KiSSim: Predicting off-targets from structural similarities in the kinome

Dominique Sydow,[†] Eva Aßmann,[†] Albert J. Kooistra,[‡] Friedrich Rippmann,[¶] and Andrea Volkamer[*,†,§]

†*In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Augustenburger Platz 1, 13353 Berlin, Germany*

‡*Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark*

¶*Computational Chemistry & Biologics, Merck Healthcare KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany*

§*Corresponding author: Andrea Volkamer; andrea.volkamer@charite.de*

E-mail: andrea.volkamer@charite.de

Abstract

Protein kinases are among the most important drug targets because their dysregulation can cause cancer, inflammatory, and degenerative diseases. Developing selective inhibitors is challenging due to the highly conserved binding sites across the roughly 500 human kinases. Thus, detecting subtle similarities on a structural level can help to explain and predict off-targets among the kinase family.

Here, we present the kinase-focused and subpocket-enhanced KiSSim fingerprint (*Ki*nase *S*tructural *Sim*ilarity). The fingerprint builds on the KLIFS pocket definition, composed of 85 residues aligned across all available protein kinase structures, which

enables residue-by-residue comparison without a computationally expensive alignment. The residues' physicochemical and spatial properties are encoded within their structural context including key subpockets at the hinge region, the DFG motif, and the front pocket.

Since structure was found to contain information complementary to sequence, we used the fingerprint to calculate all-against-all similarities within the structurally covered kinome. Thereby, we could identify off-targets that are unexpected if solely considering the sequence-based kinome tree grouping; for example, Erlobinib's known kinase off-targets SLK and LOK show high similarities to the key target EGFR (TK group) though belonging to the STE group. KiSSim reflects profiling data better or at least as well as other approaches such as KLIFS pocket sequence identity, KLIFS interaction fingerprints (IFPs), or SiteAlign. To rationalize observed (dis)similarities, the fingerprint values can be visualized in 3D by coloring structures with residue and feature resolution.

We believe that the KiSSim fingerprint is a valuable addition to the kinase research toolbox to guide off-target and polypharmacology prediction. The method is distributed as an open-source Python package on GitHub and as conda package: `https://github.com/volkamerlab/kissim`

Introduction

Protein kinases are involved in most aspects of cell life due to their role in signal transduction. Their dysregulation can cause severe diseases such as cancer, inflammation, and neurodegeneration,[1] which makes them a frequent target of drug discovery campaigns. In 2015, 30% of FDA-approved small molecules targeted kinases.[2] The roughly 500 kinases in the human genome share a highly conserved binding site, which challenges selective drug design for a single kinase or a well-defined set of kinases (polypharmacology) avoiding binding to undesired off-targets.[3,4]

Protein kinases bind adenosine triphosphate (ATP) to catalyze the transfer of its phosphate group to serine, threonine, or tyrosine residues of themselves or other proteins. ATP and most other ligands bind to the front cleft of the kinase pocket that lays between the two kinase domains, the C- and N-terminal lobes. These domains are connected via a hinge region, which is forming important hydrogen bonds to ATP as well as most studied ligands. The gate area contains the conserved DFG (aspartate-phenylalanine-glycine) motif, whose phenylalanine flips in and out of the front pocket, opening and closing a hydrophobic region in the back cleft, i.e., the DFG-in and DFG-out conformation, respectively. The back cleft also comprises the αC-helix with a conserved glutamine residue, which forms a salt bridge with a conserved lysine residue in the gate area. Such a conformation is called αC-in as opposed to αC-out.[5]

Researchers have studied kinase similarity between the full — or parts of the — kinome from many different angles. Manning et al.[6] used a multiple sequence alignment (MSA) to cluster the kinome into eight main groups of eukaryotic protein kinases (ACG, CAMK, CK1, CMGC, STE, TK, TKL, and Other) and the atypical protein kinase families. Recently, Modi and Dunbrack[7] assigned some kinases, which were left unassigned in the Other category, based on a structurally validated MSA.

While sequence comparison — and thus, evolutionary similarity — can explain many observations from kinase profiling experiments, other more distantly related off-targets remain undetected. For example, profiling Erlotinib against 48 kinases revealed high affinity against the on-target EGFR (TK group) but also the non-TK off-targets SLK, LOK, and GAK;[8] or the chemical probe SGC-STK17B-1 binds both DRAK2 and CaMMK,[9] although they are dissimilar when judged solely by their sequence.[6] Focusing on the kinase pocket instead of the whole sequence already helps: The 50 most similar kinases to EGFR are only TK kinases when ranked by full-length sequence while listing non-TK kinases when considering the pocket sequence only.[10] The KinCore phylogenetic tree produced by a kinome-wide structure-guided MSA[7,11] overall confirms the assignment from Manning et al.[6] but provides higher

precision, e.g. regarding previously unassigned kinases. Schmidt et al.[12] have recently investigated the similarities between a panel of nine kinases — EGFR, ErbB2, PIK3CA, KDR, BRAF, CDK2, LCK, MET, and p38a — based on different pocket encodings, including the pocket sequence identity, pocket structure similarity, interaction fingerprint similarity, and ligand promiscuity. Individual kinase relationships differed according to these different perspectives, while some trends could be observed such as the atypical kinase PIK3CA being an outlier amongst the otherwise typical kinases in this panel.

In an attempt to facilitate computer-aided kinase similarity studies, we here aim to add another perspective. Binding site comparison methods employed so far can be applied to any binding site regardless of the protein class. Kuhn et al.[13] have applied such a method, Cavbase, to the structurally resolved kinome and could detect expected and unexpected kinase relationships. Since kinases are highly conserved and have been aligned and annotated across the full structurally covered kinome, a binding site comparison method tailored to kinases may provide an extended perspective on kinase similarities. We make use of data in the KLIFS[14] database, a rich resource for kinase research that extracts protein kinase-focused information on structures from the PDB,[15] on inhibitors in clinical trials from the PKIDB,[16] on bioactivities from ChEMBL,[17] and much more. All kinase structures from the PDB are split into single chains and models and aligned with respect to sequence and structure across the full structurally covered kinome. The KLIFS authors defined the kinase pocket as a set of 85 residues that interact with co-crystallized ligands in the initial KLIFS dataset of more than 1200 structures.[5] Thanks to this structural alignment, it is possible to look up all 85 residues in any kinase structure, given the residue is structurally resolved and not in a gap position. This pocket alignment is the basis for the here introduced KiSSim fingerprint.

The kinase-focused and subpocket-enhanced KiSSim (Kinase Structural Similarity) fingerprint builds on the KLIFS[14] pocket, whose alignment allows a computationally inexpensive residue-by-residue comparison. The residues' physicochemical and spatial properties are

encoded within their structural context including important kinase subpockets — the hinge region, DFG region, and front pocket — building on features from previously published methods such as SiteAlign,[18] KinFragLib,[19] and Ultrafast Shape Recognition (USR).[20] We used the fingerprint to calculate all-against-all similarities within the structurally covered kinome and to generate a KiSSim-based kinome tree. Detected similarities can be used to predict off-targets or guide polypharmacology studies and to rationalize profiling observations on a structural level. We distribute the method as an open source Python package at `https://github.com/volkamerlab/kissim` and as conda package, alongside the data and analyses notebooks at `https://github.com/volkamerlab/kissim_app` to support FAIR[21] science.

Methods & Data

In the following, we outline the KiSSim methodology and implementation, the datasets used, and the method's evaluation. All data, fingerprints, and analyses are available at `https://github.com/volkamerlab/kissim_app`.

KiSSim methodology

The KiSSim methodology consists of three steps: the encoding of a set of kinase binding sites as KiSSim fingerprints (Figure 1), the all-against-all comparison of these structures using their fingerprints, and — since one kinase can be represented by multiple structures — the mapping of multiple structure/fingerprint pairs to one kinase pair.

Encoding: From structure to fingerprint

The KiSSim fingerprint encodes the 85 KLIFS pocket residues in the form of physicochemical and spatial properties as illustrated in Figure 1. We summarize the encoding procedure in the following; for a detailed description please refer to the Supplementary methods section.

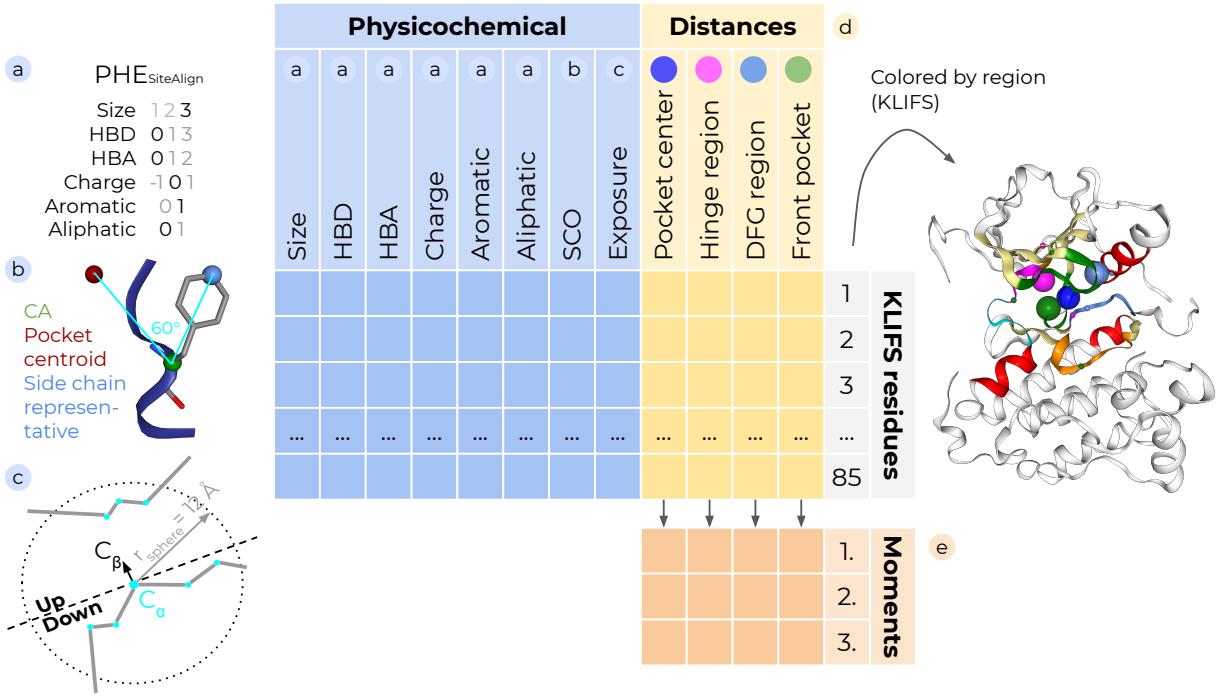

Figure 1: **KiSSim fingerprint encodes physicochemical and spatial properties of kinase pockets.** The fingerprint builds on the KLIFS[14] pocket definition, i.e. 85 residues aligned across all available protein kinase structures, which enables residue-by-residue comparison without a computationally expensive alignment. Each residue is encoded physicochemically and spatially. Physicochemical properties include the following features per residue (example: phenylalanine/PHE): (a) Pharmacophoric features and size categories are taken from the SiteAlign[18] binding site comparison methodology. (b) Side chain orientation is adapted from SiteAlign and defined as inward-facing, intermediate, or outwards-facing depending on the vertex angle between the pocket centroid, the residue's side chain representative (Table S3), and CA atom. (c) Solvent exposure is defined as high, intermediate, or low, depending on the ratio of CA atoms in the upper half of a sphere cut in half by a normal plane spanned by the residue's CA-CB vector. The implementation is based on BioPython's HSExposure.[22,23] Spatial properties are defined as follows: (d) Each residue's distance to the pocket center and important kinase subpockets, i.e., the hinge region, DFG region, and the front pocket. On the right, example locations are shown in the 3D representation of kinase EGFR (PDB ID: 2ITO, KLIFS structure ID: 783). (e) The distance distributions per pocket center and subpocket are furthermore described by their first three moments, i.e. the mean, standard deviation, and skewness.

Physicochemical properties are encoded by eight features in the form of categorical values. Pharmacophoric and size features are taken from the SiteAlign categories for standard amino acids.[18] They encode the *size* based on the number of heavy atoms, the number of hydrogen bond donors (*HBD*) and hydrogen bond acceptors (*HBA*), the *charge* (negative, neutral, or positive), and *aromatic* and *aliphatic* properties (present or not present) of a residue (Table S1). The *side chain orientation* (inward-facing, intermediate, or outward-facing) is based on the vertex angle from the residue's CA atom (vertex) to the pocket center and to the residue's outermost side chain atom, the side chain representative (Table S3). The *solvent exposure* of a residue (high, intermediate, or low) is based on the ratio of CA atoms in the upper half of a sphere that is placed around the residue's CA atom (radius 12 Å) and cut in half by a normal plane spanned by the residue's CA-CB vector, as implemented in BioPython's HSExposure module.[22,23]

Spatial properties are described by discrete values, i.e., distances and moments. *Spatial distances* are calculated from each residue's CA atom to the pocket's geometric center and to prominent subpocket centers. The *pocket center* is the centroid of all pocket CA atoms. The selected *subpocket centers* include functionally well-characterized kinase regions such as the hinge region, DFG region, and front pocket. Each subpocket center is calculated based on the centroid of three anchor residues' CA atoms (Table S4), following the idea described in the KinFragLib methodology.[19] We added the code to calculate the subpocket centers to the structural cheminformatics library OpenCADD (module `opencadd.structure.pocket`)[24] to allow for easy access in other projects. *Spatial moments* describe each of the four distributions of distances to the pocket center, hinge region, DFG region, and front pocket. In KiSSim, the first three moments are used: the mean, the standard deviation, and the cube root of the skewness. This procedure is inspired and adapted from the ligand-based Ultrafast Shape Recognition (USR)[20] method.

Fingerprint length. The final full-length fingerprint encompasses eight discrete physicochemical features (8 features x 85 residues), four continuous spatial distance features (4

features x 85 residues), and three continuous spatial moment features (3 moments x 4 distributions), resulting in a 1032 bit vector. Optionally, a subset of residues can be selected to generate a subset fingerprint emphasizing certain residues. We offer a subset of residues that is based on frequently interacting co-crystallized ligands,[25] see more details in the Supplementary methods section.

Normalization. Fingerprints are normalized to values between 0 and 1 by applying a min-max normalization. For discrete features, the minimum and maximum categorical values are used. For continuous features, the minimum and maximum values for each spatial feature are set to the minimum and maximum values observed across all structures; distance extrema are defined for each residue position individually, while moment extrema for the first, second, and third moment individually.[26]

Pairwise structure comparison

Two kinase pocket structures — encoded as two fingerprints — can be compared in two steps (Figure 2). First, we calculate for each feature the distance between the corresponding two feature vectors across the 85 residue entries, producing a *feature distances* vector of length 15 (i.e., aggregating over the columns in Figure 2 a). For example, the two fingerprints' 85-bit size feature vectors — representing the size of each of the 85 pocket residues — will be reduced to a single size feature distance. The distance between discrete features is defined as the scaled L1 norm $\|\mathbf{x}\|_1 = \frac{1}{n}\sum_{i=1}^{n}|x_i|$ (scaled Manhattan distance), whereas the distance between continuous features is defined as the scaled L2 norm $\|\mathbf{x}\|_2 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}x_i^2}$ (scaled Euclidean distance), where $\mathbf{x}$ is a vector of length n.[27] Second, we calculate the weighted sum of the 15-bit feature distance vector with feature-level weights $\alpha_{1\ldots15}$ to produce the final *fingerprint distance*. By default, the 15 features are equally weighted with a weight of $\frac{1}{15}$ each.

Summarizing both steps, the fingerprint distance $d(\mathbf{f}_i, \mathbf{f}_j)$ between two fingerprints $\mathbf{f}_i$ and $\mathbf{f}_j$ is defined in Equation 1. The different KiSSim features are denoted as m: 1=size,

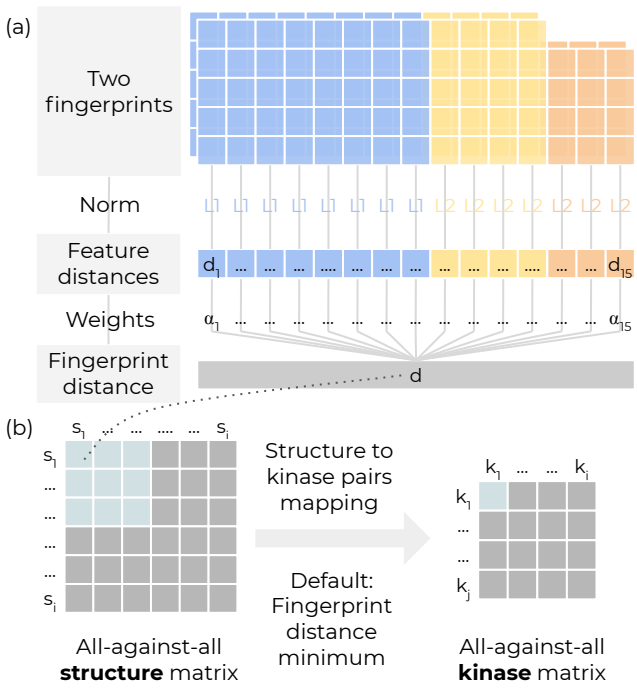

Figure 2: **Structures — encoded as KiSSim fingerprints — are compared pairwise and mapped to kinase pairs.** (a) The discrete physicochemical features (blue) are compared using the scaled L1 norm, while the continuous spatial features (yellow/orange) are compared using the scaled L2 norm, resulting in a feature distances vector composed of one distance per feature. Custom weighting of these features results in the final fingerprint distance. By default, the features are weighted equally. (b) Two kinases of interest may have multiple structures each. Thus, multiple structure/fingerprint pairs can represent the same kinase pair. By default, we select the minimum (fingerprint) distance value amongst all structure/fingerprint pairs to represent the (kinase) distance between a kinase pair.

2=HBD, 3=HBA, 4=charge, 5=aromatic, 6=aliphatic, 7=side chain orientation, 8=solvent exposure, 9=distance to pocket center, 10=distance to hinge region, 11=distance to DFG region, 12=distance to front pocket, 13=first moment, 14=second moment, and 15=third moment.

$$d(\mathbf{f}_i, \mathbf{f}_j) = \sum_{m=1}^{8} \alpha_m \frac{\left\|\mathbf{f}_i^m - \mathbf{f}_j^m\right\|_1}{85} + \sum_{m=9}^{12} \alpha_m \frac{\left\|\mathbf{f}_i^m - \mathbf{f}_j^m\right\|_2}{85} + \sum_{m=13}^{15} \alpha_m \frac{\left\|\mathbf{f}_i^m - \mathbf{f}_j^m\right\|_2}{4} \quad (1)$$

Kinome-wide comparison

The kinome-wide comparison is based on an all-against-all comparison of all available structures. Note that a kinase can be represented by multiple structures (see KLIFS data section), thus, a kinase pair can be represented by multiple structure pairs with multiple distance values. Our final goal is to assign one distance value to each kinase pair as a measure of the similarity between these two kinases (Figure 2 b). The structural coverage of kinases is highly imbalanced: Some kinases are represented by one structure only, others like EGFR or CDK2 by more than 100. We select the structure pair with the lowest distance as representative for the kinase pair, hence always picking the two closest structures in the dataset. For example, if a dataset consists of ten structures representing three kinases, the 10×10 all-against-all *structure distance matrix* will be reduced to a 3×3 all-against-all *kinase distance matrix*, consisting of the lowest distance values only after mapping structure pairs to kinase pairs.

Fingerprint and similarity visualization in 3D

Fingerprint features can be visualized in 3D using the NGLviewer[28,29] and IPyWidgets[30] for the following applications: (a) Fingerprint features of a structure can be visualized in 3D by coloring the residues by different feature values. (b) The difference between two structures can be highlighted to spot positions of high or low similarity between two structures. The differences are shown for each feature type individually. (c) The standard deviation

of spatial features between all structures available for one kinase can be mapped onto an example structure in 3D to show regions of high or low variability between different kinase conformations.

KiSSim tree

The kinase distance matrix produced as described in the Kinome-wide comparison section is submitted to a hierarchical clustering as implemented in SciPy[31] using as metric the Euclidean distance and as linkage Ward's criterion. We generate a phylogenetic tree in the Newick format based on this KiSSim kinase clustering. The tree branches are labeled with the mean of all distances belonging to that branch; the tree leaves are annotated with the kinase names and their assigned Manning kinase groups. We visualize the tree in an automatized way using BioPython's Phylo[22,32] module to be used in Jupyter Notebooks, and in a manual way using the freely available FigTree[33] software to produce publication-ready circular trees.

KiSSim implementation

The `kissim` library is implemented as an open-source Python package, which is available on GitHub at `https://github.com/volkamerlab/kissim` and as conda package at conda-forge.[34,35] Structures are retrieved via the OpenCADD-KLIFS module[24] and are encoded as fingerprints using the `FingerprintGenerator` class; fingerprints can be compared using the `FingerprintDistanceGenerator` class. We also offer quick access `encode` and `compare` functionalities as Python API and as command-line interface (CLI), see Figure 3. Lastly, the `kissim.encoding.tree` module offers an automatized all-against-all clustering and phylogenetic tree generation, while the 3D visualization of fingerprints and pairwise comparisons is implemented in the `kissim.viewer` module.

Structural data is read and processed with BioPython[22] and BioPandas;[36] computation is performed with NumPy,[37] Pandas,[38] SciPy,[39] and Scikit-learn.[40] The code for operations

that are of use outside of the KiSSim project has been added to the OpenCADD library:[24] KLIFS queries are implemented in the OpenCADD-KLIFS module and subpocket centers can be defined and visualized with the OpenCADD-pocket module.

All code is written in Python 3[41] following the PEP8 style guide. We document the code following NumPy docstrings[42] as well as format and lint the code and notebooks with black,[43] black-nb,[44] flake8,[45] and flake8-nb.[46] A detailed documentation is hosted on ReadTheDocs[47] at `https://kissim.readthedocs.io` using sphinx.[48] We test the `kissim` code using pytest[49] with a code coverage of over 90%, measured with CodeCov.[50] Notebooks are checked with nbval[51] and continuous integration is deployed with GitHub Actions[52] on a weekly basis.

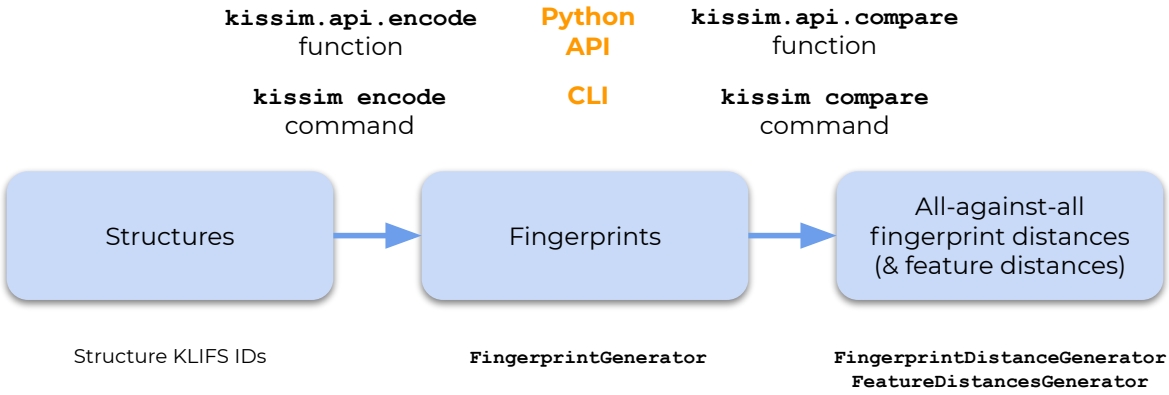

Figure 3: **The `kissim` library's Python API and CLI.** Structures from the KLIFS database can be encoded as fingerprints using the `FingerprintGenerator` class (details in Figure 1) and compared using the `FeatureDistancesGenerator` and `FingerprintDistanceGenerator` class (details in Figure 2). The package offers the wrappers `encode` and `compare` for quick and easy access from within a Python script (Python API) or from the command line (CLI). Please also refer to the `kissim` library's documentation at `https://kissim.readthedocs.io`.

Data

We are using the following sources of external data: KLIFS kinase structures[14] and the profiling datasets by Karaman et al.[8] and Davis et al.[53], filtered and processed as described in the following. All prepared datasets described here are accessible via the `src.data` module

at `https://github.com/volkamerlab/kissim_app`.

KLIFS data

We downloaded the human structural kinase dataset from the KLIFS database version 3.2[14] on 2021-09-02. This dataset contained 11806 human monomeric structures, i.e., PDB entries split into monomeric structures if consisting of multiple chains and alternate models. We filtered the dataset for human kinases with a resolution $\leq$ 3Å and a KLIFS quality score ≥ 6. The KLIFS quality score ranges from 0 (bad) to 10 (flawless) and describes the quality of the structural alignment and resolution regarding missing residues and atoms. In addition, we excluded structures with more than three pocket mutations or with more than eight missing pocket residues. In order to reduce computational costs, we selected the best structure per kinase in each PDB entry (kinase-PDB pair); the best structure per kinase-PDB pair is defined as the structure with the least missing pocket residues, the least missing pocket atoms, the lowest alternate model identifier, and the lowest chain identifier (in that order). Structures were excluded if they are flagged as problematic structures in KLIFS and if they could not be encoded as KiSSim fingerprint. We produced three final datasets of structures for KiSSim fingerprint generation and all-against-all comparison: structures in any DFG conformation, DFG-in conformation only, and DFG-out conformation only. Table 1 lists the number of structures remaining after each filtering step.

Bioactivity profiling data

To compare predicted and measured on- and off-targets, we use two kinase bioactivity datasets available through KinMap:[56] The Karaman et al.[8] and the Davis et al.[53] datasets on KinMap contain inhibition profiles (K_d values) for 38 and 72 kinase inhibitors across 317 and 442 kinases, respectively. The lower the K_d value, the higher the binding affinity, which is used as a proxy for activity. We pooled data from both datasets by taking the union of all kinase-ligand pairs. If kinase-ligand pairs have bioactivity values in both datasets,

Table 1: **KiSSim dataset.** Upper half: Filtering steps performed on the human dataset from KLIFS version 3.2[14] downloaded on 2021-09-02 to generate the KiSSim dataset. Lower half: Number of structures and kinases as well as number of structure and kinase pairs encoded and compared with the KiSSim methodology; number of structure/kinase pairs does not contain self-comparisons. See notebooks for more details.[54,55]

	Number of structures		
	all	**DFG-in**	**DFG-out**
Select species: human	11806		
Select KLIFS structures without flag	11650		
Select resolution: ≤ 3	10690		
Select quality score: ≥ 6	10236		
Select mutated pocket residues: ≤ 3	10155		
Select missing pocket residues: ≤ 8	10150		
Select conformation	10150	8982	786
Select best structure per PDB and kinase pair	4690	4120	407
Encode structures as fingerprints	4681	4112	406
Number of structures	4681	4112	406
Number of kinases	279	257	71
Number of structure pairs	10953540	8452216	82215
Number of kinase pairs	38781	32896	2485

we proceeded as follows: If both measurements $K_{d,1}$ and $K_{d,2}$ are (a) below or equal to or (b) above or equal to the chosen activity cutoff of $K_d^{\text{cutoff}} = 100\,\text{nM}$, we kept the lower K_d, i.e. the more active compound. If one of the measurements is above and the other below that cutoff, we kept the lower K_d if the difference is $|K_{d,1} - K_{d,2}| \leq 100\,\text{nM}$, otherwise, the measurements were discarded. That way we keep the measurement with the lowest K_d if both measurements agree on the ligand's activity, including a tolerance zone around our defined activity cutoff; and we remove measurements if they disagree considerably. This approach results in a 353×80 kinase-ligand matrix with 7619 measurements, named *Karaman-Davis dataset* from here on.

Evaluation

We evaluate our KiSSim results by comparison to profiling data as well as alternative similarity measures based on KLIFS pocket sequences, KLIFS pocket interaction finger-

prints (IFPs), and SiteAlign.[18] All prepared datasets and evaluation strategies described here are accessible via the `src.data` and `src.evaluation` modules at `https://github.com/volkamerlab/kissim_app`.

KiSSim evaluation using profiling data

To evaluate how well KiSSim detects kinase similarities, we need to define a ground truth of kinase similarities. We use profiling data as a surrogate for this, since it is safe to assume that kinases that are targeted by the same ligand share similar binding sites. To this end, we use the profiling Karaman-Davis dataset, which describes the activity of ligands against a panel of kinases. We assign each ligand l_i in the profiling dataset to their reported key target(s) $k_j(l_i)$ in the PKIDB,[16] ranging from one target to multiple targets, e.g. Erlotinib is assigned to EGFR only while Imatinib binds to ABL1, KIT, RET, TRKA, FMS, and PDGFRa. These examples result in the following kinase-ligand pairs: EGFR-Erlotinib, ABL1-Imatinib, KIT-Imatinib, RET-Imatinib, TRKA-Imatinib, FMS-Imatinib, and PDGFRa-Imatinib. Note that we only included (a) kinases whose name could be mapped to the KinMap kinase names and (b) ligands that are listed in the PKIDB and are FDA-approved. Furthermore, only kinase-ligand pairs were included (a) whose kinase was tested active against the ligand ($K_d \leq 100\,\text{nM}$) and (b) that share at least 10 kinases between the Karaman-Davis and KiSSim datasets, of which at least three have measured ligand activities of $K_d \leq 100\,\text{nM}$. For example, the EGFR-Erlotinib pair shares Erlotinib profiling measurements and EGFR KiSSim distances for 50 kinases, of which four are defined as active using the aforementioned K_d cutoff. Each remaining kinase-ligand pair is evaluated as demonstrated here for the EGFR-Erlotinib pair (l_1 = Erlotinib and $k_1(l_1)$ = EGFR):

1. We define the kinases in both lists as active or inactive based on the chosen activity threshold of $K_d = 100\,\text{nM}$.

2. We rank all kinases by their KiSSim distance to EGFR. These are our *KiSSim-based kinase similarities*.

3. We calculate ROC curves to demonstrate how well the profiling data is predicted by our KiSSim-based kinase similarities.

Some kinase activities measured in the profiling dataset are rather unexpected from a sequence-based similarity point of view. For the EGFR-Erlotinib example, we use the KinMap server to plot the profiling-based and KiSSim-based ranked kinases onto the kinome tree by Manning et al.[6]. For example, we highlight kinases with measured activities against Erlotinib as well as the 50 most similar kinases to EGFR as detected by KiSSim. All kinases that are part of the KiSSim dataset are shown as well to define which data points are available for similarity predictions.

KiSSim comparison to other methods

We outline here the preparation of all-against-all kinase distance matrices based on different similarity measures to be compared to the KiSSim kinase distance matrix (KiSSim dataset section): KLIFS pocket sequence, KLIFS pocket-ligand interaction fingerprint (IFP), and SiteAlign's pocket structure. All distance matrices underwent a min-max normalization[57] and can be loaded via `src.data.distances` at `https://github.com/volkamerlab/kissim_app`.

KLIFS pocket sequence. We performed an all-against-all comparison of the sequence identity within the KLIFS pocket of 85 residues. The sequence identity is defined as the number of identical pocket residues divided by all 85 pocket residues; gap positions are treated as identical if both structures show a gap. If two sequences are identical, the sequence identity is 1; if two sequences do not have a single residue in common, the sequence identity is 0. In order to make these values comparable with the kinase distance matrices, we define distance $= 1 -$ identity.

KLIFS pocket IFP. We performed an all-against-all comparison of the KLIFS IFP describing interactions between co-crystallized ligands and the KLIFS pocket. For each pocket residue, seven potential protein-ligand interaction types were defined as described by Marcou

and Rognan[58]. The presence or absence of a certain type of interaction is noted as 1 or 0 in the bit-string. This results in an $85 \cdot 7 = 595$-bit long IFP per pocket-ligand pair. The Jaccard distance is used to compare the IFPs. If multiple IFP pairs describe the same kinase pair, we selected the minimum distance as the representative measure for the kinase pair, following the same procedure as described for the KiSSim methodology.

SiteAlign. We performed an all-against-all comparison using the pocket comparison method SiteAlign[18] (version 4.0). In this approach, properties of a binding site are projected to a triangulated sphere positioned at the pocket center, stored as a fingerprint to be compared and aligned to another binding site fingerprint iteratively. Since we used the existing KLIFS alignment, a few SiteAlign parameters were adapted to reduce runtime: we decreased the number of alignment steps in SiteAlign from 3 to 1, the translational steps from 5 to 3, and reduced the rotational and translational intensity from 2π to $\frac{1}{4}\pi$ and from 4 to 1, respectively. Comparison of the SiteAlign performance for > 4000 structure pairs with the default and adjusted settings, showed that the adjusted settings resulted in lower distances (average decrease of 6%), while matching a higher number of triangles (average increase of 15%). Pocket residues with modifications (e.g. phosphorylated threonines) were excluded to avoid segmentation faults.

Results and Discussion

We present here the generated KiSSim dataset and the resulting KiSSim-based kinome tree. Furthermore, we evaluate the KiSSim results in comparison to profiling data (KiSSim evaluation using profiling data section) and other pocket encoding methods (KiSSim comparison to other methods section).

KiSSim dataset

KLIFS structures are filtered as described in detail in the KLIFS data section (Table 1), then encoded and compared as described in the KiSSim methodology section. When considering structures in DFG-in conformations only, 4112 fingerprints representing 257 kinases result in a 4112×4112 *structure distance matrix* and — after mapping structure to kinase pairs as described in the Kinome-wide comparison section — in a 257×257 *kinase distance matrix* (Table 1).

Fingerprint feature value distribution

The KiSSim fingerprint encodes the 85 KLIFS pocket residues in the form of physicochemical and spatial properties. *Physicochemical properties* include pharmacophoric and size features, side chain orientation, and solvent exposure; *spatial properties* include each residue's distance to the pocket center as well as to three subpockets and the first three moments of the resulting distance distributions (Figure 1). We investigate here the fingerprint feature value distribution across all KiSSim fingerprints.

The value distributions for pharmacophoric and size features differ depending on the feature type (Figure 4a) and the residue position (Figure S2 and S3). For example, the amino acid size is more evenly distributed than the aromatic or charge feature, since most amino acids are neither aromatic nor charged (Figure 4a, left). Since the five pharmacophoric and residue size features encode — in an abstracted manner — the pocket sequence, features are more robust at more conserved pocket positions than at other positions; examples are the conserved salt-bridge between residues 17 and 24 or the DFG residues 81–83 (Figure S2).

Spatial distances range between 2–33 Å (Figure 4a, middle), however, depending on the residue position, the values cover only a subset of this range. For example, the hinge region residues 46–48 are close to the hinge region center, while further away from the DFG region center (Figure S3). Distances from subpocket centers to regions such as the G-rich loop (residues 4–9), the αC-helix (residues 20–30), and the DFG motif vary more than for example

to the hinge region, which agrees with knowledge on more flexible vs. more stable regions in the kinase pocket. The spatial moment features describe the distance distributions between the pocket residues to the subpocket centers. They show lower variability for the mean and the standard deviation but high variability for the skewness (Figure 4a, right).

The spatial features are based on the KiSSim subpockets as described in the Encoding: From structure to fingerprint section. These subpockets are calculated for each structure individually, however, show robustness over the structural kinome. The subpocket centers occupy the same space across the aligned KLIFS structures, while the front pocket and DFG region center show higher variability than the hinge region and pocket center (Figure 4b), as to be expected. Therefore, the subpocket definition procedure seems to be robust enough to span comparable subpocket centers while fine-grained enough to encode structural differences.

In conclusion, the feature space encoded in the KiSSim fingerprint, on the one hand, reflects sequence-related similarities between kinases on a generalized level through the defined physicochemical properties and, on the other hand, incorporates information on flexible and stable regions through the defined spatial properties.

Fingerprint distances to compare structures

Moving on from the structure encoding (fingerprints) to the structure comparison (fingerprint distances), we aimed to explore if the KiSSim fingerprint can be used to discriminate between kinases and between DFG-in and DFG-out conformations.

First, we measured the discriminating power between kinases by comparing KiSSim fingerprint distances between DFG-in structures of the same kinase and of different kinases, i.e. intra-kinase and inter-kinase distances, respectively. With a median of 0.02 compared to 0.11, the (about 200000) intra-kinase distances are significantly lower than the (about 8.2 million) inter-kinase distances as shown in Figure 5a, indicating that the fingerprint can discriminate between kinases. Note that the distances between structure pairs describing the

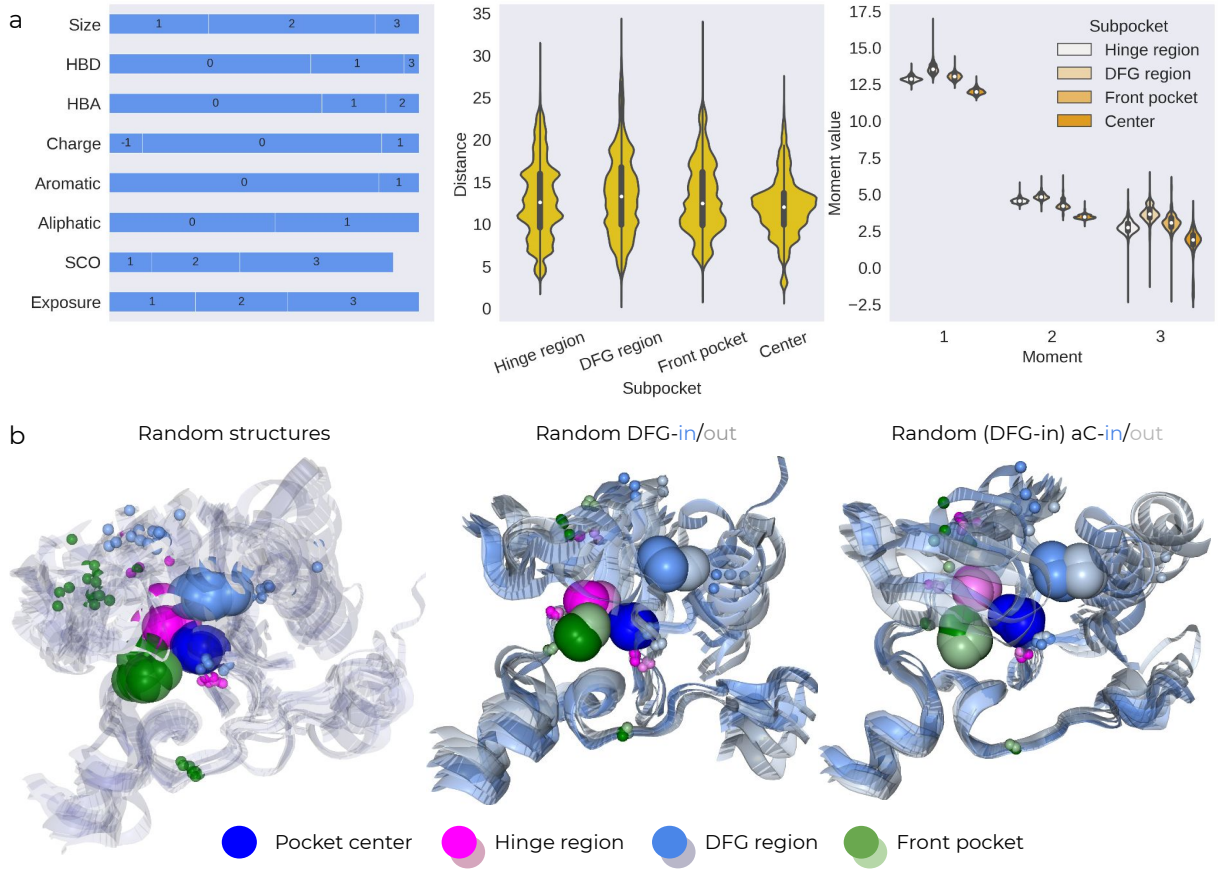

Figure 4: **Fingerprint feature and subpocket distributions.** (a) Distribution of all over 400, 000 feature values aggregated from all structures and all pocket residues. Categorical physicochemical features (in blue) include size, hydrogen bond donor count (HBD), hydrogen bond acceptor count (HBA), charge, aromatic, aliphatic, side chain orientation (SCO), and solvent exposure. Distance features (in yellow) include distances to the subpocket centers for the hinge region, DFG region and front pocket as well as the pocket centroid. Moment features (in orange) include the first three moments, i.e. mean, standard deviation, and scaled skewness, for each structure's distance distribution. (b) The subpocket centers are shown in 3D for example structures (left), highlighted by DFG conformations (middle) and αC-helix conformations for example DFG-in structures (right). See notebooks for more details.[59–61] Note: We show here unnormalized fingerprints; for the downstream fingerprint comparison, the fingerprints are normalized to values between 0 and 1 first.

same kinase pair can vary a lot (Figure S4); for the all-against-all comparison, we consider only the most similar structure pair per kinase pair.

Second, we measured KiSSim's discriminating power between DFG conformations by comparing fingerprint distances between structure pairs in DFG-in/in, DFG-out/out, and DFG-in/out conformations. For this analysis, we used the distances based on only the spatial fingerprint features to exclude the eight physicochemical features and thereby to focus on conformational information. While the distributions for the three categories are similar when considering all kinases (data not shown), they differ when split by kinase as shown exemplarily for the BRAF kinase in Figure 5b, indicating that the fingerprint can discriminate between DFG conformations. We conducted this analysis for other kinases with sufficient structural coverage for DFG-in and -out conformations and observed the same for CDK8, EphA2, MET, and p38a (see details in notebook[62]).

Before we use the KiSSim fingerprints for an all-against-all comparison, we confirmed two important properties: First, the KiSSim fingerprint distances for structures describing the same kinase are significantly lower than for structures describing different kinases (here based on DFG-in structures only). Second, the fingerprint distances for structures in the same DFG conformation are lower than for DFG-in/out structure pairs (here based on spatial features only).

KiSSim-based kinome tree

Structure is known to be more conserved than sequence,[64] and previous studies have shown that including structural information adds orthogonal information to shed light on unexpected similarities between kinases and off-target effects.[7,12] To help detect such relationships between more distantly related kinases, we generated KiSSim kinome trees based on the DFG-in conformations, as described in detail in the KiSSim tree section, to investigate all-against-all relationships between kinases compared to the sequence-based kinome tree by Manning et al.[6]. Note that we can base the comparison on structurally resolved

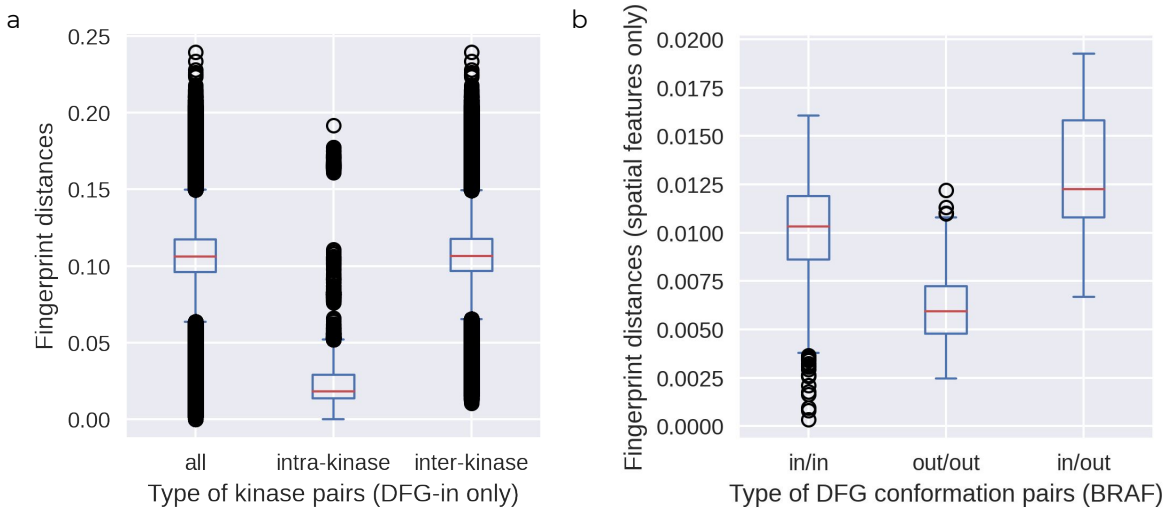

Figure 5: **KiSSim fingerprint can distinguish between kinases and DFG conformations.** (a) We compare fingerprint distances (based on all fingerprint bits) for structure pairs representing any kinase (all), the same kinase (intra-kinase), or different kinases (inter-kinase); here we use only DFG-in conformations. Dataset includes about 8.4 million pairwise structure distances, of which about 200000 and 8.2 million are intra-kinase and inter-kinase pairs, respectively. (b) We compare fingerprint distances (based on spatial distance fingerprint bits only) for structure pairs representing the BRAF kinase in different DFG-conformations. Dataset includes 28 DFG-in and 21 DFG-out structures, resulting in 378 DFG-in/in, 210 DFG-out/out, and 588 DFG-in/out pairwise structure distances. The box-and-whisker plot extends from the Q1 to Q3 quartile values of the data; the whiskers extend no more than $1.5 \cdot \mathrm{IQR}$ with $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$. See notebooks for more details.[62,63]

kinases only, i.e., 257 out of the roughly 500 human kinases.

The KiSSim-based kinome tree (structure-based) shows large overlap with most kinase groups as annotated by Manning et al.[6] (sequence-based). We will summarize the KiSSim clusters and highlight differences in comparison to Manning's kinase groups AGC, CAMK, CK1, CMGC, STE, TKL, TK, the atypical group, and the unassigned kinases (Other).

Kinases from the *TK group* build a single large cluster with two outliers, i.e., the pseudokinases TYK2-b and JAK1-b. Known highly similar kinases, which form (sub)families in the Manning tree, are grouped together, e.g. the families Erb (EGFR, Erb2, Erb3, Erb4), Eph (EphB[1,4] and EphA[1,2,3,5,7,8]), JakA (JAK1, JAK2, JAK2, TYK2), and JakB (JAK1-b, TYK2-b).

Kinases from the *CAMK group* mainly cluster together. In addition, the following kinases from other kinase groups are included in our CAMK-like cluster: (a) CaMKK2 (Other), (b) MSK1 (AGC), (c) CK2a2 (CMGC), and (d) AurA, AurC, PLK4, TTK, and MPSK1 (Other). This is partly in agreement with the findings by Modi and Dunbrack[7] who have reassigned 10 kinases from Manning's Other group to the CAMK group, of which seven are part of the KiSSim dataset (AurA, AurC, CaMKK2, PLK1, PLK2, PLK3, and PLK4) and three are not (AurB, CaMKK1, PLK5). The KiSSim-based similarity of CaMKK2 to CAMK kinases is further supported by profiling data for the chemical probe SGC-STK17B-1, which targets both CaMKK2 and DRAK2 (part of the CAMK group).[9] Note that the following kinases belong to the CAMK group but are found outside of our CAMK-like cluster: (a) Trb1, (b) LKB1, and (c) PASK, PIM1, and PIM2.

Kinases from the *STE group* are assigned mostly to a single cluster that is, however, shared with kinases from many other kinase groups. The STE kinases MAP2K[1,4,6,7] and OSR1 are separated from the other STE kinases.

Kinases from the *CMGC group* are clustered in two subgroups: kinases from the CDK, CDKL, and MAPK families build one cluster, while kinases from the DYRK, SRPK, and CLK family build another. The CK2a2 kinase (CK2 family) is an outlier.

Kinases from the *TKL* group are mainly clustered together with kinases from the Other group but some are separated from the rest (DLK, BRAK, IRAK2, and LIMK1). Kinases from the *CK1 group* build one group except for TTBK1 and TTBK2. Kinases from the *AGC group* cluster together as well; MSK1 is the only outlier that is found closer to the CAMK kinases. Lastly, only three *atypical kinases* are included in the KiSSim dataset (ADCK3, RIOK1, and RIOK2) and build their own cluster, neighboring to the CK1 kinases.

Overall, the KiSSim dataset retrieves the sequence-based kinome tree by Manning et al. [6], including subbranches as discussed for the kinases assigned to the TK and CMGC groups. This is not surprising because we do encode the sequence in an abstracted manner in the physicochemical KiSSim fingerprint bits. However, some kinases show deviating relationships, of which some can be rationalized such as the CaMKK2 and DRAK2 relationship shown also in profiling data. Thus, the addition of structural information in the KiSSim fingerprint allows us to cluster more distantly related kinases. This aspect of the KiSSim tree is of interest because it predicts novel information on kinase similarities.

KiSSim evaluation using profiling data

As discussed, the KiSSim tree shows expected and unexpected kinase (dis)similarities. In order to evaluate the specificity and sensitivity of our method, we use profiling data as a surrogate for (real) expected kinase (dis)similarities: if a ligand targets a set of kinases with high activity, these kinases have similar binding sites and are therefore treated as similar kinases.

To this end, we pooled the Karaman et al. [8] and Davis et al. [53] datasets and filtered for FDA-approved inhibitors and their targets as listed in the PKIDB. [16] The dataset preparation is described in detail in the Bioactivity profiling data section. We show the KiSSim method's performance in the form of ROC curves for each inhibitor's listed targets.

For example, Imatinib has three reported on-targets (assigned in PKIDB) and two off-targets (based on activity data in the Karaman-Davis dataset); KiSSim's performance is

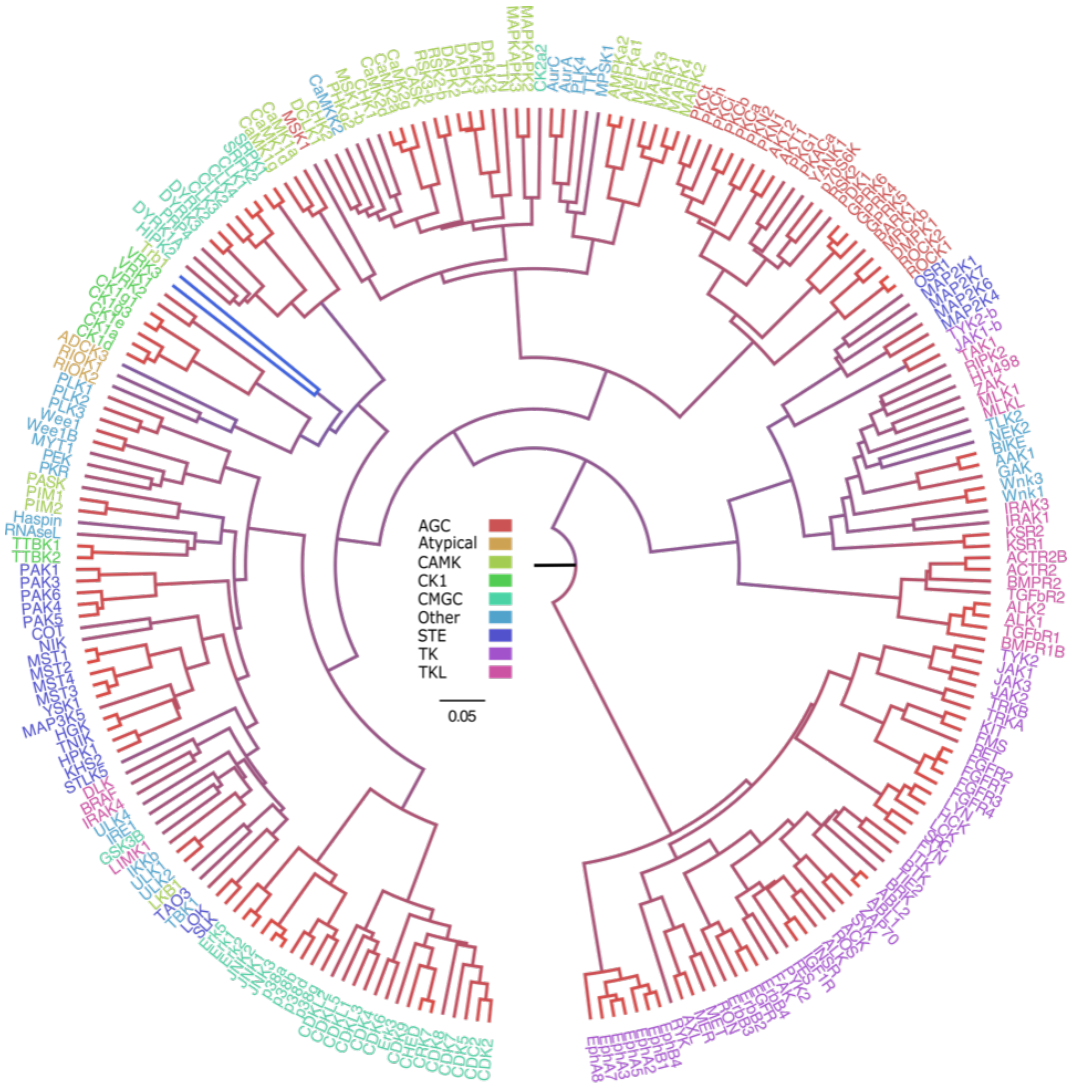

Figure 6: **KiSSim-based kinome tree** based on 257 structurally resolved kinases in the DFG-in conformation. Tree nodes are colored from red to blue showing small to large distances (0.01–0.20), describing high to low similarities; tree leaves represent kinases colored by kinase group. The tree is based on a clustering of the kinase distance matrix using as metric the Euclidian distance and as linkage Ward's criterion. The clusters are converted to the Newick format and visualized using FigTree.[33] See notebook for more details.[65]

evaluated by checking for these five Imatinib targets in KiSSim's most similar kinases to the on-targets (1) ABL1, (2) KIT, and (3) FMS, producing three ROC curves (Figure 7, first row, second plot). Details are described in the KiSSim evaluation using profiling data section. In total, we analyzed KiSSim's performance across 48 kinase-ligand pairs involving 21 ligands; the AUCs range from 0.49 to 1.0 with a mean of 0.75 ± 0.12. In the following, we discuss a few examples in Figure 7 (first row); please refer to the full set of ligands in Figure S5.

The *Erlotinib* profiling and KiSSim datasets share 50 kinases, of which 4 show high activity ($K_d \leq 100\,\text{nM}$), i.e., the on-target EGFR (TK, $K_d = 19.0\,\text{nM}$) and the off-targets SLK (STE, $K_d = 3.10\,\text{nM}$), LOK (STE, $K_d = 0.67\,\text{nM}$), and GAK (Other, $K_d = 0.67\,\text{nM}$). The top 20 KiSSim ranks for EGFR are dominated by TK kinases but include the STE kinases LOK and SLK on ranks 11 and 20 out of the 50 shared kinases, respectively; the GAK kinase is not detected by KiSSim, being found on rank 44 only (AUC = 0.641). The EGFR-GAK fingerprint pair shows many differences in their physicochemical bits, which stem from their relatively high pocket sequence dissimilarity (Figure 8). The fingerprint differences for the EGFR-GAK pair are visualized in 3D in Figure 9 for selected fingerprint features with high differences such as the HBA, aliphatic, and the hinge region features.

The *Imatinib* profiling and KiSSim datasets share 18 kinases, of which 5 TK kinases show high activity, i.e., the key target ABL1 as well as ABL2, LCK, KIT, and FMS. Compared to ABL1, all active kinases are ranked within KiSSim's top 7 most similar kinases (AUC = 0.908).

The *Bosutinib* profiling and KiSSim datasets share 108 kinases, of which 33 show high activity, mainly from the TK and STE groups. Compared to ABL1, which is one of the key targets, the TK kinases are found first in the top 35, followed by the STE kinases in the top 61 (AUC = 0.796).

The *Doramapimod* profiling and KiSSim datasets share 43 kinases, of which 8 show high activity, including the on-target p38a and four additional CMGC kinases (p38b, p38d, p38g,

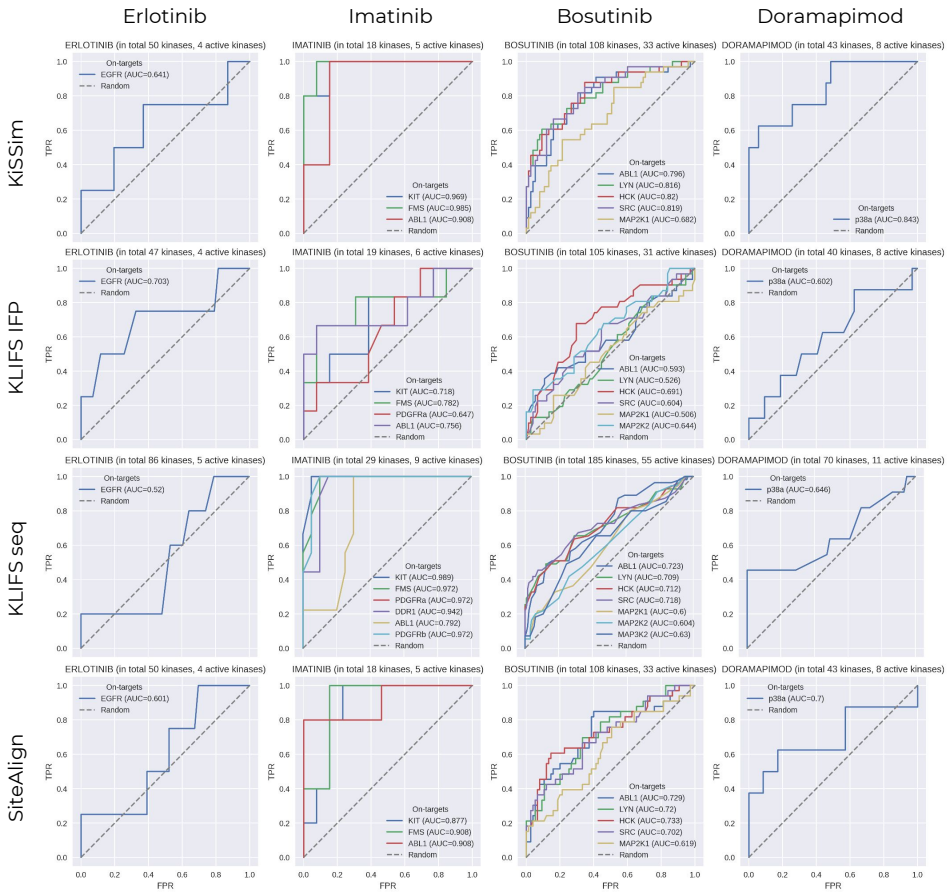

Figure 7: **Performance of KiSSim and other similarity measures against profiling data.** ROC curves comparing predicted and profiling-based kinase similarities (FPR = False positive rate; TPR = True positive rate). *Predicted similarities* against a selected kinase k are based on the KiSSim similarities (*KiSSim*), the KLIFS pocket IFP similarity (*KLIFS IFP*), the KLIFS pocket sequence identity (*KLIFS seq*), and the SiteAlign pocket structure similarity (*SiteAlign*). *Profiling-based kinase similarities* define kinases as similar if they are targeted by the same ligand with $K_d \leq 100\,\text{nM}$, including the ligand's on-target(s) as reported in the PKIDB. The kinases, for which the ligand shows lower activities with $K_d > 100\,\text{nM}$, are treated as dissimilar to the ligand's on-target(s). Find more details in the Bioactivity profiling data section. The first rank is always occupied by the kinase k. We show here only a selection of kinase-ligand pairs, please refer to Figures S5–S8 to inspect the full datasets. See notebooks for more details.[66–70]

and JNK2), two STE kinases (HGK and LOK), and the TK kinase TIE2. Compared to p38a, the CMGC kinases cover the top 7 KiSSim ranks, followed by the STE kinases and TIE2 in the top 25 (AUC = 0.845).

Furthermore, we performed the same profiling-based evaluations for subset KiSSim fingerprints, solely including residues that interact with the respective ligand were included. Ligand-interacting residues were selected from X-ray kinase structures based on the KLIFS IFP, i.e. 12, 57, 26, and 13 structures have cumulatively 21, 31, 27, and 35 interacting residues with Erlotinib, Imatinib, Bosutinib, and Doramapimod, respectively. In the case of Erlotinib and Bosutinib, the performance improves when including only the ligand-interacting residues — LOK, SLK, and the previously KiSSim-undetected GAK are all in the top 20 most kinase similarities compared to EGFR —, while the performance decreases slightly in the case of Imatinib and Doramapimod (see notebook[71] for more details). Thus, depending on the user's research question such as predicting off-target for one or multiple ligands of interest, known interaction profiles can be used to guide the selection of residues for the KiSSim fingerprint.

Note that the prediction tasks evaluated with the ROC curves may vary in difficulty: (a) Generally, only few data points are available for this analysis. (b) Erlotinib- vs. Imatinib-based evaluations stem from predictions across different kinase groups vs. within the TK group only. (c) Erlotinib- vs. Bosutinib-based evaluations are based on a dataset with a share of active kinases of 1 out of 10 and 1 out of 3, respectively.

Comparison of KiSSim to other methods

In the next step, we investigated all-against-all comparisons based on the KiSSim fingerprints, the KLIFS pocket sequence, KLIFS ligand-pocket interaction fingerprints (IFP), and the SiteAlign scores. The data preparation steps are described in detail in the KiSSim comparison to other methods section.

The KiSSim fingerprint contains physicochemical bits, which generalize the pocket sequence, and spatial bits, which consider the individual atom/residue positions in the under-

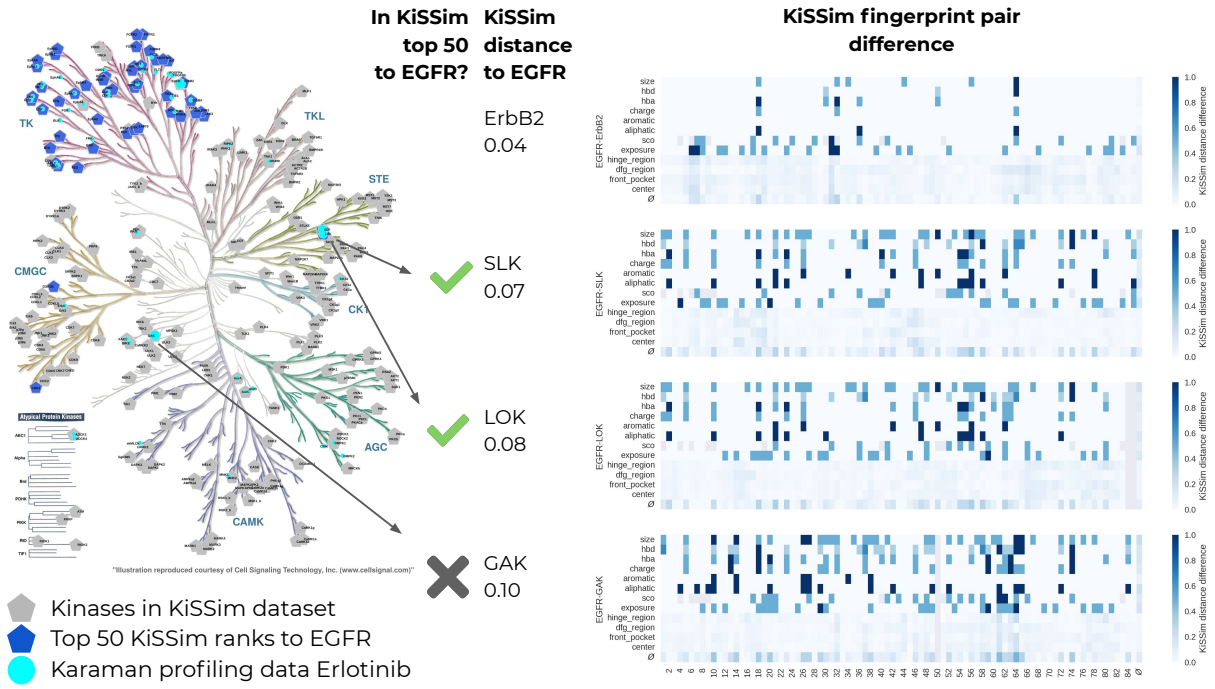

Figure 8: **KiSSim similarities between EGFR and Erlotinib's off-targets SLK, LOK, and GAK.** (left) The KinMap[56] tree shows the Karaman profiling data for Erlotinib (cyan), the top 50 most similar kinases to Erlobtinib's on-target EGFR (blue), and all kinases that are covered by the KiSSim dataset (grey). (right) KiSSim fingerprint pair differences between EGFR and selected kinases: ErB2 (as an example for highly similar kinases) as well as SLK, LOK, and GAK (unexpected off-targets for Erlotinib). Similarities between EGFR and SLK/LOK are detected by KiSSim (top 50 of all 279 kinases covered in KiSSim) while GAK stays undetected due to higher differences in the overall KiSSim fingerprints. See notebook for more details.[72]

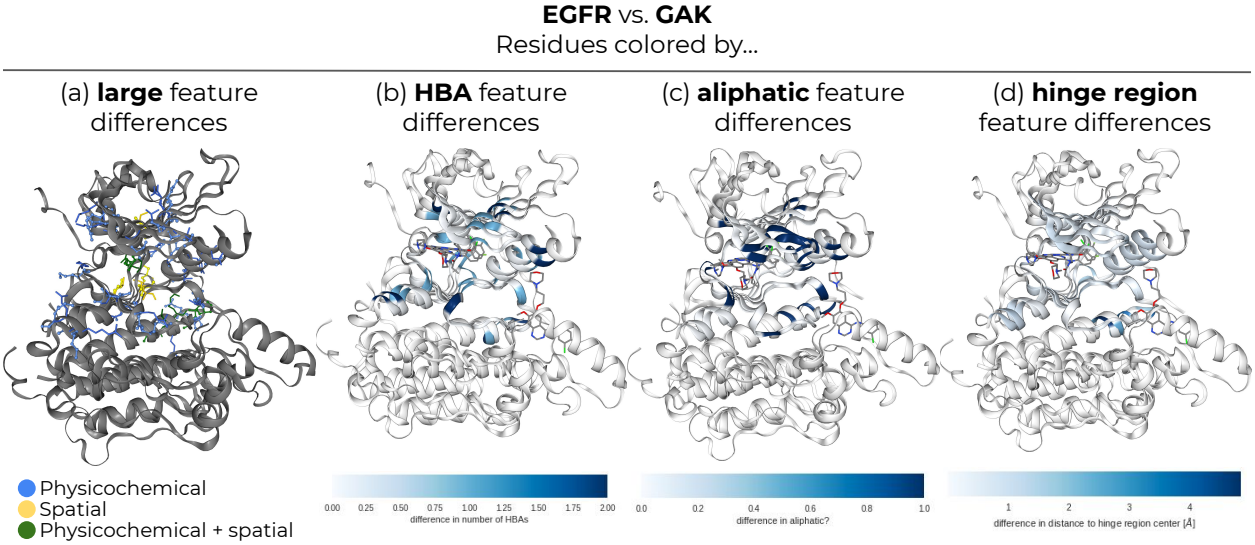

Figure 9: **3D visualization of KiSSim fingerprint differences between EGFR and GAK** (EGFR and GAK structure KLIFS IDs: 12159[73] and 10329,[74] respectively). (a) Highlight residues with at least one large difference in their physicochemical bits ($\Delta d_{normalized} = 0.6$, blue), spatial bits ($\Delta d_{normalized} = 0.2$, yellow), or both (green). Color residues by their differences in their (b) HBA, (c) aliphatic, and (d) hinge region feature, ranging from no difference (white) to highest difference (blue). See notebook for more details.[72]

lying kinase conformations. First, we use the KLIFS pocket sequence (*KLIFS seq*) to probe if the KiSSim fingerprint's generalized sequence and spatial information improve predictions compared to sequence information only. Second, we use the KLIFS pocket IFP (*KLIFS IFP*) to probe if the KiSSim fingerprint, which does not contain any information about interactions, improves kinase similarity predictions compared to interaction-based fingerprints. The advantage of IFPs is that they emphasize important residues and interactions as seen based on one or more ligands; the disadvantage is that not all possibly relevant interactions have been seen, yet. Note that combining the IFP information with KiSSim — using only interacting residues in the KiSSim fingerprint — can improve the KiSSim performance as discussed in the KiSSim evaluation using profiling data section. Third, we use kinase similarities calculated with the SiteAlign methodology (*SiteAlign*), from which we adapted some of the physicochemical KiSSim features, to confirm that the KiSSim fingerprint adds relevant kinase-focused information.

Correlation. We compared the pairwise kinase distances between the four different method setups (Figure S9). We observed a rather strong correlation between the KiSSim distances and (a) the KLIFS pocket sequence distances ($r = 0.77$), reflecting the sequence-generalizing physicochemical features in the KiSSim fingerprint, and (b) the SiteAlign distances ($r = 0.73$), reflecting the partly shared physicochemical features in KiSSim and SiteAlign (pharmacophoric and size features). In contrast, the correlation between KiSSim and KLIFS IFP distances is low ($r = 0.39$), possibly reflecting the lack of information on ligand-kinase interaction patterns.

Performance. We performed the same profiling analysis, which we discussed for KiSSim (mean AUC 0.75 ± 0.12) in the KiSSim evaluation using profiling data section, for the *KLIFS seq* (mean AUC 0.78 ± 0.15), *KLIFS IFP* (mean AUC 0.63 ± 0.12), and *SiteAlign* (mean AUC 0.71 ± 0.12) datasets, see Figure 7.

The KiSSim approach performs slightly worse compared to the KLIFS pocket sequence comparison in case of ligands like Imatinib, whose reported on-targets all belong to the TK

group, but shows better performance for Erlotinib, Bosutinib, and Doramapimod, which have known kinase targets belonging to different kinase groups. Hence, while the sequence-based approach picks up kinase group assignments as to be expected, KiSSim picks up more distant and less obvious off-targets.

The KLIFS pocket IFP comparison performs similarly to the KiSSim comparison in the case of Erlotinib, however, worse for the other three ligands. In contrast to the KiSSim approach, pocket similarities can only be detected by the IFP approach if the respective kinases have been co-crystallized with ligands that form similar interaction patterns. Such an IFP-based comparison probably can be more successful for a defined kinase set with high coverage of co-crystallized ligands in contrast to a kinome-wide comparison as performed here.

The SiteAlign methodology projects topological and chemical properties onto a sphere that sits in the center of a protein pocket. The spheres are aligned based on these projections and a similarity score is calculated between the aligned fingerprints. Finding the right alignment is a time-consuming step, hence we offered SiteAlign already the KLIFS-aligned structures as a starting point and reduced the iterations as described in the KiSSim comparison to other methods section. KiSSim outperforms the SiteAlign results in most cases, however, often not considerably much.

Runtime. The runtime for the methods discussed here differ considerably: Generating the *KLIFS seq* dataset takes about a second (based on about 500 kinases), while the *KLIFS IFP* dataset is ready within half a minute (based on about 8800 IFPs); both procedures build on the processed and curated KLIFS datasets, i.e. both the pocket sequences and the pocket interaction fingerprints are ready for use. Generating the KiSSim kinase matrix takes about 24 hours, while the all-against-all comparison with *SiteAlign* is ready after $>$ 20000 hours using the optimized SiteAlign settings (both based on over 4000 structures and a single-core/thread execution). Parallelization is built-in for the KiSSim approach to speed up the calculation.

Taking all these findings together, the KiSSim methodology compares well with established methods while often improving predictions between kinase pairs without an obvious relationship based on the sequence. The pocket sequence and IFP based methods are much faster than the structure-based methods KiSSim and SiteAlign, however, the overall kinase similarity assessment benefits from the added structural pocket information. KiSSim's setup and runtime are more convenient than for the SiteAlign method, however, KiSSim does rely on the KLIFS 85-residue pocket alignment.

Conclusion

We presented here the KiSSim (Kinase Structural Similarity) fingerprint as a novel structure-enabled pocket encoding tailored to kinase pockets. The fingerprint encodes physicochemical and spatial properties of the 85 KLIFS residues, which are aligned across the structurally covered kinome. On the one hand, the majority of physicochemical bits — size, HBD, HBA, charge, aromatic, and aliphatic, which are adapted from the SiteAlign method — encode the pocket sequence in a generalized, pharmacophoric way. On the other hand, the side chain orientation, solvent exposure, and the spatial bits — the distances to the pocket center and key subpocket centers and the distance distributions' moments — account for the structural conformation. Across all fingerprints, we saw that the fingerprint captures the physicochemical property variability (e.g., most residues are uncharged, whereas HBD/HBA features vary) and the conserved residue positions (e.g., distances to DFG region are more widely spread than to the hinge region).

We used the fingerprint to calculate all-against-all distances — small distances refer to high similarity, large distances to low similarity – within the structurally covered kinome: the DFG-in and DFG-out dataset consist of 4112 and 406 structures, representing 257 and 71 kinases, respectively. We found that the fingerprint can distinguish between intra- and inter-kinase similarities and between DFG-in and DFG-out structures.

Some kinases are represented by multiple structures, hence some kinase pairs are represented by multiple structure pairs. The distribution of structure distances for one kinase pair can be broad; we selected per kinase pair the closest structure pair that is experimentally observed. We clustered the resulting kinase distance matrix to produce a KiSSim-based kinome tree. While the tree reproduced large parts of the sequence-based Manning tree, some relationships could be observed that are unexpected from a sequence perspective only. For example, we found similarities between CaMKK2 (STE) and DRAK2 (CAMK), which are targeted by the same chemical probe SGC-STK17B-1;[9] we also could confirm the reassignment of AurA, AurC, PLK4, and CaMMK2 from the Other to the CAMK group as proposed by Modi and Dunbrack[7].

Besides the averaged tree view, we also investigated the top-ranked kinases given a query kinase to show that KiSSim can partially explain profiling data. While some ligand profiles are reflected completely in the KiSSim dataset (e.g., Imatinib), other ligand profiles are covered partially (e.g., Erlotinib's off-targets LOK and SLK are detected while GAK is not).

In comparison with other similarity measures — focusing on the pocket sequence (*KLIFS seq*), interaction profiles (*KLIFS IFP*), or topological- and chemical pocket properties (*SiteAlign*) — KiSSim performs equally or slightly better in most cases. The sequence- and IFP-based measures are easy and fast to compute thanks to the preprocessed kinase pockets available at KLIFS; we recommend to include these datasets in any case when investigating kinase similarities. SiteAlign is a powerful tool to compare pockets across all protein classes; if interested only in kinases, KiSSim is a kinase-focused and faster alternative with slightly better results in most of the investigated cases.

As for all structure-based methods, the imbalanced dataset of kinase structures is a challenge. Some kinases are structurally well represented (e.g., EGFR or CDK2), while others have only few structures available. And unfortunately still roughly half of the humane kinome has no structural information available at all. The recent breakthrough of AlphaFold2[75] could help here; predicted structures for almost all human kinases are available now on the

AlphaFold DB.[76] Modi and Dunbrack[77] have already classified the structures' conformations and found most structures in the DFG-in conformation. An AlphaFold-enhanced KiSSim tree may further increase the usefulness of the KiSSim methodology for kinome-wide similarity studies. Furthermore, the KiSSim fingerprint can be applied in machine learning, e.g. to extract the most important features in the kinase pocket.

We believe that the KiSSim fingerprint is a valuable tool for kinase research to explain and predict off-targets and polypharmacology. Since the code is open sourced and available as Python package, the KiSSim fingerprint can easily be integrated in other larger-scale workflows.

Code and data availability

- KiSSim library (`kissim`): `https://github.com/volkamerlab/kissim` and `https://kissim.readthedocs.io`
- KiSSim datasets: `https://doi.org/10.5281/zenodo.5774521`
- KiSSim application and analyses (`kissim_app`): `https://github.com/volkamerlab/kissim_app`

Author Contributions

Conceptualization, DS, AV; Data Curation, DS, AK; Formal Analysis, DS, EA, AK, AV; Funding Acquisition, AV; Investigation, DS, EA, AK, AV; Methodology, DS, EA, AV; Project Administration, DS, AV; Resources, AV; Software, DS; Supervision, AK, FR, AV; Validation, DS, AV; Visualization, DS; Writing: Original Draft, DS, AV; Writing: Review and Editing, DS, EA, AK, FR, AV.

Disclosures

Nothing to disclose.

Acknowledgement

DS thanks Talia B. Kimber for insightful and motivating discussions about the more mathematical aspects of this project. DS thanks Jaime Rodríguez-Guerra for bringing software best practices into the lab and for helpful and enthusiastic Python conversations. AV and DS gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (Grant VO 2353/1-1). AV acknowledges support from the Bundesministerium für Bildung und Forschung (Grant 031A262C). AV, DS, and EA thank the HPC service of ZEDAT, Freie Universität Berlin[78] for cluster time and support.

List of abbreviations

- KiSSim: Kinase Structural Similarity
- ATP: Adenosine triphosphate
- IFP: Interaction fingerprint
- DFG: Asparagine-phenylalanine-glycine
- MSA: Multiple sequence alignment
- SCO: Side chain orientation
- HBD: Hydrogen bond donors (here: number of HBD)
- HBA: Hydrogen bond acceptors (here: number of HBA)
- ROC: Receiver operating characteristic
- AUC: Area under the curve

References

(1) Cohen, P.; Alessi, D. R. Kinase Drug Discovery - What's Next in the Field? *ACS Chem. Biol.* **2013**, *8*, 96–104.

(2) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.* **2017**, *16*, 19–34.

(3) Cohen, P.; Cross, D.; Jänne, P. A. Kinase Drug Discovery 20 Years after Imatinib: Progress and Future Directions. *Nat. Rev. Drug Discov.* **2021**, *20*, 551–569.

(4) Morphy, R. Selectively Nonselective Kinase Inhibition: Striking the Right Balance. *J. Med. Chem.* **2009**, *53*, 1413–1437.

(5) van Linden, O. P.; Kooistra, A. J.; Leurs, R.; de Esch, I. J.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database to Navigate Kinase-Ligand Interaction Space. *J. Med. Chem.* **2014**, *57*, 249–277.

(6) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934.

(7) Modi, V.; Dunbrack, R. L. A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. *Sci. Reports* **2019**, *9*, 1–16.

(8) Karaman, M. W. et al. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(9) Structural Genomics Consortium, SGC-STK17B-1: A Chemical Probe for STK17B/DRAK2 Kinase. `https://www.thesgc.org/chemical-probes/SGC-STK17B-1`, [accessed 2021-08-16].

(10) Kooistra, A. J.; Volkamer, A. Kinase-Centric Computational Drug Development. *Annu. Rep. Med. Chem.* **2017**, *50*, 197–236.

(11) KinCore, Phylogeny of Human Protein Kinase Domains. `http://dunbrack3.fccc.edu/kincore/phylogeny`, [accessed 2021-08-11].

(12) Schmidt, D.; Scharf, M. M.; 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*, 629.

(13) Kuhn, D.; Weskamp, N.; Hüllermeier, E.; Klebe, G. Functional Classification of Protein Kinase Binding Sites Using Cavbase. *ChemMedChem* **2007**, *2*, 1432–1447.

(14) Kanev, G. K.; de Graaf, C.; Westerman, B. A.; de Esch, I. J. P.; Kooistra, A. J. KLIFS:

An Overhaul after the First 5 Years of Supporting Kinase Research. *Nucleic Acids Res.* **2021**, *49*, D562–D569.

(15) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure* **2012**, *20*, 391–396.

(16) Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23*, 908.

(17) Gaulton, A. et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

(18) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2008**, *71*, 1755–1778.

(19) Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A. KinFragLib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination. *J. Chem. Inf. Model.* **2020**, *60*, 6081–6094.

(20) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.

(21) Wilkinson, M. D. et al. The FAIR Guiding Principles For Scientific Data Management and Stewardship. *Scientific Data* **2016**, *3*, 160018.

(22) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(23) Hamelryck, T. An Amino Acid Has Two Sides: A New 2D Measure Provides a Different View of Solvent Exposure. *Proteins Struct. Funct. Bioinforma.* **2005**, *59*, 38–48.

(24) Volkamer Lab, OpenCADD. `https://github.com/volkamerlab/opencadd`, [accessed 2021-11-27].

(25) KiSSim, KLIFS Pocket Residue Subsets for DFG-in and DFG-out Conformations. `https://github.com/volkamerlab/kissim/blob/main/kissim/data/klifs_pocket_residue_subset.json`, [accessed 2021-11-11].

(26) Volkamer Lab, Extrema Used for Min-Max Normalization of KiSSim's Spatial Features. `https://github.com/volkamerlab/kissim/tree/v1.0.0/kissim/data`, Version 1.0.0.

(27) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; Chapter 2, pp 37–38, `http://www.deeplearningbook.org`.

(28) Rose, A. S.; Hildebrand, P. W. NGL Viewer: A Web Application for Molecular Visualization. *Nucleic Acids Res.* **2015**, *43*, W576–W579.

(29) Nguyen, H.; Case, D. A.; Rose, A. S. NGLView - Interactive Molecular Graphics for Jupyter Notebooks. *Bioinformatics* **2017**, *34*, 1241–1242.

(30) IPyWidgets, IPyWidgets Documentation. `https://ipywidgets.readthedocs.io/en/latest/`, [accessed 2021-10-05].

(31) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

(32) BioPython, Bio.Phylo package. `https://biopython.org/docs/latest/api/Bio.Phylo.html`, [accessed 2021-08-16].

(33) FigTree, FigTree. `http://tree.bio.ed.ac.uk/software/figtree/`, [accessed 2021-08-16].

(34) Anaconda Software Distribution, Anaconda Documentation. `https://docs.anaconda.com/`, [accessed 2021-07-30].

(35) Conda-Forge Community, The Conda-Forge Project: Community-Based Software Distribution Built on the Conda Package Format and Ecosystem. 2015.

(36) Raschka, S. BioPandas: Working with Molecular Structures in Pandas DataFrames. *The Journal of Open Source Software* **2017**, *2*.

(37) Harris, C. R. et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362.

(38) The Pandas Development Team, pandas-dev/pandas: Pandas. 2020.

(39) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

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

(41) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.

(42) numpydoc, numpydoc. `https://numpydoc.readthedocs.io/en/latest/format.html`, [accessed 2021-11-27].

(43) Python Software Foundation, Black: The Uncompromising Python Code Formatter. `https://github.com/psf/black`, [accessed 2021-10-06].

(44) Black-nb, Black-nb: The Uncompromising Code Formatter, for Jupyter Notebooks. `https://github.com/tomcatling/black-nb`, [accessed 2021-10-06].

(45) flake8, flake8. `https://flake8.pycqa.org/`, [accessed 2021-10-06].

(46) flake8-nb, flake8-nb. https://flake8-nb.readthedocs.io/, [accessed 2021-10-06].

(47) Read the Docs, Read the Docs. https://readthedocs.org/, [accessed 2021-07-31].

(48) sphinx, sphinx - Python Documentation Generator. https://www.sphinx-doc.org/, [accessed 2021-10-06].

(49) pytest, pytest. https://docs.pytest.org/, [accessed 2021-10-06].

(50) CodeCov, CodeCov. https://docs.codecov.com/docs, [accessed 2021-11-27].

(51) nbval, nbval. https://nbval.readthedocs.io/en/latest/, [accessed 2021-10-06].

(52) GitHub, GitHub Actions. https://docs.github.com/en/actions, [accessed 2021-10-06].

(53) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046–1051.

(54) Volkamer Lab, KiSSim notebook: KLIFS Data Preparation and Exploration. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/002_structures/001_prepare_dataset.ipynb, Version 1.0.0.

(55) Volkamer Lab, KiSSim Notebook: Loading KiSSim Results. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/001_quick_start/001_quick_start_kissim.ipynb, Version 1.0.0.

(56) 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*, 16.

(57) Sebastian Raschka, About Min-Max Scaling. https://sebastianraschka.com/Articles/2014_about_feature_scaling.html#about-min-max-scaling, [accessed 2021-11-27].

(58) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2006**, *47*, 195–207.

(59) Volkamer Lab, KiSSim Notebook: Feature Distributions. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/004_fingerprints/003_feature_distributions.ipynb, Version 1.0.0.

(60) Volkamer Lab, KiSSim Notebook: Subpocket Center Robustness. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/003_subpockets/002_subpocket_robustness.ipynb, Version 1.0.0.

(61) Volkamer Lab, KiSSim Notebook: Influence of Conformations on Subpockets. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/003_subpockets/003_subpocket_vs_conformations.ipynb, Version 1.0.0.

(62) Volkamer Lab, KiSSim Notebook: Can Fingerprint Distances Discriminate DFG Conformations? https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/005_comparison/004_fingerprint_distances_vs_dfg.ipynb, Version 1.0.0.

(63) Volkamer Lab, KiSSim Notebook: Fingerprint Distances Between Structures for the Same Kinase. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/005_comparison/005_structure_kinase_mapping.ipynb, Version 1.0.0.

(64) Illergård, K.; Ardell, D. H.; Elofsson, A. Structure is Three to Ten Times More Conserved Than Sequence - A Study of Structural Response in Protein Cores. *Proteins Struct. Funct. Bioinforma.* **2009**, *77*, 499–508.

(65) Volkamer Lab, KiSSim Notebook: KiSSim-Based Kinome Tree. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/005_comparison/006_kissim_kinome_tree.ipynb, Version 1.0.0.

(66) Volkamer Lab, KiSSim Notebook: Predict Ligand Profiling Using KiSSim (Pooled Karaman and Davis Dataset). https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/006_evaluation/004_profiling_karaman_davis.ipynb, Version 1.0.0.

(67) Volkamer Lab, KiSSim Notebook: Predict Ligand Profiling Using IFPs (Pooled Karaman and Davis Dataset). https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/006_evaluation/011_profiling_karaman_davis__ifp.ipynb, Version 1.0.0.

(68) Volkamer Lab, KiSSim Notebook: Predict Ligand Profiling Using Sequence (Pooled Karaman and Davis Dataset). https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/006_evaluation/012_profiling_karaman_davis__seq.ipynb, Version 1.0.0.

(69) Volkamer Lab, KiSSim Notebook: Predict Ligand Profiling Using SiteAlign (Pooled Karaman and Davis Dataset). https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/006_evaluation/013_profiling_karaman_davis__sitealign.ipynb, Version 1.0.0.

(70) Volkamer Lab, KiSSim Notebook: Compare AUC Values Between KiSSim and Other Methods. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/006_evaluation/014_comparative_analyses_auc.ipynb, Version 1.0.0.

(71) Volkamer Lab, KiSSim Notebook: KiSSim Matrix Only Based on Ligand-Interacting Residues. https://github.com/volkamerlab/kissim_app/blob/v1.0.

0/notebooks/006_evaluation/015_subset_kissim_fingerprints.ipynb, Version 1.0.0.

(72) Volkamer Lab, KiSSim Notebook: Fingerprint Bit Differences. https://github.com/volkamerlab/kissim_app/blob/v1.0.0/notebooks/005_comparison/007_fingerprint_diffs_3d.ipynb, Version 1.0.0.

(73) KLIFS, 6JRK - Chain A — Epidermal Growth Factor Receptor. https://klifs.net/details.php?structure_id=12159.

(74) KLIFS, 5Y80 - Chain A (Model A) — Cyclin G Associated Kinase. https://klifs.net/details.php?structure_id=10329.

(75) Jumper, J. et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583—-589.

(76) Tunyasuvunakool, K. et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590—596.

(77) Modi, V.; Dunbrack, R. L. Kincore: A Web Resource for Structural Classification of Protein Kinases and Their Inhibitors. *Nucleic Acids Res.* **2021**, *TBA*, TBA.

(78) Loris Bennett and Bernd Melchers and Boris Proppe, Curta: A General-Purpose High-Performance Computer at ZEDAT, Freie Universität Berlin. 2021.

Graphical TOC Entry