Kinase similarity assessment pipeline for off-target prediction [Article v1.0]

Talia B. Kimber[1†], **Dominique Sydow**[1†‡], **Andrea Volkamer**[1*]

[1] *In silico* Toxicology and Structural Bioinformatics, Institute of Physiology, Charité-Universitätsmedizin Berlin, Charitéplatz 1, 10117, Berlin, Germany

This LiveCoMS document is maintained online on GitHub at https://github.com/volkamerlab/kinase_similarity_pipeline_paper; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.

This version dated June 6, 2022

Abstract Kinases are established drug targets to combat cancer and inflammatory diseases. Despite decades of kinase research, challenges still remain, such as the under-exploration of a large fraction of the kinome and the promiscuous binding of many kinase inhibitors. Due to the highly conserved orthosteric ATP binding site in kinases, ligands may bind not only to their designated kinase (on-target) but also to other kinases (off-targets). Such promiscuous binding can cause mild to severe side effects, and the prediction of these off-targets is highly non-trivial. Therefore, we propose a pipeline that allows the study of kinase similarities from four different angles in an automated and modular fashion. The first method considers the binding site sequence. The second method uses structural information via KiSSim, a newly developed fingerprint that considers both physico-chemical and spatial properties of the binding site. The third method involves kinase-ligand interaction fingerprints as provided by KLIFS, and the last method utilizes the measured activity of ligands on kinases based on ChEMBL data. Finally, results for a given set of kinases are collected and analyzed to gain insight into potential off-targets from the different aforementioned perspectives. Since the pipeline is set up as a series of Jupyter notebooks covering both theoretical and practical aspects, the target audience ranges from beginners to advanced users working in the field of natural and computer sciences. The pipeline is part of the TeachOpenCADD project and extends it with this special kinase edition. All code is free, open-source, and made available at https://github.com/volkamerlab/teachopencadd .

***For correspondence:**
andrea.volkamer@charite.de (AV)

[†]These authors contributed equally to this work

Present address: [‡]Sosei Heptares, Steinmetz Building, Granta Park, Cambridge CB21 6DG, United Kingdom

1 Introduction

Kinases are involved in most cellular processes by phosphorylating—and thereby activating—themselves or other proteins. This family is among the most frequently mutated proteins in tumors and has been successfully studied as drug targets for many decades [1]. Thanks to the longstanding

research, a plethora of kinase data is freely available, i.e. as part of databases such as UniProt [2], PDB [3] or ChEMBL [4], and has been made easily accessible via kinase resources such as the KLIFS—Kinase-Ligand Interaction Fingerprints and Structures—database [5]. As of February 2022, 5 911 X-ray structures of human kinases have been resolved (see the KLIFS database [6]) and 70 FDA-approved small molecule protein kinase inhibitors are on the market [7]. Most of the approved drugs bind in the ATP-binding pocket and intermediate surroundings (orthosteric binding site).

Although structural information provides rich information, kinases have been widely classified based on sequence. Manning et al. [8] clustered the human protein kinases based on their sequence similarity into eight major groups (AGC, CAMK, CK1, CMGC, RGC, STE, TK, TKL) and one "Other" group for unassigned kinases, as well as atypical kinases. The resulting Manning kinome tree depicts kinase clustering (see Figure 1).

Despite decades of kinase research, challenges still remain [9]. For example:

1. A large fraction of the kinome is un-/underexplored. Figure 1a shows the number of PDB structures per kinase, unveiling a vast imbalance between structurally resolved kinases and unexplored ones. For example, CDK2 has been resolved in 426 PDB structures, while only 313 kinases [6] out of approximately 540 in the kinome [9] have been structurally resolved.
2. Many kinase inhibitors are promiscuous binders causing off-target effects or enabling polypharmacology [1, 10]. For example, the Epidermal Growth Factor Receptor (EGFR) inhibitor erlotinib shows affinities to other kinases in the highly sequentially-similar TK kinase group, but also strongly affects off-targets in more remote kinase groups (see Figure1b).

Therefore, assessing kinase similarity from different angles may be a crucial step in understanding and predicting off-targets to help designing more selective drugs and avoiding side effects.

1.1 Scope

In this study, similarities between a set of kinases are investigated based on methods offering different perspectives on this challenging topic with a focus on orthosteric binding sites (here referred to as binding sites), as summarized in Table 1. The first method considers the binding site sequence as deposited in the KLIFS database. The second method uses KiSSim [11], a recently developed fingerprint that considers physico-chemical as well as spatial properties of the binding site. The third method involves protein-ligand interaction fingerprints as provided in the KLIFS database, and the last method utilizes the measured activity of ligands against kinases based on ChEMBL data [4]. The different methods are preceded by a general introduction to kinases and the challenges faced during kinase-centric drug design, and succeeded by a comparison between the different kinase similarity methods.

Please note that this study focuses on the similarities between ATP binding sites. Therefore, kinase polypharmacology and off-targets can only be assessed within the scope of orthosteric binding sites, even though the promiscuity of some ligands may be explained by binding to allosteric binding sites (potential allosteric binding sites are summarized in the Kinase Atlas [12]).

This study has been put together into a modular pipeline that enables the research of kinase similarities in an automated fashion, allowing users to simply use it out of the box, or adapt it to their needs.

This workflow is integrated in the context of TeachOpenCADD [15, 16], a teaching platform for computer-aided drug design (CADD) using open-source packages and data. Specific tasks in cheminformatics and structural bioinformatic are described and solved using Python-based Jupyter notebooks [17] as interacting platform. All code has been deposited on GitHub, see https://github.com/volkamerlab/teachopencadd. And the project website can be found at this link, https://projects.volkamerlab.org/teachopencadd/.

2 Prerequisites

2.1 Target audience

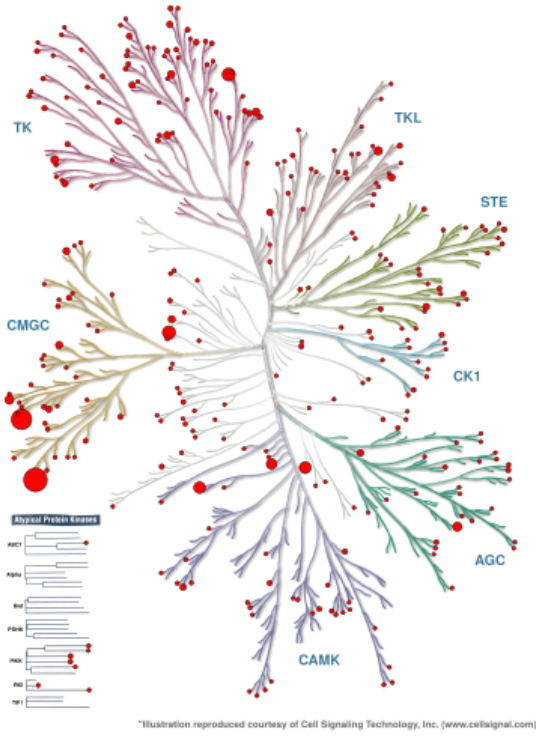
The notebooks were developed to support researchers interested in kinase-centric computational drug design with a focus on understanding and predicting kinase off-targets. As this collection is part of the TeachOpenCADD [15, 16] training material, we also recommend the notebooks to teachers as pedagogical interactive material in structural bioinformatics and cheminformatics.

2.2 Background knowledge

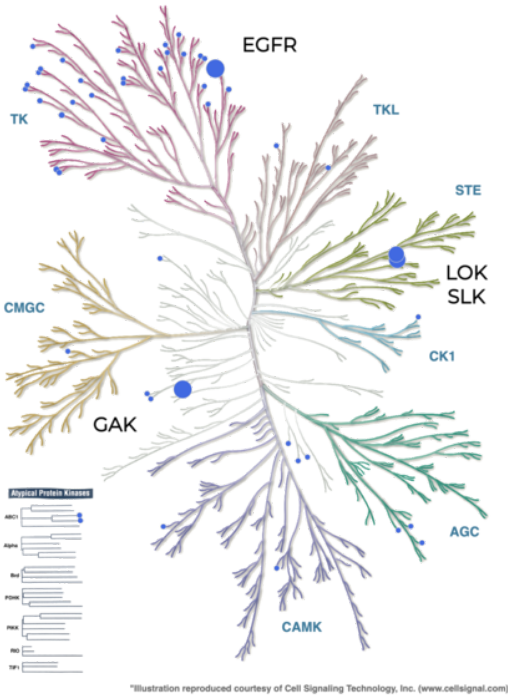
The notebooks are constructed in a way that no in depth prior knowledge besides an affinity for the natural or computer sciences is required. Each notebook eases into the topic of kinase drug development and kinase similarity with a lot of theoretical background and comments on all content as well as programming-related steps in great detail. Nevertheless, users will benefit from a basic understanding of the Python programming language and the usage of Jupyter notebooks. If such basic introduction is needed, please refer to training material as listed on the TeachOpenCADD website [18].

Topic	Description	Hyperlink
What is a kinase?	Introduction to kinases and challenges in drug discovery.	T023
Pocket sequence	Pairwise similarities/identities between 85 residue long KLIFS pocket sequences.	T024
Pocket structure	Pairwise similarities between 1032–bit long KiSSim fingerprints, which encode spatial and physico-chemical pocket properties.	T025
Pocket-ligand interactions	Pairwise similarities between 595–bit long KLIFS kinase-ligand interaction fingerprints (IFP).	T026
Ligand profile	Similarity based on the ratio of compounds tested active against kinase pairs.	T027
Kinase similarity comparison	Comparison of predicted off-targets based on calculated kinase similarities using aforementioned methods.	T028

Table 1. TeachOpenCADD kinase edition overview: Notebook topics, description, and index with a hyperlink to the associated notebook.



(a) Number of PDB structures per kinase. The figure shows the imbalance between highly explored kinases, e.g. groups: TK, CMGC, and CAMK, and less explored ones, e.g. the CK1 group. CDK2 has the most structures, with 426. The red circle is proportional to the number of PDB structures for each kinase, such that the greater is the circle, the higher is the number of structures.



(b) Developing selective kinase inhibitors is non-trivial since kinases are highly conserved in the ATP binding site. EGFR inhibitor erlotinib binds not only to its intended target EGFR, but also to kinases in remote groups, such as SLK/LOK in the STE group and GAK in the CMGC group. The blue circle is proportional to the K_d value in nM taken from the Karaman et al. [13] dataset.

Figure 1. Visual representation using the Manning tree of existing challenges in kinase research: un-/underexplored kinase groups (left) and the promiscuity character of kinases (right). The figure is taken from https://projects.volkamerlab.org/teachopencadd/talktorials/T023_what_is_a_kinase.html and is generated using KinMap [14].

2.3 Software requirements

The notebooks are written in Python and rely on open-source packages such as pandas [19], numpy [20], scipy [21], matplotlib [22], seaborn [23], scikit-learn [24], rdkit [25], biotite [26], opencadd [27], requests [28], and pillow [29].

The user only needs to install the *teachopencadd* conda-forge package [30] (see installation [31]), which will install all relevant packages and save a copy of all TeachOpenCADD notebooks—including the kinase edition discussed in this paper—on the user's local machine. A read-only mode of the notebooks is accessible via the TeachOpenCADD website at https://projects.volkamerlab.org/teachopencadd/. Online execution can be done via Binder [32], using the following link https://mybinder.org/v2/gh/volkamerlab/TeachOpenCADD/master.

3 Method

In this section, the four methods that are introduced to quantify kinase similarity are described, namely the pocket sequence, the KiSSim fingerprint, the interaction fingerprint, and the ligand profile. Note that the theoretical and practical aspects of each method are also covered in great detail in the individual notebooks of this kinase collection (Table 1). As discussed in the "Scope" section of this manuscript, we focus on kinase similarities based on orthosteric binding sites.

3.1 Pocket sequence

The full amino acid sequence is often used to assess similarities between kinases (see the phylogenetic tree developed by Manning et al. [8]). Since binding sites are often more conserved than the whole protein, van Linden et al. [33] defined as part of KLIFS a 85-long pocket sequence that is aligned across the kinome. Using a sequence that focuses on the binding site seems appropriate in the case of kinases, since this is where the ligand is likely to bind. Moreover, working with a fixed length sequence is practical from a computational point of view.

In this study, two methods are used to compute relationships based on sequence, namely the sequence identity and the sequence similarity, which are described below.

3.1.1 Sequence identity

The pairwise sequence identity, or simply sequence identity, is a similarity based on character-wise discrepancy, in other terms, the number of residues that match in two aligned sequences [34]. More formally, given two kinase sequences S and S' of same lengths L, the sequence identity can be defined as

$$\text{sequence identity}(S, S') = \frac{1}{L}\sum_{n=1}^{L} I\left(S[n], S'[n]\right), \qquad (1a)$$

where I is the identity matrix of the amino acids, and $S[n]$ the amino acid at position n of the kinase sequence S. Note that not all kinases have residues present at each of the 85 alignment positions. Such gaps are represented by "-" and count as mismatch to any amino acid.

3.1.2 Sequence similarity

Unlike sequence identity which treats all residues uniformly, pairwise sequence similarity, or sequence similarity, takes into account the change of the amino acids over evolutionary time, thus reflecting relationships between amino acids. It is based on a substitution matrix M, where each entry gives a score between two amino acids. In this study, the BLOSUM substitution matrix [35], as implemented in biotite [36], is used. Formally, the following is defined:

$$\text{sequence similarity}(S, S') = \frac{1}{L}\sum_{n=1}^{L} M'\left(S[n], S'[n]\right), \qquad (1b)$$

where M' is the translated and rescaled version of the substitution matrix M.

For both the sequence identity and similarity, the closer the value is to 1, the more similar are the kinases.

Figure 2 shows the sequence similarity between the KLIFS pocket sequence of EGFR and MET kinases. Sequence similarity is used by default in the pipeline for further analysis.

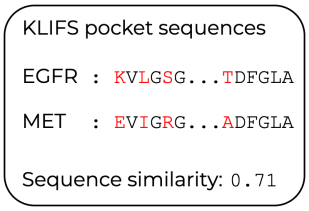


Figure 2. Sequence similarity between EGFR and MET. The 85 residue pocket sequence is retrieved from KLIFS. The pairwise sequence similarity takes into account the change of the amino acids over evolutionary time.

3.2 The KiSSim fingerprint

In order to assess the pairwise similarity of kinases from a structural point of view, the newly developed KiSSim (**Ki**nase **S**tructure **Sim**ilarity) fingerprint [11, 37] is used. This fingerprint describes the physico-chemical and spatial properties of structurally resolved kinases, while focusing on the KLIFS pocket residues. Each structure is mapped to a fingerprint

composed of 1032 bits, the first 680 (= 85 × 8) describing physico-chemical features and the remaining 352 (= 85 × 4 + 12) spatial information (see Figure 3).

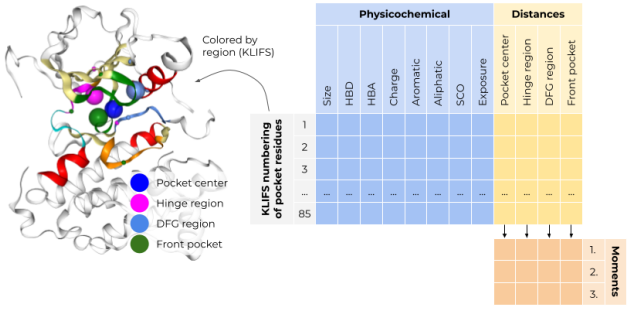


Figure 3. The 1032-long KiSSim fingerprint encodes physico-chemical and spatial properties of the kinase's pocket, adding a structural perspective on kinases. The figure is adapted from [37].

3.2.1 From several structures to one kinase

A kinase can be represented by one or even a hundred resolved crystal structures in the PDB (see Figure 1a). In this study, we aim at comparing different kinases and not individual structures. Since KiSSim generates a fingerprint for each structure, the following mapping from structures to kinase is applied:

Given two kinases K and K', all available structures in KLIFS for these kinases are fetched using opencadd [27], namely $s_1, \ldots, s_m$ for kinase K, and $s'_1, \ldots, s'_n$ for kinase K', noting that the number of structures might be different for each kinase. Each structure s_i, s'_i is then mapped to its corresponding KiSSim fingerprint fp_i, fp'_i, see Figure 4. The fingerprints fp, fp' corresponding to kinases K, K' respectively, are the ones for which the Euclidean distance is minimized (Figure 4). Note that these *minimal distance* fingerprints vary for each kinase depending on the compared K, K' pair.

Finally, two kinases K, K' are compared based on their respective *minimal distance* between KiSSim fingerprint fp, fp' using the Euclidean norm:

$$\text{KiSSim dissimilarity (fp, fp')} = \|\text{fp} - \text{fp}'\|_2 . \qquad (2)$$

In this case, the closer the value to 0, the more similar the kinases.

3.3 The interaction fingerprint

Interaction fingerprints (IFPs) encode the binding mode of a ligand in a binding site, i.e., the protein-ligand interactions that are present in a structurally resolved complex. If a ligand can form similar interaction patterns in proteins other than its designated protein (off- vs. on-target), it is possible

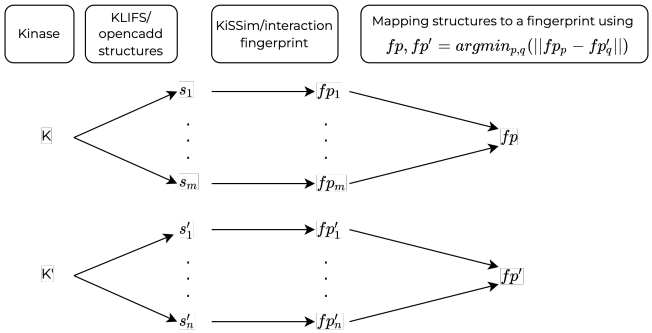


Figure 4. Associating one structural fingerprint per kinase. All available structures are retrieved for two given kinases and all fingerprints are computed. The fingerprints selected to be associated with the kinase in the present kinase pair are the ones for which the computed distance is minimized.

that this ligand will cause unintended side effects. Knowledge about binding mode similarities can therefore help to avoid such off-target effects.

The KLIFS interaction fingerprint describes seven possible interactions for each of the 85 residues in the binding pocket. Interactions include 1. hydrophobic contacts, 2. aromatic interactions, face to face, 3. aromatic interactions, edge to face, 4. H-bond donors, 5. H-bond acceptors, 6. cationic interactions, and 7. anionic interactions. The 595-bit long vector describes the presence or absence of such interactions for all 85 residues (see Figure 5).

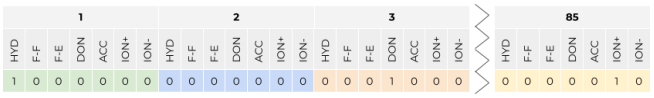


Figure 5. The KLIFS interaction fingerprint encodes seven interaction types for each of the 85 residues in the binding site. Interaction types include: hydrophobic contacts (HYD), face to face aromatic interactions (F-F), face to edge aromatic interactions (F-E), protein H-bond donors (DON), protein H-bond acceptors (ACC), protein cationic interactions (ION+), and protein anionic interactions (ION-). The figure is taken from [38].

Similarly to the KiSSim comparison, given two kinases K and K', all available structures in KLIFS for these kinases are fetched using opencadd [27]. Each structure is mapped to its corresponding IFP. The interaction fingerprints fp, fp' corresponding to kinases K, K' respectively are the ones for which the Jaccard distance [39] is minimized (Figure 4). Note that the Euclidean distance is used in case of the KiSSim fingerprint, which contains continuous and discrete values, while the Jaccard distance is employed in case of the binary IFPs.

Finally, two kinases K, K' are compared using their respective *minimal distance* between interaction fingerprint fp, fp' and calculating the Jaccard distance:

$$\text{IFP dissimilarity (fp, fp')} = d_J(\text{fp, fp'}), \quad (3)$$

where d_J is the Jaccard distance.

In this case, the closer the value to 0, the more similar the kinases.

3.4 Ligand profile

In the context of drug design, the following assumption is often made: if a compound was tested active on two different kinases, it is suspected that these two kinases may have some degree of similarity [40]. This is the rationale behind the ligand profile similarity. Given bioactivity data for a set of compounds measured against a set of targets—in this case kinases—and two kinases K, K', ligand profile similarity is defined as

$$\text{lig. profile similarity(K, K')} = \frac{\text{\# actives on both K and K'}}{\text{\# tested on both K and K'}}. \quad (4)$$

The closer the value is to 1, the more similar are the kinases. If no compounds were commonly tested on two kinases, then the similarity is set to 0. Computing the similarity between a kinase and itself may be interpreted as kinase promiscuity, where the similarity described above would therefore represent the fraction of active compounds over all tested compounds for this kinase.

3.4.1 Bioactivity data

The bioactivity data used for this method comes from Kinodata [41], from the Openkinome organization [42]. It is a preprocessed kinase subset of the ChEMBL data [4], version 29. Further processing includes keeping only IC_{50} values given in nM, and converted to pIC_{50}s. If there are several measurements for a kinase-compound pair, then the most active value, i.e. the entry with the highest pIC_{50}, value is kept. Finally, the pIC_{50} values are binarized using a 6.3 cutoff to discriminate between an active or inactive compound as described in [43].

In the pipeline, one can additionally compute the non-reduced ratio of number of active compound against the total number of compounds to gain insight into the actual number of measurements for each kinase pair.

3.5 Kinase comparison and clustering

To assess kinase similarities based on the calculated (dis)similarity matrices, two visualization methods are used, namely heatmaps and dendrograms.

3.5.1 Heatmaps

The heatmaps are generated using Matplotlib [22] to depict the similarity between a set of kinases. The maximum value is 1, indicating exact similarity, as is the case for diagonal entries. The value 0 indicates total dissimilarity. Plotting such figures allows to see and extract patterns thanks to the gradient of colors, see top row in Figure 6.

3.5.2 Dendrograms

Clustering algorithms are used to identify groups such that the similarities within clusters are higher than compared to other clusters [44]. In this study, hierarchical clustering is used, and, unlike heatmaps, it is based on distance (or dissimilarity). Hierarchical clustering can be graphically displayed using a dendrogram (see bottom row in Figure 6), where the height of each node is proportional to the dissimilarity between its two daughter clusters. The clustering and plotting is done using Scikit-learn [24] and Matplotlib [22], respectively.

For fair comparison, the distance matrices for all four methods are normalized so that each entry lives between 0 and 1. Similarity matrices—as used for the heatmaps—are then computed using 1–distance matrix. Contrary to the dendrograms, that use the distance matrix.

4 Pipeline

Measuring kinase similarity is a non-trivial task; distinct measures can provide different insights, which can be complementary, confirmatory, or contradictory, and therefore expand our knowledge on the target(s) at hand. However, implementing multiple methods can be time-consuming and comparing results across many output types can be laborious. Turning such processes into a functional pipeline helps to avoid the scattering of scripts and to speed up iterations of the design-make-test-analyze cycle [45] of drug design campaigns. Moreover, following the findable, accessible, interoperable, and reusable (FAIR) principles [46] makes such pipelines long-lasting and available to the community.

In the pipeline presented herein, we implement the different methods once and streamline each method's results into a standardized output with a pre-defined set of visualization tools for easy comparison, while leaving the pipeline flexible enough so that adding new methods or new visualization tools is effortless, making the whole process easy to understand, maintain, and expand.

4.1 Means of the pipeline

The proposed pipeline is a collection of six Jupyter notebooks [17] that allows the study of kinase similarities from four different angles in an automated and modular fashion (Figure 7).

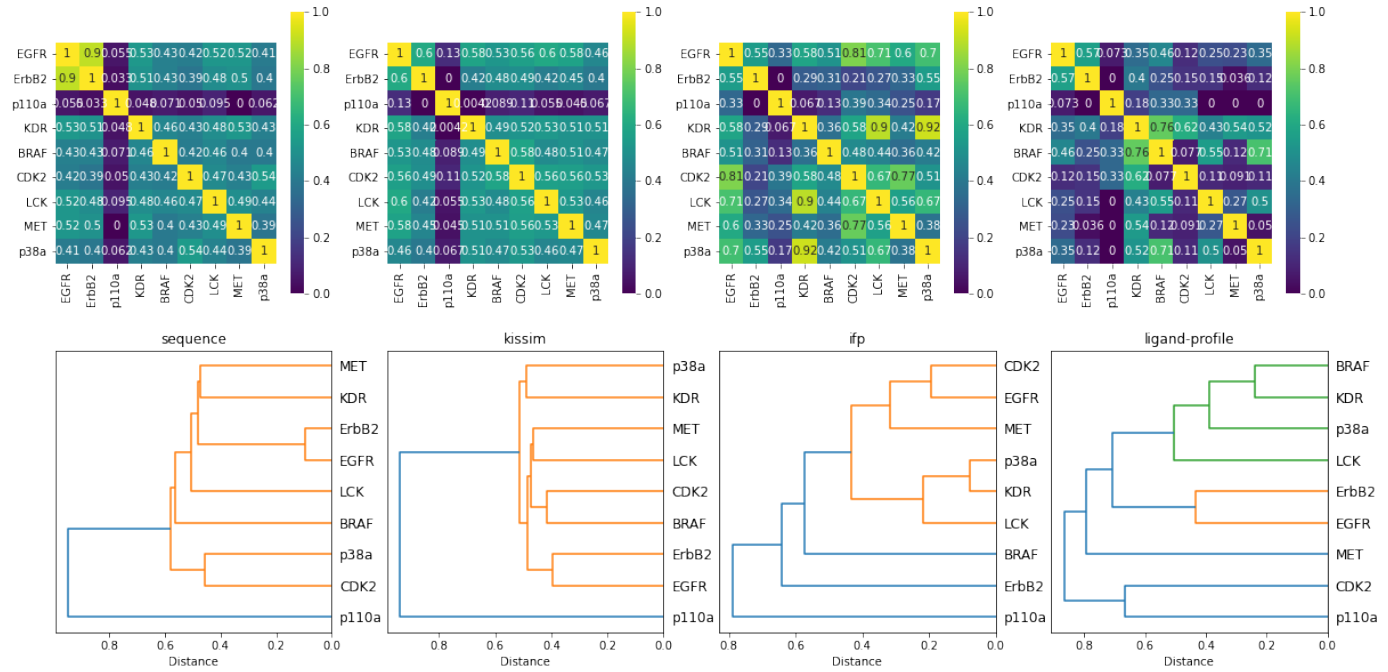


Figure 6. Visualization of kinase similarity from four different angles: sequence, KiSSim, interaction fingerprint (ifp) as well as ligand profile. The top, bottom row shows four heatmaps, dendrograms respectively for a set of nine study kinases.

4.2 Structure of the notebooks

The structure of all notebooks is as follows: the first section covers the theory written in Markdown and summarizes the necessary concepts to understand the task. Relevant references are also mentioned. The second part of a notebook deals with the actual implementation of the task in a pedagogical manner, including motivation for practical steps and detailed comments on coding decisions. Finally, a discussion and a quiz section wrap up the notebook. This structure is very well suited from a teaching perspective, since it contains both theory and hands on programming. The notebook can easily be used as a medium for a presentation, and it allows for self-study and usage in own research projects.

4.3 About the code

The programming part is done in Python exclusively and the code follows the latest software best practices. It is written pythonically and contains lots of code comments. Thanks to the continuous integration (CI), all outputs and results are fully reproducible and the pipeline's maintenance is facilitated.

4.4 Content of the pipeline

As mentioned previously, the proposed pipeline contains six notebooks, described below:

The first notebook sets the stage with a kinase introduction and references/tools on where to find kinase-related information. It is also in this first notebook that a set of kinases of interest is defined. In this study, nine kinases are selected, the same nine as in the paper by Schmidt et al. [47], where the authors discussed the challenges and advantages of tackling kinase similarities from multiple perspectives. Table 2 summarizes the information used for these kinases. The pipeline can be executed out-of-the box with the defined set of kinases, but it can equally be run with a different user defined set of kinases. The only condition is that the uploaded CSV file with the kinases of interest contains two mandatory columns, namely `kinase_klifs`, which is the KLIFS name of the kinase, and `uniprot_id`, the Uniprot identifier (ID) [2] of the kinase (Figure 7).

The four following notebooks describe one similarity method at a time as discussed in Section 3: the pocket sequence, the KiSSim fingerprint, the interaction fingerprint, and the ligand profile.

The final notebook collects the information from the previous ones and compares the different perspectives with easy-to-understand visualization such as heatmaps and dendrograms (see Section 3.5). Additionally, an equally weighted average to combine distance and similarity matrices of all four perspectives can be computed, yielding a single heatmap, and a single dendrogram. The user can easily extend this to a knowledge-informed weighted scheme based on their own research focus.

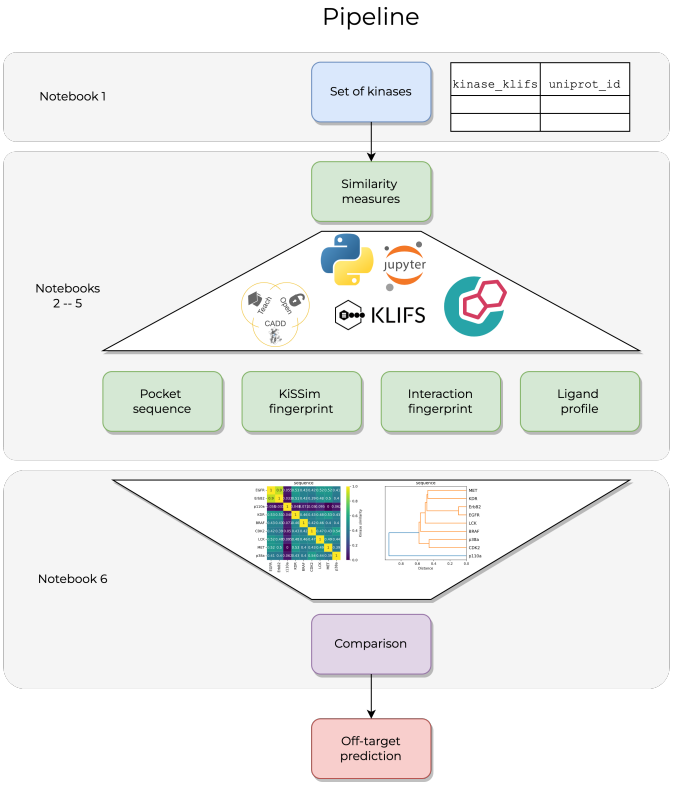


Figure 7. The proposed pipeline consists of six Jupyter notebooks [17]. Given a set of kinases in a CSV format, four similarity measures are implemented, and kinases are compared using heatmaps and dendrograms. The project is part of TeachOpenCADD [15, 16] and uses open-source tools and databases such as KLIFS [5] and ChEMBL [4].

4.5 Features of the pipeline

The developed pipeline contains many useful features. Firstly, it is part of the TeachOpenCADD project [15, 16] and extends it with this special kinase edition. Being part of TeachOpenCADD has the following advantages:

1. TeachOpenCADD is open-source and freely available at https://github.com/volkamerlab/teachopencadd, the project being licensed under the Attribution 4.0 International (CC BY 4.0).
2. A dedicated conda package [48] facilitates installation.
3. Online execution is possible via the Binder project [32].
4. The teaching approach makes the notebooks easy to follow.

Moreover, the pipeline is easily adaptable to new sets of kinases defined by an external user, as well as new similarity methods.

5 Conclusion

In this study, a full pipeline for the assessment of kinase similarity is presented, using four methods of comparison. The pipeline is composed of six Jupyter notebooks:

1. An introduction to kinases and their central role in drug discovery, as well as collecting the kinase set for the downstream notebooks.
2. The similarity from a pocket sequence point of view.
3. The similarity based on the KiSSim fingerprint, which encodes physico-chemical and spatial properties of the kinase pocket.
4. The similarity based on KLIFS interaction fingerprints between the kinase pocket residues and a co-crystallized ligand.
5. The similarity based on ligand profiling data collected from ChEMBL, measuring a compound's activity on a kinase.
6. An analysis notebook which collects the proximity matrices calculated for the four methods, visualizes the similarities with heatmaps and the clusters with dendrograms, and finally discusses the results.

We encourage users to develop their own similarity methods and to contribute to the existing pipeline.

This paper could be of interest to

1. researchers who want to gain insights into off-target prediction and kinase similarities, and integrate their new comparison methods to a working workflow,
2. beginners in software development who need inspiration to set up a fully functional pipeline,
3. teachers who want a starting point for lecture material,
4. students with a background in bioinformatics, cheminformatics, and the life sciences in general,
5. anyone who is curious.

Author Contributions

Conceptualization: TBK, DS, AV; Methodology: TBK, DS, AV; Software: TBK, DS, AV; Validation: TBK, DS, AV; Formal Analysis: TBK, DS, AV; Investigation: TBK, DS, AV; Writing – Original Draft: TBK, DS, AV; Writing – Review & Editing: TBK, DS, AV; Visualization: TBK, DS, AV; Project Administration: TBK, DS, AV; Funding Acquisition, Supervision: AV.

For a more detailed description of author contributions, see the GitHub issue tracking and changelog at https://github.com/volkamerlab/kinase_similarity_pipeline_paper.

Potentially Conflicting Interests

The authors declare no conflict of interests.

Abbreviations

List of abbreviations used in the paper.

kinase	**kinase_klifs**	**uniprot_id**	group	full kinase name
EGFR	EGFR	P00533	TK	Epidermal growth factor receptor
ErbB2	ErbB2	P04626	TK	Erythroblastic leukemia viral oncogene homolog 2
PI3K	p110a	P42336	Atypical	Phosphatidylinositol-3-kinase
VEGFR2	KDR	P35968	TK	Vascular endothelial growth factor receptor 2
BRAF	BRAF	P15056	TKL	Rapidly accelerated fibrosarcoma isoform B
CDK2	CDK2	P24941	CMGC	Cyclic-dependent kinase 2
LCK	LCK	P06239	TK	Lymphocyte-specific protein tyrosine kinase
MET	MET	P08581	TK	Mesenchymal-epithelial transition factor
p38a	p38a	Q16539	CMGC	p38 mitogen activated protein kinase alpha

Table 2. Set of defined kinases. The table lists the kinases used in the pipeline, the same nine as in the study by Schmidt et al. [47]. It is noteworthy that the pipeline is applicable to an arbitrary set of kinases, the only condition being that the input CSV file should contain two columns, `kinase_klifs` and `uniprot_id`, displayed in bold.

KLIFS	Kinase-Ligand Interaction Fingerprints and Structures
EGFR	Epidermal Growth Factor Receptor
KiSSim	Kinase Structure Similarity
IFP	Interaction Fingerprint
ID	Identifier
CI	Continuous Integration

Funding Information

TBK received funding from the Stiftung Charité in the context of the Einstein BIH Visiting Fellow Project, DS from the Deutsche Forschungsgemeinschaft (grant VO 2353/1-1), and AV from the Bundesministerium für Bildung und Forschung (grant number 031A262C).

Author Information

ORCID:

Talia B. Kimber: 0000-0002-8881-920X
Dominique Sydow: 0000-0003-4205-8705
Andrea Volkamer: 0000-0002-3760-580X

References

[1] **Cohen P**, Cross D, Jänne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. Nature Reviews Drug Discovery. 2021; 20(7):551–569. https://doi.org/10.1038/s41573-021-00195-4.

[2] **Consortium TU**. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2020; 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100.

[3] **Berman HM**, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research. 2000; 28(1):235–242. https://doi.org/10.1093/nar/28.1.235.

[4] **Gaulton A**, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Research. 2016; 45(D1):D945–D954. https://doi.org/10.1093/nar/gkw1074.

[5] **Kanev GK**, de Graaf C, Westerman BA, de Esch IJP, Kooistra AJ. KLIFS: an overhaul after the first 5 years of supporting kinase research. Nucleic Acids Research. 2020; 49(D1):D562–D569. https://doi.org/10.1093/nar/gkaa895.

[6] KLIFS; 2022. [Online; accessed 01-February-2022]. https://klifs.net/.

[7] **Blue Ridge Institute for Medical Research in Horse Shoe NCU**, FDA-approved small molecule protein kinase inhibitors; 2022. [Online; accessed 01-February-2022]. http://www.brimr.org/PKI/PKIs.htm.

[8] **Manning G**, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The Protein Kinase Complement of the Human Genome. Science. 2002; 298(5600):1912–1934. https://doi.org/10.1126/science.1075762.

[9] **Kooistra AJ**, Volkamer A. Kinase-Centric Computational Drug Development. In: *Annual Reports in Medicinal Chemistry* Elsevier; 2017.p. 197–236. https://doi.org/10.1016/bs.armc.2017.08.001.

[10] **Morphy R**. Selectively Nonselective Kinase Inhibition: Striking the Right Balance. J Med Chem. 2009; 53(4):1413–1437. https://doi.org/10.1021/JM901132V.

[11] **Sydow D**, Aßmann E, Kooistra AJ, Rippmann F, Volkamer A. KiSSim: Predicting off-targets from structural similarities in the kinome. ChemRxiv. 2021; https://doi.org/10.26434/chemrxiv-2021-n3288.

[12] **Yueh C**, Rettenmaier J, Xia B, Hall DR, Alekseenko A, Porter KA, Barkovich K, Keseru G, Whitty A, Wells JA, Vajda S, Kozakov D. Kinase Atlas: Druggability Analysis of Potential Allosteric Sites in Kinases. J Med Chem. 2019; 62(14):6512–6524. https://doi.org/10.1021/acs.jmedchem.9b00089.

[13] **Karaman MW**, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, Faraoni R, Floyd M, Hunt JP, Lockhart DJ, Milanov ZV, Morrison MJ, Pallares

G, Patel HK, Pritchard S, Wodicka LM, et al. A quantitative analysis of kinase inhibitor selectivity. Nature Biotechnology. 2008; 26(1):127–132. https://doi.org/10.1038/nbt1358.

[14] **Eid S**, Turk S, Volkamer A, Rippmann F, Fulle S. KinMap: a web-based tool for interactive navigation through human kinome data. BMC Bioinformatics. 2017; 18(1). https://doi.org/10.1186/s12859-016-1433-7.

[15] **Sydow D**, Morger A, Driller M, Volkamer A. TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data. Journal of Cheminformatics. 2019; 11(1). https://doi.org/10.1186/s13321-019-0351-x.

[16] **Sydow D**, Rodríguez-Guerra J, Kimber TB, Schaller D, Taylor CJ, Chen Y, Leja M, Misra S, Wichmann M, Ariamajd A, Volkamer A. TeachOpenCADD 2022: open source and FAIR Python pipelines to assist in structural bioinformatics and cheminformatics research. Nucleic Acids Research. 2022; https://doi.org/10.1093/nar/gkac267, gkac267.

[17] **Kluyver T**, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, development team J. Jupyter Notebooks - a publishing format for reproducible computational workflows. In: Loizides F, Scmidt B, editors. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* IOS Press; 2016. p. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.

[18] List of Python introduction resources; 2022. [Online; accessed 21-February-2022]. https://github.com/volkamerlab/teachopencadd#python-programming-introduction.

[19] **The pandas development team**, pandas-dev/pandas: Pandas. Zenodo; 2020. https://doi.org/10.5281/zenodo.3509134.

[20] **Harris CR**, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, et al. Array programming with NumPy. Nature. 2020; 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2.

[21] **Virtanen P**, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020; 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.

[22] **Hunter JD**. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007; 9(3):90–95. https://doi.org/10.1109/mcse.2007.55.

[23] **Waskom ML**. seaborn: statistical data visualization. Journal of Open Source Software. 2021; 6(60):3021. https://doi.org/10.21105/joss.03021.

[24] **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

[25] **RDKit**, RDKit: Open-Source Cheminformatics;. http://www.rdkit.org, [Online; accessed 2022-02-02].

[26] **Kunzmann P**, Hamacher K. Biotite: a unifying open source computational biology framework in Python. BMC Bioinformatics. 2018; 19(1):346. https://doi.org/10.1186/s12859-018-2367-z.

[27] **Sydow D**, Rodríguez-Guerra J, Volkamer A. OpenCADD-KLIFS: A Python package to fetch kinase data from the KLIFS database. Journal of Open Source Software. 2022; 7(70):3951. https://doi.org/10.21105/joss.03951.

[28] **requests**, requests. https://docs.python-requests.org/;. https://docs.python-requests.org/, [Online; accessed 2022-02-02]. https://docs.python-requests.org/.

[29] **pillow**, pillow. https://python-pillow.org/;. https://python-pillow.org/, [Online; accessed 2022-02-02]. https://python-pillow.org/.

[30] TeachOpenCADD conda-forge package; 2022. [Online; accessed 2022-02-02]. https://anaconda.org/conda-forge/teachopencadd.

[31] **TeachOpenCADD**, TeachOpenCADD installation instructions. https://volkamerlab.org/;. [Online; accessed 2022-02-02]. https://projects.volkamerlab.org/teachopencadd/installing.html.

[32] **Project Jupyter**, Matthias Bussonnier, Jessica Forde, Jeremy Freeman, Brian Granger, Tim Head, Chris Holdgraf, Kyle Kelley, Gladys Nalvarte, Andrew Osheroff, Pacer M, Yuvi Panda, Fernando Perez, Benjamin Ragan Kelley, Carol Willing. Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. In: Fatih Akici, David Lippa, Dillon Niederhut, Pacer M, editors. *Proceedings of the 17th Python in Science Conference*; 2018. p. 113 – 120. https://doi.org/10.25080/Majora-4af1f417-011.

[33] **van Linden OPJ**, Kooistra AJ, Leurs R, de Esch IJP, de Graaf C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. Journal of Medicinal Chemistry. 2013; 57(2):249–277. https://doi.org/10.1021/jm400378w.

[34] **Rost B**. Twilight zone of protein sequence alignments. Protein Engineering, Design and Selection. 1999; 12(2):85–94. https://doi.org/10.1093/protein/12.2.85.

[35] **Henikoff S**, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992; 89(22):10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

[36] **Kunzmann P**, Hamacher K. Biotite: a unifying open source computational biology framework in Python. BMC Bioinformatics. 2018; 19(1). https://doi.org/10.1186/s12859-018-2367-z.

[37] **Volkamerlab**, KiSSim open-source Python package; 2022. [Online; accessed 01-February-2022]. https://github.com/volkamerlab/kissim.

[38] **TeachOpenCADD**, TeachOpenCADD website. https://volkamerlab.org/;. [Online; accessed 2022-02-02]. https://projects.volkamerlab.org/teachopencadd/.

[39] **Kosub S**. A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters. 2019; 120:36–38. https://doi.org/https://doi.org/10.1016/j.patrec.2018.12.007.

[40] **Barelier S**, Sterling T, O'Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins. ACS Chemical Biology. 2015; 10(12):2772–2784. https://doi.org/10.1021/acschembio.5b00683.

[41] Kinodata; 2022. [Online; accessed 01-February-2022]. https://github.com/openkinome/kinodata.

[42] OpenKinome; 2022. [Online; accessed 01-February-2022]. http://openkinome.org/.

[43] **Merget B**, Turk S, Eid S, Rippmann F, Fulle S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. Journal of Medicinal Chemistry. 2017; 60(1):474–485. https://doi.org/10.1021/acs.jmedchem.6b01611.

[44] **Hastie T**, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer New York; 2009. https://doi.org/10.1007/978-0-387-84858-7.

[45] **Schneider P**, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, Fisher J, Jansen JM, Duca JS, Rush TS, Zentgraf M, Hill JE, Krutoholow E, Kohler M, Blaney J, Funatsu K, Luebkemann C, Schneider G. Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery. 2019; 19(5):353–364. https://doi.org/10.1038/s41573-019-0050-3.

[46] **Wilkinson MD**, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3(1). https://doi.org/10.1038/sdata.2016.18.

[47] **Schmidt D**, Scharf MM, Sydow D, Aßmann E, Martí-Solano M, Keul M, Volkamer A, Kolb P. Analyzing Kinase Similarity in Small Molecule and Protein Structural Space to Explore the Limits of Multi-Target Screening. Molecules. 2021; 26(3):629. https://doi.org/10.3390/molecules26030629.

[48] **conda-forge community**, The conda-forge Project: Community-based Software Distribution Built on the conda Package Format and Ecosystem. Zenodo; 2015. https://doi.org/10.5281/zenodo.4774216.