JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

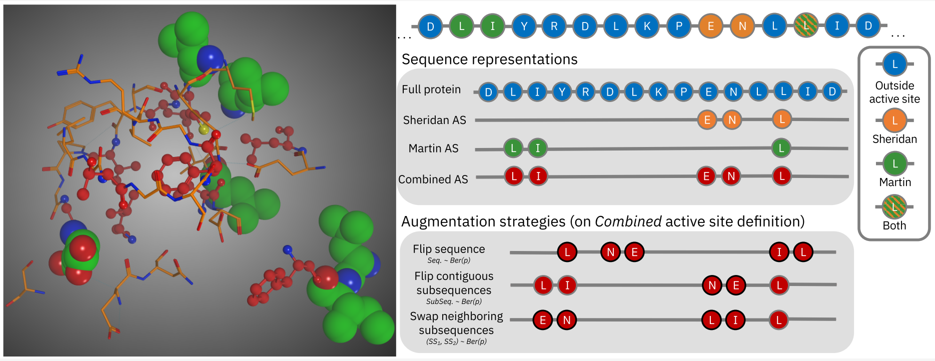
pubs.acs.org/jcim Letter

On the Choice of Active Site Sequences for Kinase-Ligand Affinity Prediction

Jannis Born,* Yoel Shoshan, Tien Huynh, Wendy D. Cornell, Eric J. Martin, and Matteo Manica

Cite This: *J. Chem. Inf. Model.* 2022, 62, 4295−4299 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Recent work showed that active site rather than full-protein-sequence information improves predictive performance in kinase-ligand binding affinity prediction. To refine the notion of an "active site", we here propose and compare multiple definitions. We report significant evidence that our novel definition is superior to previous definitions and better models of ATP-noncompetitive inhibitors. Moreover, we leverage the discontiguity of the active site sequence to motivate novel protein-sequence augmentation strategies and find that combining them further improves performance.

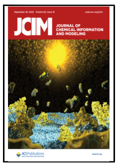
1. INTRODUCTION

The human kinome is indispensable for the regulation of cell function and comprises many widely studied drug targets due to its key role in a multitude of diseases such as cancer. Therefore, proteochemometric models that can predict protein−ligand interaction, kinetic energies, or binding affinities have received growing interest.[1] Most efforts either rely on structure-based[2,3] or sequence-based[4,5] deep learning models. While structure-based approaches can, in principle, model binding dynamics more realistically, their practical superiority is questionable: recent work evidenced that incorporating noncovalent interactions does not give benefits compared to simple protein/ligand descriptors.[6]

Sequence-based models for affinity prediction are usually trained on prohibitively long protein sequences that consist predominantly of residues irrelevant for binding. Recently, however, we demonstrated that using only residues of the ATP-binding site rather than the full protein increases the signal-to-noise-ratio in the protein representation and improves significantly the performance in protein−ligand affinity prediction for human kinases.[7] All experiments in that work were based on an active site definition from Sheridan et al. (ref 8) which comprises 29 residues surrounding the ATP-binding site that were identified using MSA.

The superiority of the active site representation manifested consistently across all ligand types, with the sole exception of one drug class: MEK/MAPK inhibitors.[7] Notably, this class contains many allosteric binders, in particular ATP-non-competitive MAPK inhibitors that bind to a unique site near the ATP-binding pocket.[9] One goal of the presented work is to address this systematic limitation in modeling allosteric binders and refine the definition of an "active site" for binding affinity prediction. Therefore, we leverage an alternative active site definition comprising 16 residues from ref 10 that includes 6 residues farther away from the immediate binding site (see Figure 1A). These two representations are compared to a broader *Combined* definition (cf. Figure 1B). Last, we explore additional mechanisms to leverage the knowledge about the active site, in particular how it can inspire data augmentation. We propose two new protein sequence augmentation techniques and find that they have complementary positive effects.

Received: July 11, 2022
Published: September 13, 2022



https://doi.org/10.1021/acs.jcim.2c00840

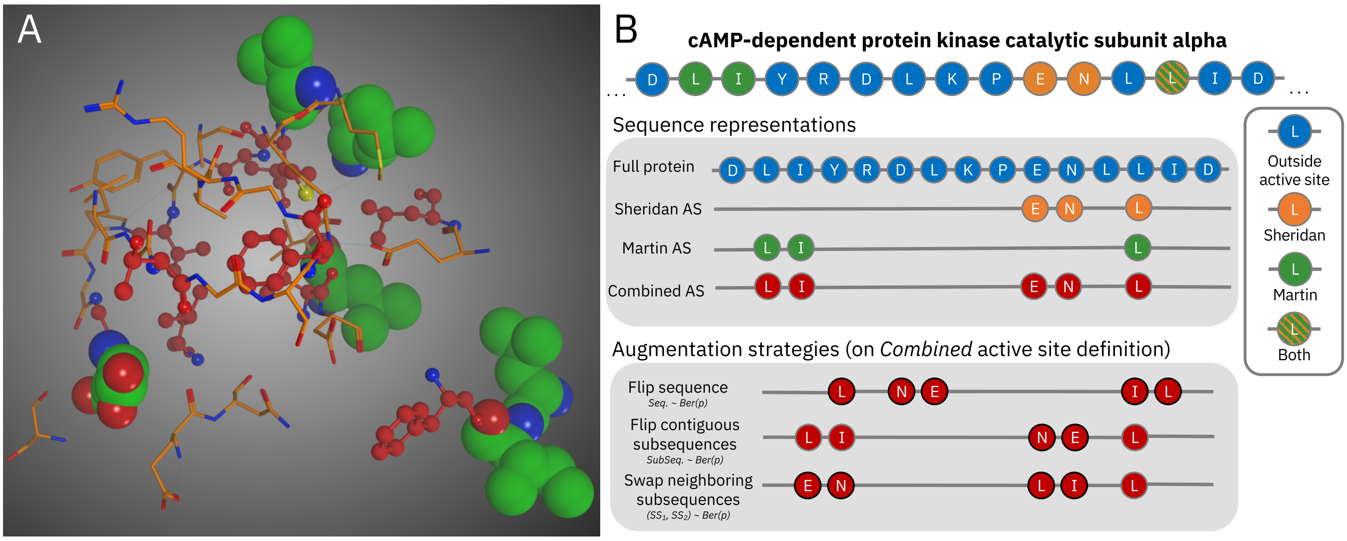

A
B
cAMP-dependent protein kinase catalytic subunit alpha
D L I Y R D L K P E N L L I D
Sequence representations
Full protein
D L I Y R D L K P E N L L I D
Sheridan AS
E N L
Martin AS
L I L
Combined AS
L I E N L
Augmentation strategies (on Combined active site definition)
Flip sequence
Seq. ~ Ber(p)
L N E I L
Flip contiguous subsequences
SubSeq. ~ Ber(p)
L I N E L
Swap neighboring subsequences
(SS1, SS2) ~ Ber(p)
E N L I L
L
Outside active site
L
Sheridan
L
Martin
L
Both

Figure 1. Overview of active site site definitions and representations. A) Visualization of cAMP-dependent protein kinase catalytic subunit alpha (P17612). Residues unique to the active site definitions of refs 8 and 10 are colored in orange and green, respectively. Residues contained in both definitions are shown in red. B) Partial amino acid sequence (residues 48−62) of the same kinase. The upper gray panel displays the four kinase sequence representations examined in this work. The lower gray panel visualizes three kinase augmentation strategies, exemplified on the "combined" active site definition: flipping (i.e., reversing) the entire sequence, flipping contiguous subsequences, and swapping neighboring subsequences. Residues affected by the augmentation are encircled in black.

Table 1. Results on Validation and Test Data (Ligand Split)[a]

		RMSE (↓)		Pearson (↑)	
data	config	BiMCA	BiMCA-pre	BiMCA	BiMCA-pre
val.	full sequence	$0.908_{\pm 0.01}$	$0.848_{\pm 0.01}$	$0.748_{\pm 0.00}$	$0.782_{\pm 0.01}$
	AS (Sheridan)	$0.829_{\pm 0.01}$	$0.821_{\pm 0.01}$	$0.794_{\pm 0.00}$	$0.797_{\pm 0.01}$
	AS (Martin)	$0.839_{\pm 0.01}$	$0.813_{\pm 0.01}$	$0.791_{\pm 0.00}$	$\mathbf{0.804}_{\pm 0.01}$
	AS (combined)	$\mathbf{0.828}_{\pm 0.01}$	$\mathbf{0.811}_{\pm 0.01}$	$\mathbf{0.797}_{\pm 0.01}$	$\mathbf{0.804}_{\pm 0.01}$
test	full sequence	$0.912_{\pm 0.01}$	$0.863_{\pm 0.01}$	$0.744_{\pm 0.00}$	$0.774_{\pm 0.01}$
	AS (Sheridan)	$\mathbf{0.832}_{\pm 0.01}$	$0.826_{\pm 0.01}$	$0.792_{\pm 0.01}$	$0.795_{\pm 0.01}$
	AS (Martin)	$0.842_{\pm 0.01}$	$0.818_{\pm 0.01}$	$0.789_{\pm 0.01}$	$0.801_{\pm 0.01}$
	AS (combined)	$\mathbf{0.832}_{\pm 0.01}$	$\mathbf{0.816}_{\pm 0.01}$	$\mathbf{0.795}_{\pm 0.01}$	$\mathbf{0.802}_{\pm 0.01}$

[a] 10-fold cross-validation results on kinase data from BindingDB. For each model and data partition, we show mean and standard deviation across 10 folds and mark the best representation in bold.

2. KINASE SEQUENCE REPRESENTATION

2.1. Active Site Definitions. In our previous work,[7] the active site representation relied on 29 residues defined originally in Sheridan et al. [ref 8, Table 1]. These residues are short contiguous subsequences that lie discontiguously in the original sequence (cf. Figure 1B top). Here, the predictive power of this *Sheridan* definition is compared to 16 residues that were found most relevant for kinase kernel models by Martin et al.[10] These *Martin* residues were identified from a starting set of 46 residues based on how frequently they were picked with a variable selection algorithm for a large set of kinase-kernel models. Since only 10 of these 16 residues are overlapping with the *Sheridan* definition, we also examine a *Combined* active site definition with 35 residues. For a table with the PKA numbering of all residues, see subsection S1.3.

2.2. Kinase Sequence Augmentation. While the MSA guarantees a meaningful and consistent ordering of the residues (and their physical roles), the sequences do not provide explicit 3D information on protein conformation. Especially, proximity in the sequence likely but not necessarily corresponds to proximity in 3D space. We therefore hypothesized that sequence augmentation strategies could assist to learn general binding patterns for two reasons: 1) There may be 1D representations that align better with the 3D relation of residues than the original sequence. Representing a kinase as a distribution of sequences reflects this lack of knowledge, might regularize the model, and thus improves generalization, especially to unseen target families. 2) Static roles of specific residue positions may induce overfitting in practice as the model might memorize too specific patterns.

A natural augmentation technique for protein sequences is **flipping** (**F**) the entire reduced residue set ($p = 0.5$). Moreover, we leveraged the knowledge about the location of the active site residues and exploited their discontiguity in the full sequence to motivate two additional augmentation strategies (cf. Figure 1B bottom). First, **flipping contiguous subsequences** (**FS**): Since subsequences of the active site that are contiguous in the full sequence are close together in space, reading such sequences from either direction should not affect model predictions ($p = 0.5$). Second, **swapping neighboring contiguous subsequences** (**SS**): This strategy relies on the assumption that neighboring contiguous sequences have a higher probability to be closer in space than distant active site subsequences ($p = 0.2$). Last, we

https://doi.org/10.1021/acs.jcim.2c00840

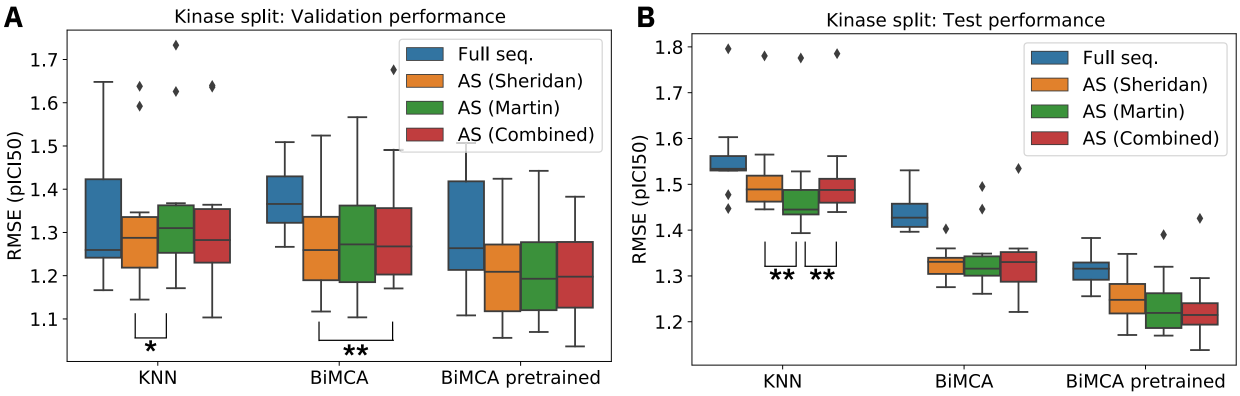


Figure 2. RMSE in affinity prediction for kinase split on validation and test data. 10-fold cross-validation results on kinase data from BindingDB. Performance of validation (A) and test data (B) is shown. Statistically significant differences between the three different active site configurations are marked with a star.

also explore combinations of these augmentation strategies. For details, see Supporting Information S1.2.

3. EXPERIMENTAL SETUP

The experimental setup is largely identical to the binding affinity prediction task described in ref 7. We take data from BindingDB[11] and examine two types of models, a *k*-nearest-neighbor (KNN) model that builds a joint similarity space of protein and ligand distances and a deep neural network called BiMCA (Bimodal Multiscale Convolutional Attention encoder[12]) that ingests protein and ligand sequences (SMILES strings) and consists of convolutional and attention layers. The remaining methods (data source and preprocessing, model definitions) can be found in Supporting Information S1.

4. RESULTS

4.1. Ligand Split. This split corresponds to the classical discovery setting: Kinases are shared across train and validation data, and thus, we measure generalization in the ligand space. The results on the ligand split confirm the superiority of using active sites compared to full sequences, irrespective of the exact definition of the active site (cf. Table 1). The table clearly indicates that the *Combined* representation yields consistently the best results for both models, both metrics and validation and test data (cf. Table 1). These improvements are statistically significant (Wilcoxon signed-rank test, $W+$) compared to at least one active site definition for all settings (see Figure S1 and Figure S2).

There are several kinase inhibitor classes with notable performance differences: First, the conspicuous inferiority of the *Sheridan* definition to the full protein sequence for MEK inhibitors [ref 7, Figure 7], caused by allosteric MAPK inhibitors that cannot be modeled using an ATP-based active site definition, was a limitation of our previous work. Importantly, this can be resolved using the *Martin* or the *Combined* active site definition with 6 more distant residues (cf. Figure S1 panel C). These definitions include residues distant from the ATP-binding site and around the "hydrophobic spine", hypothesized to affect the stability of binding site features or the active and inactive forms.[13] Second, the *Martin* definition also includes T51, a residue that builds an important salt bridge with residues in the same loop in many CDK kinases, another class where *Martin*/*Combined* is better than *Sheridan*.

4.2. Kinase Split. This split tests the ability of the model to predict the binding affinity for an unseen protein kinase. Since it induces high heterogeneity across each fold/chunk of data, care has to be taken in drawing conclusions, especially from the test data results. The results for the KNN and the BiMCA on the validation and test data are shown in Figures 2A and B, respectively.

On the validation data, no clear trend is visible when comparing the three active site configurations across models, data splits, and metrics. Notably, however, *all* active site definitions significantly outperform the full sequence representations across all models, splits, and metrics. While the *Sheridan* representation is significantly superior to the *Martin* representation for the KNN ($p < 0.05$, $W+$) and to the *Combined* representation for the BiMCA, this trend does not persist in the test data. During testing, the *Martin* representation consistently obtained the highest Pearson correlation, irrespective of the model (cf. Table S1). However, this finding does not corroborate when using the RMSE as a response metric (cf. Figure 2B). Notably, our best model (the pretrained BiMCA) obtained the best performance with the *Combined* representation in all but one case.

In Supporting Information S2.3, we report additional results on a subset of samples where both kinases *and* ligands are unseen. The results on this strict split evidence the higher generalization capabilities of the BiMCA compared to the KNN and underline the superiority of the active site sequence representations.

Kinase Sequence Augmentation. To further improve performance, we systematically investigated different kinase sequence augmentation strategies. The results demonstrate that all augmentation techniques improved model performance (cf. Table 2). Interestingly, the structure-motivated techniques of swapping (SS) and flipping subsequences (FS) exhibited a similar performance boost to simple flipping (F). However, the benefit of flipping is statistically insignificant, whereas FS and SS yield significant benefits ($p < 0.01$, $W+$) in several configurations. Moreover, their performance boost is roughly additive as combining all three strategies yields the best results in seven out of eight cases ($p < 0.01$, $W+$, RMSE on validation data). We hypothesize that the pretrained model is harder to improve because it partially learned invariance to the applied transformations.

5. DISCUSSION

In this work, we corroborate the finding that "less is more" in sequence-based kinase-ligand affinity prediction models.

Table 2. Results of Sequence Augmentation (Kinase Split)[a]

		RMSE (↓)		Pearson (↑)	
data	augmentation	BiMCA	BiMCA-pre	BiMCA	BiMCA-pre
val.	none	$1.32_{\pm0.16}$	$1.20_{\pm0.12}$	$0.438_{\pm0.08}$	$0.489_{\pm0.09}$
	flip (F)	$1.25_{\pm0.13}$	$1.19_{\pm0.13}$	$0.463_{\pm0.08}$	$0.502_{\pm0.08}$
	flip subseq (FS)	$1.28_{\pm0.12}$	$\mathbf{1.18}_{\pm0.12}$	$0.431_{\pm0.11}$	$\mathbf{0.521}_{\pm0.08}$
	swap subseq (SS)	$1.28_{\pm0.17}$	$\mathbf{1.18}_{\pm0.12}$	$0.443_{\pm0.11}$	$0.511_{\pm0.09}$
	FS + SS	$1.27_{\pm0.11}$	$\mathbf{1.18}_{\pm0.12}$	$0.444_{\pm0.09}$	$0.508_{\pm0.09}$
	F + FS + SS	$\mathbf{1.22}_{\pm0.10}$	$\mathbf{1.18}_{\pm0.11}$	$\mathbf{0.468}_{\pm0.11}$	$0.505_{\pm0.09}$
test	none	$1.33_{\pm0.08}$	$1.23_{\pm0.08}$	$0.431_{\pm0.06}$	$0.505_{\pm0.07}$
	flip (F)	$1.28_{\pm0.05}$	$1.23_{\pm0.07}$	$0.478_{\pm0.04}$	$0.515_{\pm0.06}$
	flip subseq (FS)	$1.32_{\pm0.09}$	$1.22_{\pm0.04}$	$0.444_{\pm0.08}$	$0.516_{\pm0.04}$
	swap subseq (SS)	$1.28_{\pm0.04}$	$1.23_{\pm0.03}$	$\mathbf{0.479}_{\pm0.01}$	$0.506_{\pm0.06}$
	FS + SS	$1.29_{\pm0.06}$	$1.22_{\pm0.07}$	$0.469_{\pm0.04}$	$0.526_{\pm0.05}$
	F + FS + SS	$\mathbf{1.27}_{\pm0.06}$	$\mathbf{1.21}_{\pm0.05}$	$\mathbf{0.479}_{\pm0.06}$	$\mathbf{0.531}_{\pm0.05}$

[a] All models were used the *Combined* active site definition.

Our experiments show that the 16 residues identified by ref 10 yield similar results to the *Sheridan* residues.

We report evidence that a novel, *Combined* kinase representation is superior to the *Sheridan* and the *Martin* representation for predicting binding affinity in unseen ligands. To predict unseen kinases, we not only corroborate our previous results on the superiority of active sites to full kinase sequences but also find that no residue composition is strictly advantageous. While we have previously found that incorporating *fewer* residues yields better results, we find here that bringing back specific residues that are more distant from the the ATP pocket significantly increases performance, especially for allosteric binders. Although these residues (cf. subsection S1.3) were identified algorithmically, Martin et al.[10] discuss in a post hoc analysis dynamical roles of residues in the "hydrophobic spine"[14] as well as other residues important in loop dynamics and activation−deactivation mechanisms of kinases that do not interact directly with the ligand (T51, L103, V119, G126, I163). Other residues might be involved in both direct and indirect interactions (F54, L95, L106, F187, L162).

Even though the ideal set of residues for sequence-based kinase affinity prediction models remains unclear, our results are a step forward in compactly modeling kinase-ligand binding. As shown in our previous work,[7] improved affinity predictors can be leveraged to drive molecular generative models toward generating molecules with higher binding affinity to specific kinases. Lastly, the knowledge about the location of the active site motivates multiple novel sequence augmentation techniques that demonstrated further, complementary performance improvement.

6. DATA AND SOFTWARE AVAILABILITY

The data processing and augmentation pipelines are implemented in the `pytoda` package.[12] The source code has been released on https://github.com/PaccMann/paccmann_kinase_binding_residues#choosing-active-site-sequences, and the preprocessed BindingDB data is available via https://ibm.biz/active_site_data.[7]

Moreover, in the Generative Toolkit for Scientific Discovery (GT4SD), we provide an example on leveraging the affinity predictor as a reward function in a protein-driven molecular generative model:[15] https://github.com/GT4SD/gt4sd-core/tree/main/examples/protein_driven_molecule_generation.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00840.

Additional details about data processing and model implementations and hyperparameters as well as continued results for affinity prediction including strict split (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jannis Born − *Accelerated Discovery, IBM Research Europe, 8803 Rüschlikon, Switzerland; Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland;* orcid.org/0000-0001-8307-5670; Email: jab@zurich.ibm.com

Authors

Yoel Shoshan − *IBM Research Haifa, Haifa 31905, Israel*

Tien Huynh − *IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, United States*

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

Eric J. Martin − *Novartis Institutes for BioMedical Research, Emeryville, California 94608, United States;* orcid.org/0000-0001-7040-5108

Matteo Manica − *Accelerated Discovery, IBM Research Europe, 8803 Rüschlikon, Switzerland;* orcid.org/0000-0002-8872-0269

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00840

Funding

The authors received no external funding.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Abbasi, K.; Razzaghi, P.; Poso, A.; Ghanbari-Ara, S.; Masoudi-Nejad, A. Deep learning in drug target interaction prediction: current and future perspectives. *Curr. Med. Chem.* **2021**, *28*, 2100−2113.

(2) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E.

Improved protein−ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **2021**, *61*, 1583−1592.

(3) 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.

(4) 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*, 308−322.

(5) 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.

(6) Volkov, M.; Turk, J.-A.; Drizard, N.; Martin, N.; Hoffmann, B.; Gaston-Mathé, Y.; Rognan, D. On the frustration to predict binding affinities from protein−ligand structures with deep neural networks. *J. Med. Chem.* **2022**, *65*, 7946.

(7) Born, J.; Huynh, T.; Stroobants, A.; Cornell, W. D.; Manica, M. Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction And Inhibitor Generation: 3D Effects in a 1D Model. *J. Chem. Inf. Model.* **2022**, *62*, 240−257.

(8) 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.

(9) Wu, P.-K.; Park, J.-I. MEK1/2 inhibitors: molecular activity and resistance mechanisms. *Seminars in oncology* **2015**, *42*, 849−862.

(10) 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.

(11) 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.

(12) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakarajan, N.; Cardinale, A.; Laino, T.; Martinez, M. R. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Machine Learning: Science and Technology* **2021**, *2*, 025024.

(13) Shaw, A. S.; Kornev, A. P.; Hu, J.; Ahuja, L. G.; Taylor, S. S. Kinases and pseudokinases: lessons from RAF. *Mol. Cell. Biol.* **2014**, *34*, 1538−1546.

(14) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Ten Eyck, L. F. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 17783−17788.

(15) Manica, M.; Cadow, J.; Christofidellis, D.; Dave, A.; Born, J.; Clarke, D.; Teukam, Y. G. N.; Hoffman, S. C.; Buchan, M.; Chenthamarakshan, V.; Donovan, T.; Hsu, H. H.; Zipoli, F.; Schilter, O.; Giannone, G.; Kishimoto, A.; Hamada, L.; Padhi, I.; Wehden, K.; McHugh, L.; Khrabrov, A.; Das, P.; Takeda, S.; Smith, J. R. GT4SD: Generative Toolkit for Scientific Discovery. 2022. *arXiv preprint*. https://arxiv.org/abs/2207.03928 (accessed 2022-09-11), DOI: 10.48550/arxiv.2207.03928.