**, *11*, No. 241.
(50) Parate, S.; Kumar, V.; Danishuddin; Hong, J. C.; Lee, K. W. Computational Investigation Identified Potential Chemical Scaffolds for Heparanase as Anticancer Therapeutics. *Int. J. Mol. Sci.* **2021**, *22*, No. 5311.
(51) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.
(52) Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future Med. Chem.* **2019**, *11*, 567−597.
(53) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 1496−1505.
(54) 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.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.
(55) Born, J.; Manica, M. Trends in Deep Learning for Property-driven Drug Design. *Curr. Med. Chem.* **2021**, *28*, 7862.
(56) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(57) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **2020**, *11*, 577−586.
(58) Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, 707−710.
(59) Nazarshodeh, E.; Sheikhpour, R.; Gharaghani, S.; Sarram, M. A. A novel proteochemometrics model for predicting the inhibition of nine carbonic anhydrase isoforms based on supervised Laplacian score and k-nearest neighbour regression. *SAR QSAR Environ. Res.* **2018**, *29*, 419−437.
(60) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(61) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corporation, 1958.
(62) Weber, A.; Born, J.; Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **2021**, *37*, i237−i244.
(63) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.
(64) Bahdanau, D.; Cho, K.; Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*, 3rd International Conference on Learning Representations, ICLR2015, San Diego, CA, 2015.
(65) Markert, G.; Born, J.; Manica, M.; Schneider, G.; Rodriguez Martinez, M. In *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.
(66) 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 Res.* **2016**, *44*, D1045−D1053.
(67) Modi, V.; Dunbrack, R. L. A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. *Sci. Rep.* **2019**, *9*, No. 19790.
(68) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. In *Pytorch: An Imperative Style, High-performance Deep Learning Library*, Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019; pp 8026−8037.
(69) Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Markert, G.; Rodríguez Martínez, M. PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **2021**, *24*, No. 102269.
(70) Kingma, D. P.; Ba, J. L. In *Adam: A Method for Stochastic Optimization*, 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, 2015.
(71) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: An update. *Nucl. Acids Res.* **2014**, *42*, D1083−D1090.
(72) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100. *Mach. Learn.: Sci. Technol.* **2020**, *1*, No. 045024.
(73) Kingma, D. P.; Welling, M. In *Auto-Encoding Variational Bayes*, 2nd International Conference on Learning Representations, ICLR, 2014.
(74) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Global Optim.* **1998**, *13*, 455−492.
(75) Head, T.; Kumar, M.; Nahrstaedt, H.; Louppe, G.; Shcherbatyi, I. scikit-optimize/scikit-optimize, 2020. https://doi.org/10.5281/zenodo.4014775.
(76) Wang, H.; Qiu, J.; Liu, H.; Xu, Y.; Jia, Y.; Zhao, Y. HKPocket: Human kinase pocket database for drug design. *BMC Bioinf.* **2019**, *20*, No. 617.
(77) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.
(78) Hanks, S. K.; Hunter, T. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification 1. *FASEB J.* **1995**, *9*, 576−596.
(79) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.
(80) 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. *J. Biol. Chem.* **2009**, *284*, 3183−3194.
(81) Knippschild, U.; Krüger, M.; Richter, J.; Xu, P.; Garcia-Reyes, B.; Peifer, C.; Halekotte, J.; Bakulev, V.; Bischof, J. The CK1 Family: Contribution to Cellular Stress Response and Its Role in Carcinogenesis. *Front. Oncol.* **2014**, *4*, No. 96.
(82) Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* **2016**, *103*, 26−48.
(83) Wu, P.-K.; Park, J.-I. MEK1/2 inhibitors: Molecular activity and resistance mechanisms. *Semin. Oncol.* **2015**, 849−862.
(84) Abbasi, K.; Razzaghi, P.; Poso, A.; Amanlou, M.; Ghasemi, J. B.; Masoudi-Nejad, A. DeepCDA: Deep cross-domain compound−protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics* **2020**, *36*, 4633−4642.
(85) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(86) Zhou, Z.-H. *Machine Learning*; Springer, 2021; pp 181−210.
(87) Truong, T. F., Jr. Interpretable Deep Learning Framework for Binding Affinity Prediction. Ph.D. Thesis, Massachusetts Institute of Technology, 2020.
(88) Dhillon, A. S.; Hagan, S.; Rath, O.; Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **2007**, *26*, 3279−3290.
(89) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. In *BERTology Meets Biology: Interpreting Attention in Protein Language Models*, 9th International Conference on Learning Representations, ICLR 2021, 2021.
(90) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016−8024.
(91) Cui, Y.; Dong, Q.; Hong, D.; Wang, X. Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC Bioinf.* **2019**, *20*, No. 93.

https://doi.org/10.1021/acs.jcim.1c00889
J. Chem. Inf. Model. 2022, 62, 240−257