Protein Domain-Based Prediction of Compound–Target Interactions and Experimental Validation on LIM Kinases

Tunca Doğan[1,2,3,*], Ece Akhan Güzelcan[3,4], Marcus Baumann[5], Altay Koyas[3], Heval Atas[3], Ian Baxendale[6], Maria Martin[7] and Rengul Cetin-Atalay[3,8,*]

[1] Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

[2] Institute of Informatics, Hacettepe University, 06800 Ankara, Turkey

[3] CanSyL, Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey

[4] Center for Genomics and Rare Diseases & Biobank for Rare Diseases, Hacettepe University, 06230 Ankara, Turkey

[5] School of Chemistry, University College Dublin, D04 N2E2 Dublin, Ireland

[6] Department of Chemistry, University of Durham, DH1 3LE Durham, UK

[7] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, CB10 1SD Hinxton, Cambridge, UK

[8] Section of Pulmonary and Critical Care Medicine, University of Chicago, Chicago IL, 60637, USA

* To whom correspondence should be addressed.

E-mail: tuncadogan@hacettepe.edu.tr & rengul@uchicago.edu

Abstract

Predictive approaches such as virtual screening have been used in drug discovery with the objective of reducing developmental time and costs. Current machine learning and network-based approaches have issues related to generalization, usability, or model interpretability, especially due to the complexity of target proteins' structure/function, and bias in system training datasets. Here, we propose a new computational method "DRUIDom" to predict bio-interactions between drug candidate compounds and target proteins by utilizing the domain modularity of proteins, to overcome problems associated with current approaches. DRUIDom is composed of two methodological steps. First, ligands/compounds are statistically mapped to structural domains of their target proteins, with the aim of identifying physical or functional interactions. As such, other proteins containing the mapped domain or domain pair become new candidate targets for the corresponding compounds. Next, a million-scale dataset of small molecule compounds, including the ones mapped to domains in the previous step, are clustered based on their molecular similarities, and their domain associations are propagated to other compounds within the same clusters. Experimentally verified bioactivity data points, obtained from public databases, are meticulously filtered to construct datasets of active/interacting and inactive/non-interacting compound–target pairs (~2.9M data points), and used as training data for calculating parameters of compound–domain mappings, which led to 27,032 high-confidence associations between 250 domains and 8,165 compounds, and a finalized output of ~5 million new compound–protein interactions. DRUIDom is experimentally validated by syntheses and bioactivity analyses of compounds predicted to target LIM-kinase proteins, which play critical roles in the regulation of cell motility, cell cycle progression, and differentiation through actin filament dynamics. We showed that LIMK-inhibitor-2 and its derivatives significantly block the cancer cell migration through inhibition of LIMK phosphorylation and the downstream protein cofilin. One of the derivative compounds (LIMKi-2d) was identified as a promising candidate due to its action on resistant Mahlavu liver cancer cells. The results demonstrated that DRUIDom can be exploited to identify drug candidate compounds for intended targets and to predict new target proteins based on the defined compound–domain relationships. The datasets, results, and the source code of DRUIDom are fully-available at: https://github.com/cansyl/DRUIDom.

Author Summary

Drug development comprises several interlinked steps from designing drug candidate molecules to running clinical trials, with the aim to bring a new drug to market. A critical yet costly and labor-intensive stage is drug discovery, in which drug candidate molecules that specifically interact with the intended biomolecular target (mostly proteins) are identified. Lately, data-centric computational methods have been proposed to aid experimental procedures in drug discovery. These methods have the ability to rapidly assess large molecule libraries and reduce the time and cost of the process; however, most of them suffer from problems related to producing reliable biologically relevant results, preventing them from gaining real-world usage. Here, we have developed a new method called DRUIDom to predict unknown interactions between drugs/drug candidate compounds and biological targets by utilizing the modular structure of proteins. For this, we identify the domains, i.e., the evolutionary and functional building blocks of proteins, where these potential drug compounds can bind, and utilize this information along with protein domain annotations to predict new drug targets. We have tested the biological relevance of DRUIDom on selected proteins that play critical roles in the progression of numerous types of cancer. Cell-based experimental results indicated that predicted inhibitors are effective even on drug-resistant cancer cells. Our results suggest that DRUIDom produces novel and biologically relevant results that can be directly used in the early steps of the drug discovery process.

1. Introduction

Drug development is an expensive and lengthy process, the cost of developing a new drug in the USA has been estimated at about $1.8 billion and it takes on average 13 years [1]. One of the major factors affecting the cost is the attrition rate of drug candidates in late-stage development due to unexpected side effects and toxicity problems, arising from previously unknown off-target interactions [2]. Indeed, the identification of molecular interactions between drug compounds and the intended target biomolecule(s) is the key to understanding and generating improved molecular designs leading to greater specificity. In the last decades, systematic high throughput screening (HTS) of large collections of chemical compounds has been widely utilized with the purpose of efficient lead identification, as well as efficacy evaluation and toxicity assessment [3]. Despite its advantages over previous strategies, HTS is an expensive technique that can only be afforded by big pharma. Furthermore, considering the combinations between millions of small molecule drug candidate compounds and thousands of potential protein targets, the combinatorial number of experiments is extremely high, which is not possible to experimentally evaluate.

Over the last two decades, computational approaches have been developed with the objective of aiding experimental studies in drug discovery, defining a new field entitled "virtual screening" or "drug/compound – target protein interaction (DTI) prediction" [4-6]. Here, the aim is to predict unknown compound – target interactions with the construction and application of statistical models, using various types of molecular descriptors [7]. There are two distinct approaches to virtual screening. In the ligand-based approach, new chemical substances are predicted as binders of the intended target biomolecules. This is usually done by calculating molecular similarities between the drug/compound that is known to interact with the intended protein and other chemical substances in the library, thus, returning the most similar ones as predictions via “guilt by association” [8]. Since the predicted ligands of a target are usually limited to the compounds that are highly similar to its known ligands, discovering new scaffolds is difficult with this approach. In structure-based virtual screening methods, 3-D structural information of known ligand – receptor complexes are used to model the interactions and predict new DTIs with similar interactive properties [9]. Structure-based virtual screening is a costly process due to both highly intensive computational processes and challenges associated with obtaining 3-D structures of both protein and receptor-ligand complexes [2]. As a result, they are mostly limited to the well-characterized portion of the target protein space. New computational approaches have emerged to address these issues by adopting machine learning and/or network analysis techniques [10-14]. There are cases where the drug candidate compounds, first discovered by virtual screening, or via computer-aided drug discovery in general, became approved drugs [4,15].

DTI prediction methods usually require large training datasets (i.e., experimentally verified interaction information between compounds and proteins), to build accurate models. Bioactivity databases such as PubChem [16] and ChEMBL [17] curate and publish *in vitro* and *in vivo* bioassays, in the form of compound – target bioactivity measurements, which are used by DTI predictors as training data. The open-access data presented in these resources are extremely valuable for the research community; however, it is still difficult to find data concerning less-studied targets, which prevents building predictive models for these less common targets. Besides, the information in these databases is typically incomplete, meaning that there are many unknown interactions for the compounds and the targets presented in these resources, an aspect that is especially critical for estimating the off-target effects of the drug candidate compounds. Nevertheless, computational predictions concerning under-studied targets and never-before-targeted proteins is an important topic that may help researchers to assess the druggability of these proteins and develop new therapeutic approaches.

Modelling the interaction between compounds and proteins is a difficult task especially due to the fact that molecular interactions between proteins and compounds are complex, also, many proteins expressed by the human genome are yet to be structurally characterized. In this sense, it is critical to reduce the complexity to a level where the modelling is feasible, the required data is available at large scale and the results produced are biologically relevant. Proteins have modular structures made up of functional building blocks called domains. Domains can fold, function, and evolve independently from the rest of the protein [18]. Protein regions that correspond to domains are evolutionarily highly conserved since mutations in these functionally critical regions may lead to adverse consequences for the organism. Once they are identified on the structures of characterized proteins, domains can be detected (i.e., predicted) on structurally uncharacterized proteins by constructing domain sequence profiles and by searching for these profiles on the amino acid sequences of uncharacterized proteins [19,20]. Thanks to this application, domain/family annotation coverage is considerably high on the documented protein sequence space in the UniProt Knowledgebase (UniProtKB), i.e., 96.7% for UniProtKB/Swiss-Prot and 81.3% for UniProtKB/TrEMBL. A few literature studies have investigated the relationship between domains and small molecules within the perspective of drug discovery and repositioning. For instance, Li *et al.* characterized the experimentally known binding interactions between domains and small molecules using data from Protein Data Bank (PDB). Consequently, they constructed a drug-domain network and used this to interpret modules of similar ligands and domains [21]. Kruger *et al.* proposed a simple heuristic to map Pfam domains to small molecules using ChEMBL bioactivity data as the source. The authors investigated the structural relevance of the idea of mapping domains to Pfam profiles with statistical tests and concluded that their heuristic produced accurate results [22,23]. In a recent study, Kobren and Singh identified interactions between Pfam family/domain entries and various types of ligands using PDB co-complex structures. Their system InteracDome, employs the positional correspondence between Pfam HMMs and amino acid sequences of the protein chains in PDB structures, together with known ligand-binding regions on the same protein chains, to predict the interacting receptor-ligand pairs [24]. Despite generating highly accurate mappings, InteracDome's coverage is limited on the small molecule ligand side due to its reliance on PDB co-complex structures. These studies laid the foundation for the idea of associating small molecule binding to protein domains but they have neither proposed a complete end-to-end prediction pipeline, nor leveraged the advantage of using large-scale experimental bioactivity data accumulated in public databases such as PubChem and ChEMBL. Consequently, there is a clear requirement for new computational DTI prediction methods/tools, capable of producing reliable and consistent results by using all available data in data resources to aid experimental procedures in the field of drug discovery and repositioning.

In this study, we propose a new computational method called DRUIDom (DRUg Interacting Domain prediction) for the comprehensive prediction of interactions between drugs/drug-like compounds and target proteins to aid experimental and computational research in drug discovery and repositioning. DRUIDom is based on associating compounds (i.e., small molecule ligands) with complementary protein domains. The assumption behind the mapping between domains and compounds is that, either the binding region of the ligand is on the mapped structural domain(s), or there is a functional relationship between the two, so that the mapped domain is required for the corresponding bioactivity to occur. Consequently, it is highly probable that other proteins containing the mapped domain (or combination of domains) will possess the required structural/functional properties to interact with the compound of interest. DRUIDom employs a supervised modelling approach, where the manually curated DTI information in ChEMBL and PubChem databases are used in combination with the protein sequence and annotation information in the UniProtKB [25] and the InterPro databases [20], for the construction of the predictive model. The resulting predictions cover compound and human target protein spaces recorded in the above-listed data repositories. In DRUIDom, we also evaluated compound to domain pair mappings, in order to account for the cases where multiple domains are required for the indented ligand interaction.

Our focus here was developing a complete chemogenomics-based drug/compound – target protein interaction prediction system with a global perspective without focusing on certain target families. For this, we constructed a large source bioactivity dataset and applied a scoring-based heuristic to generate the compound – domain associations, which are then propagated to other drug-like compounds and potential target proteins in the massive chemogenomics space to produce DTI predictions at large scale. We believe this study will provide valuable information for estimating both novel on-target and off-target effects of drugs and drug candidate compounds.

With the aim of validating DRUIDom, we selected the PI3K/AKT/mTOR signalling pathway for our experimental use-case study. PI3K/AKT/mTOR pathway is altered during the progression of various cancer types [26]. Therefore, it is therapeutically relevant to target this pathway. In this sense, we analyzed interacting compound predictions for PI3K/AKT/mTOR pathway proteins, resulting in 116 novel ligand predictions for four targets (i.e., MDM2, VEGFA, LIMK1, and LIMK2).

The invasiveness of cancer cells is based on the changes in control mechanisms that regulate cytoskeletal remodeling and cell migration. LIMK proteins (i.e., serine/threonine-protein kinases) play important roles in metastasis by phosphorylating cofilin proteins which are involved in the dynamic remodeling of actin filaments [27]. Recent studies have shown that inhibition of LIMKs, combined with other kinase inhibitors, is effective for various tumor cells in terms of decreasing their proliferative and metastatic features [28]. LIMKs are required for the collective invasion by taking roles in invadopodium formation and extracellular matrix degradation in cancer cells [29,30]. It has been reported that an overexpressed LIMK1 in breast and prostate cancer cells resulted in increased cell motility, and invasion capacity was attenuated when the inhibitors of upstream regulators of LIMKs are administered [31]. Therefore, we focused on LIMK1 and LIMK2 proteins for the *in vitro* experimental validation of the proposed method. We synthesized both the 4 initially predicted compounds and their 4 novel derivatives. The bioactivities of these small molecule compounds were analyzed on transformed normal cells and cancer cell lines. The results of these experimental assays, which are described in the following sections, validated the computational predictions and indicate potential novel inhibitors for LIMK1 and LIMK2 proteins that can be further investigated for their anti-migratory effects.

2. Results

Our source/training dataset is composed of 2,869,943 drug/compound – target protein pair data points (1,637,599 actives and 1,232,344 inactives) between 1,033,581 compounds and 3,644 target proteins. Using drug/compound – target associations contained in this dataset, we first mapped compounds to domains, then, we produced DTI predictions by propagating mappings to new compounds and new proteins (Figure 1). Detailed information about the procedure is given under 4.2.1 of the Methods section. Below, we first explained the conducted main test together with its results (section 2.1), serving both as a guide to determine the mapping parameters/thresholds and as a predictive performance analysis of DRUIDom. This is followed by the detailed analysis of compound – domain pair mappings in comparison with single domain mappings (section 2.2), large-scale production of new drug/compound – target protein interaction predictions (section 2.3), a validation use-case study on hepatocellular carcinoma disease (section 2.4) with molecular docking of selected novel inhibitor predictions for LIMK proteins as an *in silico* validation of DRUIDom (section 2.4.1), and the wet-lab *in vitro* analysis of LIMK inhibition with the treatment of predicted inhibitors via chemical syntheses and cell-based assays (section 2.4.2).

Figure 1. (a) The overall representation of the drug/compound – target protein interaction prediction approach used in DRUIDom (the diagram only depicts the relationship in terms of physical binding; however, DRUIDom also covers functional relationships between domains and compounds); **(b)** drug/compound – domain mapping procedure and its scoring over two representative (c_1, c_2) toy examples.

2.1 Predictive Performance Analysis

The performance of DRUIDom was measured over the success of the mappings between the compounds and domains, since compound – domain mappings are at the core of the whole predictive process. As the reference benchmark (i.e., performance test) dataset, experimentally identified binding between proteins and small molecule compounds (i.e., co-complex structures) has been employed. For this, we used InteracDome (the non-redundant representable list - v0.3) mappings [24] as our reference (i.e., gold-standard / benchmark) dataset, and calculated the performance of our compound – domain mapping procedure, for arbitrarily selected mapping score threshold values. In the InteracDome representable non-redundant set, there are 15,593 high-quality mappings indicating the interactions between 2,375 Pfam family/domain entries and 1,522 drug-like small molecules. It is important to note that InteracDome focuses on the cases of physical binding, whereas we aimed to account for both physical and functional relationships between domains and small molecule compounds. The main reasons behind using InteracDome as the reference dataset for the performance analysis of DRUIDom was first, cases of physical binding obtained from PDB are reliable, and second, there is no ground-truth/reference dataset for functional relationships between domains and small molecule ligands, as far as we are aware.

To prepare the performance analysis dataset, we first extracted the intersecting domain entries and compounds between the InteracDome benchmark and our source bioactivity dataset, to carry out the performance analysis on the intersecting set. Out of the total 2,375 Pfam family/domain entries in the InteracDome, 1,043 were included in the target proteins in our source dataset, and thus, constitute the intersecting domain set. Pfam-InterPro entry relationships were used for the conversion from Pfam to InterPro. Two main contributing factors to the reduced intersecting domain set are, we only used domain type entries in InterPro (leaving family type entries out since there is no structural correspondence to family entries), whereas InteracDome included family type entries along with domains; and second, there were several Pfam entries without any correspondence in InterPro and many InterPro entries without corresponding Pfam signatures. Out of 1,522 compounds in the non-redundant representable InteracDome dataset, a total of 1,144 were included in our mappings, and thus, constitute the intersecting compounds set. The main reason behind the difference in numbers is that many of the ligands in the InteracDome were not drug-like small molecules; whereas, in our mappings, all of the ligands/compounds were drug-like, as they were obtained from ChEMBL and PubChem. Next, we extracted all compound – domain pairs in InteracDome that include the intersecting compounds and domains. Following the construction of the finalized benchmark dataset, we compared our compound – domain mappings constructed at different mapping score thresholds with the benchmark mappings, to observe what portion of the benchmark mappings can be retrieved. Thresholds were applied on the performance scores of our mappings, calculation of which are described in the Methods section 4.2.1. Thus, a threshold of 0.7 means all compound – domain mappings with a mapping score recall, precision, accuracy, and F1-score less than 0.7 are discarded. At each threshold, if a compound – domain pair in the benchmark dataset is also retrieved in our mappings, it is counted as a true positive (TP). If a benchmark pair could not be retrieved in our mappings, it is counted as a false negative (FN). If a pair in our mappings could not be found in the benchmark dataset, it is counted as a false positive (FP). Finally, if a potential compound – domain pair could not be found both in our mappings and in the benchmark dataset, it is counted as a true negative (TN).

Table 1 displays the results of the compound – domain mapping performance analysis. As shown, performance increases with the increasing mapping score thresholds; however, the coverage of the mappings, with respect to InteracDome, decreases simultaneously. This was expected since increasing the confidence thresholds eliminates more and more compound – domain mappings from our set, but the remaining mappings are more reliable. The coverage can be considered low even with the lowest confidence score threshold (i.e., coverage for ligands: 31% and for domains: 16.5%) due to the fact that experimental data sources behind InteracDome and our mappings are different from each other (i.e., co-crystal structures and measured bioactivities, respectively). Since the performance was calculated considering the intersecting compounds and domains at each score threshold, the performance gradually increases with the increasing threshold, in terms of all metrics. Both the ligand and domain coverage, at the score threshold (0.9) that yielded the highest performance, was around 1% of the InteracDome. Considering the trade-off between coverage and performance, we selected the confidence threshold of 0.5, which provided an acceptable performance (i.e., accuracy: 0.95 and MCC: 0.78) and an InteracDome coverage of compounds: ~5% and domains: ~6%. At this score threshold, our approach produced 27,032 mappings between 250 domains and 8,165 compounds/ligands. It is also important to check the coverage extensions yielded by our mappings over the InteracDome, which corresponds to the percentage of new domains and new ligands added to the mapping set. These new ligands and domains were not presented in the InteracDome dataset. For the selected confidence threshold (0.5), our mappings enriched the InteracDome dataset by

~19% for domains and ~707% for ligands. The extended coverage values indicate the added value of our approach. In this study, all of the steps followed after this point were carried out using the mapping set generated with the mapping score threshold of 0.5. However, in order to allow users to select other score thresholds, we have also shared a file in our repository that includes raw/non-filtered compound – domain mappings together with their mapping scores.

Table 1. Compound – domain mapping performance analysis results.

Mapping score threshold	# of retrieved:			Domain coverage (% of InteracDome)	Compound coverage (% of InteracDome)	Domain coverage extension (% of InteracDome)	Compound coverage extension (% of InteracDome)	Performance analysis results								
	Mappings	Domains	Compounds					TP	FP	FN	TN	Recall	Precision	Accuracy	F1-Score	MCC
0	3,245,943	1,018	215,432	31.0	16.5	66.6	18814.9	163	3,235	116	9,414	0.58	0.05	0.74	0.09	0.11
0.1	1,872,420	894	193,538	23.8	15.9	61.9	16901.7	120	453	68	5,362	0.64	0.21	0.91	0.32	0.33
0.2	548,679	759	95,934	15.7	13.2	57.0	8372.6	96	170	36	2,328	0.73	0.36	0.92	0.48	0.48
0.3	143,332	590	36,887	10.5	9.9	46.1	3214.5	87	82	10	1,127	0.90	0.51	0.93	0.65	0.65
0.4	36,112	299	13,408	6.5	7.8	22.1	1164.2	80	54	4	787	0.95	0.60	0.94	0.73	0.73
***0.5**	**27,032**	**250**	**8,165**	**4.8**	**6.4**	**19.2**	**707.3**	**72**	**37**	**2**	**622**	**0.97**	**0.66**	**0.95**	**0.79**	**0.78**
0.6	21,592	197	4,752	3.1	4.5	15.8	410.8	65	22	1	457	0.98	0.75	0.96	0.85	0.84
0.7	17,207	115	2,476	2.2	3.2	8.8	213.2	55	9	0	215	1.00	0.86	0.97	0.92	0.91
0.8	6,846	93	1,155	1.3	1.8	7.6	99.1	36	3	0	81	1.00	0.92	0.98	0.96	0.94
0.9	2,783	70	372	1.2	1.0	5.6	31.5	21	1	0	38	1.00	0.95	0.98	0.98	0.96
1	174	54	119	0.8	0.0	4.4	10.4	0	0	0	0	-	-	-	-	-

*The selected threshold and its results are shown in bold font.

2.2 Domain pair to compound mappings

Here, our aim was to observe if it would be possible to identify the cases where the presence of a single domain is not sufficient for the occurrence of the interaction with the intended compound, instead, an interface composed of multiple domains are required. Other possible explanations for the requirement of multiple domains would be the allosteric binding/regulation phenomenon [32], or just a complex functional relation. To analyze this process, we generated compound – domain pair mappings using the procedure explained at the end of Methods section 4.2.1. For this procedure, we used the "bag of domains" approach where the order of the domains on the protein sequence was not taken into account and all possible pair combinations were then generated and tested. The reason for this evaluation is that domains that are quite far away from each other on the linear protein sequence can be located very close to each other upon folding of the protein.

Following the procedure described in the Methods section 4.2.1 and the thresholding/filtering of mappings with the selected parameter values described in the Results section 2.1, 3,721 mappings were obtained between 1,456 compounds and 270 domain pairs. Next, these pairs were compared with single domain pairings of the same compounds, in terms of the mapping performance scores (e.g., C_1 – D_xD_y is compared to C_1-D_x and C_1-D_y where C1 represents a compound and D_xD_y represents a domain pair composed of the domains: D_x and D_y), to observe if there is any performance improvement by mapping a pair instead of a single domain (which is expected to provide more specific/defined interaction properties). In most of the cases, the performance of the domain pair mapping was the same as the mapping of the same compound to one of the single domains presented in the corresponding domain pair, which indicates that only a single domain is sufficient for the binding, and the other domain in the domain pair is just an extra (i.e., the second domain does not play a detectable role in the binding). We called these domain pair mappings "neutral domain pair associations". However, there were a few cases that domain pair mapping actually increased the association performance, namely "positive domain pair associations". To prepare the finalized compound – domain pair mapping set, all of the neutral associations were discarded, yielding only 22 positive associations between 10 compounds and 12 domain pairs. Below, we investigated one example from positive domain pair associations as a case study. The experimental bioactivity results of the case study were obtained from the ChEMBL database (document link: https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3621091), which was previously curated from the study by England *et al.* where the authors investigated potent inhibitors for KDM protein subfamilies [33].

The compound with the ChEMBL id “CHEMBL3621867” (link: https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL3621867) was mapped to a single InterPro domain record named: "JmjN domain" (id: IPR003349, description: domains frequently found in the jumonji family of transcription factors, link: https://www.ebi.ac.uk/interpro/entry/IPR003349) with the confusion matrix values TP:3, FN:0, FP:1 and TN:2 (recall:1.00, precision:0.75, accuracy:0.83, F1-core:0.86, and MCC:0.71), the false positive hit indicates that there is one protein that contains IPR003349 (gene: KDM4E, protein: "Lysine-specific demethylase 4E" in human, UniProt protein accession: B2RXH2, link: https://www.uniprot.org/uniprot/B2RXH2), which was recorded to be inactive against CHEMBL3621867 in ChEMBL database with a bioactivity value of IC_{50} = 79.4 μM (and thus reported as a false positive in our analysis since the above mentioned single domain mapping predicted B2RXH2 as a target of CHEMBL3621867). Similarly, the same compound (CHEMBL3621867) was mapped to another single InterPro domain record named: "Zinc finger, PHD-type" (id: IPR001965, description: a C4HC3 zinc-finger-like motif found in nuclear proteins thought to be involved in chromatin-mediated transcriptional regulation, link: https://www.ebi.ac.uk/interpro/entry/IPR001965) with values TP:3, FN:0, FP:1 and TN:2 (recall:1.00, precision:0.75, accuracy:0.83, F1-core:0.86 and MCC:0.71), indicating that, again, there is one protein that contains IPR001965 (gene: KDM2A, protein: "Lysine-specific demethylase 2A" in human, UniProt protein accession: Q9Y2K7, link: https://www.uniprot.org/uniprot/Q9Y2K7), which was recorded to be inactive against CHEMBL3621867 in ChEMBL database with a bioactivity value of IC_{50} = 50.1 μM (and thus reported as a false positive in our analysis since the above mentioned single domain mapping would predict Q9Y2K7 as a target of CHEMBL3621867). However, the mapping between CHEMBL3621867 and the domain pair IPR003349-IPR001965 yielded an excellent mapping performance with metrics TP:3, FN:0, FP:0 and TN:3 (recall:1.00, precision:1.00, accuracy: 1.00, F1-core: 1.00 and MCC: 1.00), by eliminating the false positive target predictions of B2RXH2 and Q9Y2K7 for CHEMBL3621867. The domain pair IPR003349-IPR001965 is presented in 3 reviewed human protein entries among 6 proteins with measured activities against CHEMBL3621867 (i.e., Lysine-specific demethylases 4C, 5C and 4A, genes: KDM4C, KDM5C, and KDM4A, UniProt protein accessions: Q9H3R0, P41229, and O75164), all of which were targets of the corresponding compound verified in their respective binding assays with bioactivities of IC_{50} = 7.9, 6.3 and 5.0 μM, respectively. The protein that was accurately predicted as inactive by both single domain and domain pair mappings (i.e., as a true negative) was "Lysine-specific demethylase 6B" (gene: KDM6B, UniProt protein accession: O15054), which neither possessed IPR003349 nor IPR001965. This target also received a bioactivity measurement of IC_{50} = 63.1 μM against CHEMBL3621867. IPR003349 domain is annotated to 10 reviewed human protein entries in the UniProtKB/Swiss-Prot database, also, IPR001965 domain is annotated to 88 reviewed human protein entries. Whereas together, IPR003349-IPR001965 domains are annotated to 7 reviewed human protein entries. Due to sequence differences between KDM subfamily proteins (i.e., only 6 identical positions and 39 similar positions out of more than 1500 positions in the multiple sequence alignment of 6 KDM subfamily proteins), their domain annotations are different from each other, which is possibly reflected in their 3-D structure (although it is not possible to be sure without a crystal structure), and thus, the interaction with the corresponding compound (i.e., CHEMBL3621867).

It is important to note that, proteins annotated with only one of the domains listed above (i.e., IPR003349 or IPR001965) are also targeted by CHEMBL3621867; however, corresponding IC50s are way beyond plausible bioactivity values accepted for potential drug candidates (i.e., < 10 μM). On the other hand, the presence of both domains on the target protein yielded IC50 values that are within the acceptable range. This predicted domain pair – compound mapping (or any association predicted by DRUIDom) does not directly state a true physical binding between the mapped domains and the compound, it only suggests a relationship between the two entities (i.e., either physical or a functional interaction), where the interaction is stronger in the cases with the presence of both domains. Thus, targeting KDM subfamily proteins containing both IPR003349 and IPR001965 with CHEMBL3621867 would have a higher chance of success in a drug discovery study.

It is probable for Q9Y2K7 (KDM2A) protein to partially possess the IPR003349 domain at the N-terminal side. If this is the case, the InterProScan tool might not report the hit due to obtaining a low score under the default statistical cut-off value. To analyze the case, we locally aligned (using Smith-Waterman with default parameters of gap open: 10, gap extend: 0.5, and scoring matrix: BLOSUM62) the first 100 N-terminal residues of Q9Y2K7 (KDM2A) and O75164 (KDM4A), which is reported to possess IPR003349 between the positions 13 and 56 according to InterPro (https://www.ebi.ac.uk/interpro/protein/UniProt/O75164/). The output alignment reported a statistically significant hit (with 53.6% similarity between two sequences along the alignment length of 28 residues) between KDM4A sequence positions 11 and 38, which roughly spans the half of the IPR003349 domain, indicating the partial existence of the domain on Q9Y2K7 (KDM2A). Nevertheless, the partial existence of the domain may be the reason behind observing interaction with a rather high bioactivity value (i.e., IC_{50} = 50.1 µM). It is not possible for us to further comment on the physical binding as there is no co-crystal structure of a KDM subfamily protein with CHEMBL3621867.

Besides single domains and domain pairs, it is also possible for some of the compound – target interactions to require (either physically or functionally) three or even more domains to be presented at the target protein. We could not account for these cases in DRUIDom since they dramatically increase the complexity of the analysis, as a result, we chose to omit the cases with more than 2 domains.

2.3 Predicting New Drug/Compound – Target Protein Interactions

Drug/compound – target protein interaction predictions were generated by propagating the drug/compound – single domain (or domain pair) mappings to proteins and other compounds, using the procedure explained in Methods section 4.2.2. The crossing of new compounds and targets for each mapping has led to a geometric increase in the number of associations/predictions. Finally, a simple post-processing filter was applied to predictions for removing the known/recorded compound – target protein interactions from the prediction set.

First, 3,672,076 novel interactions (between 8,158 compounds and 5,563 proteins) were generated with the propagation of single domains to proteins (i.e., 250 domains to 5,563 proteins). Also, 631 novel interactions (between 9 compounds and 286 proteins) were produced with the propagation of domain pairs to proteins (i.e., 12 domain pairs to 286 proteins). The low number of predictions with domain pairs was due to the elimination of the domain pair mappings that did not display a performance increase over the single domain mappings of the same compound. At this point, the merged prediction dataset contained 3,672,220 novel interactions between 8,163 compounds and 5,563 proteins, after the removal of duplicates. The finalized prediction dataset was obtained following the propagation of the compounds in the previous prediction set to significantly similar compounds according to molecular similarity-based compound clusters, which yielded 5,050,841 novel interactions between 10,944 compounds and 5,461 proteins in the finalized prediction dataset, following the removal of known interactions. One notable observation is that there was only a slight increase in the number of compounds (from 8,163 to 10,944) after the pairwise molecular similarity-based propagation, which can be explained by the strict Tanimoto threshold of 0.8, which only passes the most reliable predictions.

2.4 Validation of Predicted Molecular Interactions

To select inhibitory compound predictions for *in silico* and *in vitro* experimental validation, we first checked our large-scale drug/compound – target interaction prediction dataset and found 116 inhibitor predictions for PI3K/AKT/mTOR signalling pathway proteins (Table 2), mainly due to the critical role of this pathway in various types of cancer [26]. Out of these, 4 compounds have been predicted as inhibitors of both LIMK1 and LIMK2 proteins (serine/threonine-protein kinases taking important roles in metastasis by phosphorylating cofilin proteins [27]). Structures of these compounds are given in Figure 2 together with their ChEMBL database identifier and short names as used in this study. These compounds are associated with LIMKs over their "Serine-threonine/tyrosine-protein kinase, catalytic domain" (InterPro domain id: IPR001245). In addition, we designed, synthesized, and tested 4 novel derivatives of the most active compound LIMKi-2 (Figure 2, compounds LIMKi-2a-d).

Table 2. Inhibiting compound predictions for PI3K/AKT/mTOR pathway proteins: MDM2, VEGFA, LIMK1 and LIMK2; given as ChEMBL molecule identifiers and gene names of the corresponding targets.

Predicted Compound (ChEMBL id)	Target Protein (Gene Name)	Predicted Compound (ChEMBL id)	Target Protein (Gene Name)
CHEMBL1316589	LIMK1	CHEMBL505899	MDM2
CHEMBL1512352	LIMK1	CHEMBL506261	MDM2
CHEMBL516650	LIMK1	CHEMBL506263	MDM2
CHEMBL518653	LIMK1	CHEMBL506507	MDM2
CHEMBL1316589	LIMK2	CHEMBL506623	MDM2
CHEMBL1512352	LIMK2	CHEMBL506646	MDM2
CHEMBL516650	LIMK2	CHEMBL506647	MDM2
CHEMBL518653	LIMK2	CHEMBL506740	MDM2
CHEMBL1241424	MDM2	CHEMBL507004	MDM2
CHEMBL1241425	MDM2	CHEMBL507649	MDM2
CHEMBL1241426	MDM2	CHEMBL508126	MDM2
CHEMBL1243385	MDM2	CHEMBL508377	MDM2
CHEMBL1242922	MDM2	CHEMBL508398	MDM2
CHEMBL458791	MDM2	CHEMBL508486	MDM2
CHEMBL514738	MDM2	CHEMBL508491	MDM2
CHEMBL515347	MDM2	CHEMBL508564	MDM2
CHEMBL515848	MDM2	CHEMBL508902	MDM2
CHEMBL516172	MDM2	CHEMBL508983	MDM2
CHEMBL475670	MDM2	CHEMBL509409	MDM2
CHEMBL481213	MDM2	CHEMBL509666	MDM2
CHEMBL481421	MDM2	CHEMBL510017	MDM2
CHEMBL1791379	MDM2	CHEMBL510066	MDM2
CHEMBL1791380	MDM2	CHEMBL510233	MDM2
CHEMBL1791382	MDM2	CHEMBL510473	MDM2
CHEMBL219860	MDM2	CHEMBL510817	MDM2
CHEMBL434556	MDM2	CHEMBL511030	MDM2
CHEMBL427239	MDM2	CHEMBL524509	MDM2

CHEMBL1791381	MDM2	CHEMBL524659	MDM2
CHEMBL445253	MDM2	CHEMBL524691	MDM2
CHEMBL505051	MDM2	CHEMBL524856	MDM2
CHEMBL503520	MDM2	CHEMBL524887	MDM2
CHEMBL207341	MDM2	CHEMBL524908	MDM2
CHEMBL443697	MDM2	CHEMBL525014	MDM2
CHEMBL446284	MDM2	CHEMBL525018	MDM2
CHEMBL450322	MDM2	CHEMBL525040	MDM2
CHEMBL451424	MDM2	CHEMBL525045	MDM2
CHEMBL451944	MDM2	CHEMBL525060	MDM2
CHEMBL454229	MDM2	CHEMBL525201	MDM2
CHEMBL486090	MDM2	CHEMBL525263	MDM2
CHEMBL499121	MDM2	CHEMBL525265	MDM2
CHEMBL499749	MDM2	CHEMBL525594	MDM2
CHEMBL499766	MDM2	CHEMBL525614	MDM2
CHEMBL500441	MDM2	CHEMBL525624	MDM2
CHEMBL500788	MDM2	CHEMBL525635	MDM2
CHEMBL501541	MDM2	CHEMBL525636	MDM2
CHEMBL503191	MDM2	CHEMBL526187	MDM2
CHEMBL503489	MDM2	CHEMBL526336	MDM2
CHEMBL503730	MDM2	CHEMBL526337	MDM2
CHEMBL503983	MDM2	CHEMBL526381	MDM2
CHEMBL504226	MDM2	CHEMBL526861	MDM2
CHEMBL504266	MDM2	CHEMBL527080	MDM2
CHEMBL504423	MDM2	CHEMBL527084	MDM2
CHEMBL504493	MDM2	CHEMBL1089944	VEGF
CHEMBL504855	MDM2	CHEMBL1689394	VEGF
CHEMBL504919	MDM2	CHEMBL499790	VEGF
CHEMBL505501	MDM2	CHEMBL501558	VEGF
CHEMBL505622	MDM2	CHEMBL508411	VEGF
CHEMBL505790	MDM2	CHEMBL509774	VEGF

Figure 2. Structures, database identifiers, and 2-D representations of predicted LIMK inhibitory compounds (LIMKi-1, 1a, 2, and 3) and derivatives (LIMKi-2a, b, c, and d).

2.4.1 Molecular Docking of Novel LIMK Inhibitors

For *in silico* validation of computationally predicted LIMK inhibitors, molecular docking analyses were conducted. LIMK proteins (LIMK1 and LIMK2) are serine/threonine kinases with multidomain structures including 2 LIM zinc-binding domains, 1 PDZ domain and 1 protein kinase domain. Multi-kinase inhibitor staurosporine and previously described LIMK inhibitor 9D8 have published crystal structures with the kinase domains of LIMK1 and LIMK2 proteins. These molecules were used as reference for docking to compare their binding free energies (ΔG) with the computationally predicted and novel LIMK inhibitors. In addition to computationally predicted compounds (i.e., LIMKi-1, LIMKi-1a, LIMKi-2 and LIMKi-3), novel derivatives of LIMKi-2 (i.e., LIMKi-2a, LIMKi-2b, LIMKi-2c and LIMKi-2d) were also docked against kinase domains of LIMK1 and LIMK2 proteins. AutoDock grid box parameters used in these analyses are displayed in Table 3a, and the docking results of each LIMK protein – compound combination are shown in Table 3b, which displays the lowest binding free energy calculation at the best pose obtained either from rigid or flexible docking in AutoDock. All files and results of the docking analysis, including the ones for online MTiAutoDock and SwissDock docking runs, are available in the data repository of this study. Based on the results in Table 3b; LIMKi-2, LIMKi-2d, and LIMKi-3 have binding free energy values close to that of the reference ligand staurosporine ("staurosporine" ΔG=-10.55 kcal/mol, Ki=18.47 nM; "9D8" ΔG=-12.38 kcal/mol, Ki=0.837 nM) for the LIMK1 protein, where the lower values indicate stronger interactions. As for the LIMK2 protein, binding free energy values for all ligands, except LIMKi-1 and LIMKi-1a, were around the generally accepted thresholds to assume a potential activity (i.e., -10 to -12 kcal/mol), which were close to the value of reference ligand 9D8 (i.e., -12.38 kcal/mol). In Figure 3, the best poses of LIMKi-2 and LIMKi-3 dockings against kinase domain binding sites of LIMK proteins are visualized along with the docking of reference molecules. The results indicate computationally predicted LIMK inhibitors, especially LIMKi-2 (including its derivatives) and LIMKi-3, could be promising candidate molecules for targeting LIM kinases.

Table 3. (a) Grid box parameters for AutoDock in the molecular docking analysis; **(b)** molecular docking results of computationally predicted LIMK inhibitors and their derivatives against kinase domains of LIMK proteins in terms of binding free energy (ΔG) and inhibition constant (*Ki*) estimations at the best pose.

(a)

	# of points in x-y-z dimension	**Spacing (angstrom)**	**x, y, z centers**
LIMK1 rigid docking	60-60-40	0.375	14.878, 6.646, 34.402
LIMK1 flexible docking	80-80-60	0.375	14.878, 6.646, 34.402
LIMK2 rigid docking	60-60-40	0.375	25.016, -13.952, 17.984
LIMK2 flexible docking	80-80-60	0.375	25.016, -13.952, 17.984

(b)

	ΔG (kcal/mol)		***Ki* (nM)**	
	LIMK1	**LIMK2**	**LIMK1**	**LIMK2**
Native ligands*	-10.55	-12.38	18.47	0.837
LIMKi-1	-7.68	-9.9	2340	55.14
LIMKi-1a	-7.47	-9.34	3330	142.42
LIMKi-2	-10.11	-12.07	38.73	1.43
LIMKi-2a	-9.74	-11.32	72.38	5.01
LIMKi-2b	-9.13	-11.01	203.95	8.52
LIMKi-2c	-9.67	-11.92	82.22	1.83
LIMKi-2d	-10.28	-12	28.94	1.61
LIMKi-3	-10.03	-11.92	44.34	1.82

*Native ligands correspond to small molecule compounds staurosporine and 9D8 for LIMK1 and LIMK2, respectively.

Figure 3. Visualization of the docked complex structures of **(a)** LIMK1 kinase domain in complex with the reference molecule staurosporine (green), LIMKi-2 (violet), and LIMKi-3 (red), and **(b)** LIMK2 kinase domain in complex with the reference molecule 9D8 (dark cyan), LIMKi-2 (violet), and LIMKi-3 (red) at the best poses. Hydrogen bonds are displayed with dark blue lines. Gold and pink colors represent LIMK1 and LIMK2 protein residues interacting with the corresponding compounds.

2.4.2 *In vitro* Experimental Analysis of LIMK Inhibition

LIMKi Compounds have inhibitory effects on human cancer cells

To address whether predicted inhibitors have cytotoxic effects on transformed normal human (HEK-238) and various epithelial cancer cell lines (e.g., MCF-7, HCT116, Huh7, and Mahlavu), cells were treated with LIMKi compounds with a concentration gradient of 40 μM to 2.5 μM for 72 hours. The resulting cytotoxic IC_{50} values are given in Table 4a. While there is no cytotoxicity observed on normal cells, LIMKi-2 and LIMKi-3 compounds display cytotoxic activities between 5.5-17.3 μM on cancer cells. Since LIMKi-2 showed the most potential bioactivity, we synthesized four novel derivatives of LIMKi-2 and assessed their bioactivities on Huh7 and Mahlavu liver cancer cells. LIMKi-2 derivatives; 2c, 2d displayed cytotoxic activities on Huh7 and Mahlavu cells (~8μM and <20μM, respectively), while LIMKi-2a had no effect (Table 4b).

Table 4. Cytotoxic bioactivities of LIMKi molecules on human cells: **(a)** LIMKi-1,3 compounds **(b)** LIMKi-2 derivatives.

(a)

LIMKi molecules	**IC_{50} Values (μM)**			
	LIMKi-1	**LIMKi-1a**	**LIMKi-2**	**LIMKi-3**
HEK-293 (Transformed Normal Human Embryonic Kidney Cell Line)	NI	NI	NI	NI
MCF-7 (Breast Cancer Cell Line)	NI	NI	6.4 ± 1.0	5.5 ± 0.3
HCT116 (Colon Cancer Cell Line)	NI	NI	5.6 ± 1.3	6.8 ± 1.2
Huh7 (Liver Cancer Cell Line)	NI	NI	7.9 ± 0.7	9.4 ± 1.2
Mahlavu (Liver Cancer Cell Line)	NI	NI	13.8 ± 0.8	17.7 ± 0.3

(b)

LIMKi-2 derivatives	**IC_{50} Values (μM)**

	LIMKi-2a	LIMKi-2b	LIMKi-2c	LIMKi-2d
Huh7 (Liver Cancer Cell Line)	NI	28.4 ± 2.5	8.2 ± 1.4	7.06 ± 0.8
Mahlavu (Liver Cancer Cell Line)	NI	24.6 ± 1.0	15.9 ± 3.1	15.3 ± 1.3

As stated above, phosphorylated LIMK proteins are involved in actin cytoskeleton dynamics through cofilin phosphorylation, hence we performed experiments on the migration and invasion properties of liver cancer cells in the presence of LIMK inhibitors. We focused on Huh7 and Mahlavu liver cancer cells for the rest of the study, because primary liver cancer (hepatocellular cancer, HCC) usually presents with multiple tumors within the liver and intrahepatic metastatic spread is a major problem for this cancer [34].

LIMKi compounds are effective in vitro by reducing the level of cofilin phosphorylation

Cofilin is a downstream molecule and its function is regulated by LIMK. Hence, we assessed phospho-Cofilin protein levels in Huh7 and Mahlavu cells in the presence of LIMK inhibitors. Phosphorylation of cofilin by LIMKs is significantly reduced upon treatment with LIMK inhibitors in both Huh7 and Mahlavu cells except for LIMKi-1 and LIMKi-2d, respectively (Figure 4a, b). Mahlavu cells are reported to have a resistant phenotype due to PTEN tumor-suppressive protein deficiency for migration [35]. Therefore, the differential response against LIMK inhibitors by well-differentiated Huh7 cells and poorly differentiated drug-resistant Mahlavu cells are as expected and allows us to better assess the dose-response of LIMK inhibitors.

The ratio of phosphorylated to non-phosphorylated Cofilin protein levels, together with LIMK protein phosphorylation was previously reported as an indication of the metastatic potential of a cell [27]. Therefore, we also checked the ratio of phospho- to total Cofilin levels for both Huh7 and Mahlavu cells (Figure 4a, b) and found that LIMK inhibitors decreased the phospho-Cofilin ratio significantly. These results may lead to the discovery of novel therapeutic agents against the metastatic capacity of hepatocellular carcinoma cancer cells.

Figure 4. Phospho-Cofilin protein expression. **(a)** Huh7 and **(b)** Mahlavu cells were cultured with LIMK inhibitors (20 μM) for 48 hours and expression of active p-Cofilin and total Cofilin levels were assessed with western blot analysis. The bar graph indicates the relative intensity of p-Cofilin levels compared to untreated DMSO controls. The equal loading control was analyzed based on the total protein staining normalization protocol. The ratio of phospho- and total Cofilin levels for both Mahlavu and Huh7 cell lines were calculated.

LIMK inhibitors significantly reduce migration and invasion of HCC cells in vitro

LIMK/Cofilin/ADF cascade has been described as one of the major regulators for actin cytoskeleton dynamics and reorganization [36]. Bioactivities of LIMKi compounds were tested for their effects on the migration and invasion capacity of HCC cell lines by wound healing and real-time cell invasion Transwell assays, respectively. First, Huh7 cell migration was analyzed in the presence of predicted LIMK inhibitors 1, 1a, 2, and 3. Huh7 cells have less migration capability compared to Mahlavu cells, so Huh7 migration was only tested with the originally predicted molecules. LIMKi-2 and LIMKi-3 strongly reduced the migration (2% gap closure) of Huh7 cells when compared to DMSO controls (48% gap closure) within 10 hours (Figure 5a). Then LIMKi-1, LIMKi-1a, LIMKi-2, LIMKi-3 and LIMKi-2 derivatives were tested on the migration of Mahlavu cells. LIMKi-2 derivatives reduced the resistant Mahlavu cell migration by 2.6-3.7 folds when compared to DMSO controls (Figure 5b).

We also tested the bioactivities of predicted compounds and their derivatives by real-time cell invasion for 48 hours on Huh7 and Mahlavu cells. Figure 6 indicates that LIMKi-2d was the most significant compound in terms of reducing the invasion capacity of both Mahlavu and Huh7 cell lines after 12 hours of treatment and throughout 48 hours. LIMKi-2c also significantly reduced Huh7 cell invasion.

Figure 5: Wound healing assay. *In vitro* “wound” was created by a straight-line scratch across the monolayer **(a)** Huh7, **(b)** Mahlavu cells. Then cells were treated with indicated concentrations of LIMKi compounds for 10 hours and % wound gap closures were

calculated. Bar graphs represent percent-based wound healing for Huh7 and Mahlavu cell lines.

Figure 6: Cell invasion assay. Average cell index values are normalized according to DMSO, which is represented by the horizontal dashed line for; **(a)** Huh7, and **(b)** Mahlavu cell lines, in the presence of LIMK inhibitors. The serum-free media containing 20 μM of each LIMKi compound were used and invasion progress of cells was monitored via xCelligence DP RTCA System (*: p-value < 0.05, ****: p-value < 0.0001).

3. Discussion

In this study, the main objective was to develop a computational method for predicting drug (or drug candidate compound) – target protein interactions with high confidence, for the purposes of improved drug discovery and repurposing. Here, we aimed to cover both physical and functional relationships between small molecule ligands and target proteins, to account for bio-interactions at higher levels, such as the inhibition of a cell with a drug/drug candidate compound. In DRUIDom, we assumed a data-driven approach and used experimentally validated interactions at large scale to build and optimize our model. For this, we utilized ChEMBL and PubChem databases and carefully filtered the bioactivity data points to construct our source dataset of compound – target protein interactions, which is one of the largest curated, high-quality experimental bioactivity datasets ever built, as far as we are aware (composed of 2,869,943 interaction data points between 3,644 target proteins and 1,033,581 compounds). This dataset is available in the data repository of the study and can be used by researchers working in the fields of drug discovery and repurposing, both as a training and benchmark dataset for the construction of new computational predictive models.

The idea behind DRUIDom's methodology is to identify the protein domains that are required for the interaction to occur (either physically or functionally), and propagating these associations to proteins that possess those domains. Thus, it was critical to successfully separate mappings that indicate a true relationship from the ones observed by chance. For this, we incorporated known/verified compound – target protein relations with undesired bioactivity levels (i.e., high xC_{50} values: $> 20\ \mu M$) as "inactives" even though they also are interactors, along with "actives" (compound – target protein pars with the desired levels of bioactivity: $xC_{50} < 10\ \mu M$), as two different datasets. This approach enabled us to score compound – domain mappings in terms of potential true-false positives and true-false negatives (as explained in the Methods section 4.2.1), and to identify pairs with a practical potential to ultimately become new treatment options.

One limitation of our data-centric methodological approach is penalizing a compound – domain mapping with a false negative count if one of the known active target proteins does not contain the mapped domain. It is known that a small molecule can be the ligand of different proteins and different domains, especially when the structural features of the corresponding binding sites are similar to each other. In cases like this, penalizing a mapping leads to the underestimation of its mapping score. In order to minimize this effect, we took the InterPro domain hierarchy into account while calculating the mapping scores. InterPro combines domains from the same functional family under distinct hierarchical trees. There are also significant similarities between the sequence profiles of domains from the same hierarchy. In DRUIDom, while scoring a mapping, we checked whether the known active and inactive target proteins of the intended compound possess domains from the same hierarchy. As such, we counted an active target protein containing a domain from the same hierarchy (but not the actual mapped domain) as a true positive (instead of false negative) and counted an inactive target protein containing a domain from the same hierarchy as a false positive (instead of true negative). In this way, domain similarity has been incorporated in DRUIDom. However, there are also cases where a single compound binds to domains from completely different hierarchies. Our approach does not currently take these cases into account.

During the parameter optimization and performance analyses of DRUIDom, it was important to make sure that there was no data leak from the benchmark test dataset to our training set. This condition has been automatically satisfied since the source of the mappings in the InteracDome benchmark dataset (i.e., PDB co-complex structures) and the source of the mappings in our training dataset (i.e., assay-based biological activity measurements obtained from ChEMBL and PubChem databases) are completely independent from each other.

In our analysis, we observed that only a small portion of the InterPro domain entries appear in the finalized compound – domain mappings, with the total number of 250 domains, as opposed to 8,165 compounds, at the selected mapping score threshold. The main reason behind this observation may lie in the data distribution in the source bioactivity dataset, as members from the same protein families have been targeted in most of the experimental bioassays (e.g., kinases, GPCRs). The distribution of the number of compounds mapped to each domain reveals that the top 10 domains constitute 56.7% of 27,032 mappings in total (i.e. “IPR000719 - Protein kinase domain”, “IPR001245 - Serine-threonine/tyrosine-protein kinase, catalytic domain”, “IPR017452 - GPCR, rhodopsin-like, 7TM”, “IPR020635 - Tyrosine-protein kinase, catalytic domain”, “IPR028174 - Fibroblast growth factor receptor 1, catalytic domain”, “IPR030611 - Aurora kinase A”, “IPR034670 - Checkpoint kinase 1, catalytic domain”, “IPR035588 - Janus kinase 2, pseudokinase domain”, “IPR035589 - Janus kinase 2, catalytic domain”, “IPR039192 - Glycogen synthase kinase 3, catalytic domain”). Overall, eight out of ten of these domains belong to kinases.

We examined the difference in target proteins between our source bioactivity dataset and the resulting predicted DTIs dataset, to observe if it was possible to produce predictions for under-studied proteins through the approach outlined in this study. The unique number of target proteins in our source bioactivity dataset is 3,644, whereas, this number is 5,563 for our finalized DTI prediction dataset, which indicates that there is a 52.7% increase in target proteins thanks to the domain-based association approach. We also checked the protein family distribution of the targets in the original and the predicted interaction datasets, considering 5 main classes of proteins as enzymes, membrane receptors, ion channels, transcription factors, and others (i.e., a combination of transporters, epigenetic regulators, secreted proteins, other cytosolic proteins, other nuclear proteins, and other categories), according to the first level (L1) of ChEMBL protein classification (https://www.ebi.ac.uk/chembl/g/#browse/targets). For this, we compared the target protein family distribution in the original bioactivity dataset (i.e., 64% enzymes, 11% membrane receptors, 5% ion channels, 4% transcription factors, and 16% others) with our DTI prediction dataset (i.e., 50% enzymes, 25% membrane receptors, 7% ion channels, 8% transcription factors, and 10% others). Although dominating families in the source bioactivity dataset prevail in the predicted DTIs dataset, we were able to produce interacting compound predictions for a critically higher number of proteins from membrane receptor, ion channel, and transcription factor families with a 248%, 114%, and 238% increase, respectively. These results, again, demonstrate the effectiveness of the domain-based approach in predicting new target proteins.

In this study, we aimed to validate our drug/compound – target protein interaction prediction method by targeting the PI3K/Akt/mTOR pathway by focusing on the predicted LIM kinase inhibitors. The importance of selecting LIMKs as targets come from their unique kinase domains which have longer activation loops compared to many kinases, allowing the design of specific inhibitors against cancer invasion and metastasis [31]. Furthermore, LIMK1 knockout was not embryonically lethal in mice making this protein a good candidate for drug design [37]. Another study showed that LIMK activity is beneficial for cancer cells in terms of coping with chemotherapeutics and ionizing radiation, which renders cells resistant to these treatments [38-41]. Therefore, LIMKs are promising candidates due to their essential role in cytoskeletal remodeling leading to cell migration and invasion. Hence, the lack of cytotoxicity of our predicted compounds on normal transformed HEK-238 cells is in parallel with the above-mentioned cellular LIMK activities, which is prominent in cancer cells.

For the validation study, we initially examined the binding properties of 4 originally predicted compounds (i.e., LIMKi-1, 1a, 2, and 3) by computational docking and comparing with the crystal structures of multi-kinase inhibitor staurosporine and previously identified LIMK ligand 9D8 in complex with LIMK1 and LIMK2 proteins, respectively. LIMKi-2, its derivatives, and LIMKi-3 had the most significant binding energies. During the *in vitro* validation stage of the study, we performed bioactivity experiments on liver cancer cells because intrahepatic metastatic migration/invasion is a major problem for patient survival and the specific selection of treatment is dependent on the number of distinct cancer nodules within the organ [42]. Our observations from the docking analysis were further supported by cytotoxicity and migration/invasion experiments where LIMKi-2 was the most significant compound regarding its action on cancer cells. Our promising results with LIMKi-2 directed us to synthesize 4 novel derivatives of this compound (i.e., LIMKi-2a, b, c, and d). Among these derivative compounds, LIMKi-2c and LIMKi-2d displayed highly significant anti-migratory and anti-invasive properties on liver cancer cells, together with strong docking binding affinities. The increased activity for LIMKi-2c and 2d is interesting and seems to point to a favorable change in conformation due to the bromide substituent that twists the benzene ring against the thiadiazol and causes loss of coplanarity. Finally, our evaluation singled out the novel LIMKi-2d compound as a promising candidate therapeutic agent due to its action on mesenchymal Mahlavu cells which are highly aggressive in terms of drug resistance for cytotoxicity, motility, and migration [43].

As future work, we plan to further develop our predictive approach by identifying associations between ligands and experimentally characterized protein structures (from Protein Data Bank) and high-quality structure models generated by cutting-edge structure prediction methods [44]. Additionally, we plan to develop a web-based tool that contains the entire pipeline, where researchers from various fields can both browse pre-computed associations/predictions, and generate interacting drug/compound predictions for their proteins of interest on the fly, using the provided interface. We also plan to extend the work on LIMK inhibition with additional *in vitro* experiments and *in vivo* studies, with the ultimate aim of contributing to the development of new cancer drugs.

The computational drug/compound – target protein interaction prediction approach proposed in this study led to the identification of novel interactions, a selected subset of which were then validated by both *in silico* and *in vitro* experiments. Results of the cell-based validation experiments indicate DRUIDom has the ability to generate generalized predictions that are well-translated into higher organizational levels such as the cell. Also based on these results, it is possible to state that the approach proposed here is producing biologically relevant results that can be utilized in drug discovery and repurposing studies beyond PI3K/Akt/mTOR pathway and cancer, especially for pathological conditions where specific domain-based targeting may be critical, such as metabolic disorders.

4. Methods

4.1 Dataset Construction

Bioactivity data points, each of which indicates the experimentally verified interaction between a compound and a target biomolecule (i.e., protein), were downloaded from open-access bioassay databases and divided into 2 classes as active (i.e., interacting) and inactive (i.e., non-interacting, or more precisely: “non-interacting at the desired level”) pairs. For the selection of active data points, we used a bioactivity value threshold of < 10 μM xC_{50} (i.e., IC_{50} or equivalent). For inactives, we used a bioactivity value threshold of > 20 μM xC_{50}. The data points between 10 and 20 μM were discarded, since their classification to either class was considered to be ambiguous.

ChEMBL bioactivity database [17] and PubChem bioassay database [16] were used as the bioactivity data source. The bioactivity data was acquired from the ChEMBL database (v23) via SQL queries with specified parameters (i.e., assay type: binding, target type: single protein, taxon: metazoa, standard value: < 10 μM for active/interacting pairs and > 20 μM for inactive/non-interacting pairs). We only selected the data points with a pChEMBL value, which corresponds to a calculated activity measure of half-maximal response concentration/potency/affinity (e.g., IC_{50}, EC_{50}, AC_{50}, XC_{50}, Ki, Kd, and potency) in the negative logarithmic scale. pChEMBL value of 5 is equal to an IC_{50} measurement of 10 μM. The presence of a pChEMBL value indicates that the data point has been checked by a curator. Following the elimination of duplicates, the final ChEMBL set contained 718,102 bioactivity data points (627,353 actives and 90,749 inactives) between 3,533 target proteins and 467,658 compounds.

Due to the structural organization of the PubChem bioassay database, it was not straightforward to obtain a bioactivity dataset with desired properties. However, the developers of ExCAPE-DB solved this problem by extensively filtering and organizing PubChem bioactivity data (together with ChEMBL bioactivity data) and presented the results in a database [45]. ChEMBL v20 and the PubChem bioassay database (January 2016) are incorporated in ExCAPE. In our study, we incorporated PubChem bioactivities directly using the ExCAPE-DB. We discarded the PubChem data points where the actual bioactivity values were missing. These points could have been included using the assay outcome field, where each data point is already marked as either "active" or "inactive"; however, the test concentrations for these data points are not available, and it is probable that many of them do not obey the thresholds we determined. Following the elimination of data points with activity values between 10 and 20 μM, the final ExCAPE bioactivity dataset contained 2,514,439 bioactivity values between 1,648 target proteins and 856,216 compounds. The reason behind the low number of target proteins compared to the ChEMBL dataset was that, in ExCAPE, only three organisms (i.e., human, mouse and rat) were included. Finally, ChEMBL v23 and ExCAPE datasets were merged to obtain the finalized bioactivity training dataset of the study. Since ExCAPE-DB incorporates ChEMBL data (from v20, which is an older version compared to the one we used) along with PubChem, many duplicates were added to our dataset following merging, which were eliminated by simply deleting repeat data points. Our finalized source bioactivity dataset contains 2,869,943 data points between 3,644 target proteins and 1,033,581 compounds. 1,637,599 of these data points are in the actives class, and the remaining 1,232,344 are in the inactives class. The contradictions between active and inactive classes (i.e., compound – protein pairs that are listed both as active and inactive) are low, with only 1,574 cases (< 0.06%).

UniProt Knowledgebase -UniProtKB- v2019_01 [25] and InterPro v72 database [20] were employed as the source for target protein sequences and their domain annotations, respectively. InterPro integrates sequence signatures with functional significance from 13 different manually curated and automated databases presenting functional and structural protein information. In InterPro, domain content, order and positions are pre-computed for each UniProtKB protein sequence using the InterProScan tool and the sequence profiles/HMMs and presented within a public dataset. We downloaded InterPro annotations for all of the target proteins in our dataset (i.e., 3,644) and eliminated the InterPro hits for non-domain type entries such as families and sites. A total of 3,118 target proteins had at least one InterPro domain hit, and thus, could be further used in our study. The average number of domains in these target proteins was 2.44. We also generated domain architectures, which can be defined as the linear arrangement of the domain hits on the protein sequence, for each multi-domain protein in our dataset. The domain architecture information is later used for mapping compounds to domain pairs, to account for the cases where multiple domains are required to be presented in the protein to have an interaction with the corresponding compound (the detailed procedure is described below).

Canonical SMILES notations were employed to represent the compounds. SMILES is a widely used system that defines the structures of chemical species as line notations [46].

SMILES representations of all compounds in our dataset were directly downloaded from ChEMBL and PubChem databases. Extended-Connectivity Fingerprints (ECFP4) [47] were generated for all compounds in our bioactivity dataset (i.e., 1,033,581), using SMILES as the input. Pairwise molecular similarities were measured between all compound pair combinations using the Tanimoto coefficient. Python RDKit module [48] and ChemFP library [49] were employed to generate the fingerprints and to calculate the pairwise molecular similarities.

4.2 DTI Prediction System

The proposed prediction system contains two modules: compound – domain mapping (section 4.2.1) and the propagation of associations to other proteins and compounds (section 4.2.2). In the mapping module, small molecule drugs/compounds are probabilistically associated to single domains (or domain pairs) on target proteins, using experimentally verified compound – target interaction data in bioactivity data resources. In the second module, for each compound – domain pair, all proteins that contain the mapped domain and all compounds that are significantly similar to the mapped compound (in terms of molecular similarity) are crossed with each other to produce new drug/compound – target protein predictions.

4.2.1 Compound – domain mapping

Figure 1a displays the overall methodology within a schematic representation. In this example, a compound (C_i) and its target protein (P_1) is reported to be interacting/bioactive (i.e., according to our definition of active; $xC_{50} < 10\ \mu M$) in ChEMBL and/or PubChem. In this toy example, it has been identified from the InterPro database that P_1 has one domain annotation (i.e., blue domain), on which the binding site/region of C_i (with the desired bioactivity) is assumed to reside. It may also be possible that there is a functional relationship between the blue domain and C_i. This makes other human proteins containing the blue domain (i.e., P_2, $P_{3,}$ and P_4) candidate targets for C_i and for other drug-like compounds that are significantly similar to C_i with Tanimoto similarity greater than or equal to 0.8 (i.e., C_x, $C_{y,}$ and C_z).

To quantize the association between a compound and a domain, we calculated mapping scores for each compound – domain combination, using verified active and inactive compound – target protein data points in our source ChEMBL + PubChem bioactivity dataset. For this, precision, recall, accuracy, F1-score, and Matthew's correlation coefficient (MCC) metrics are employed. MCC successfully measures the quality of binary classifications when there is a class imbalance [50], such as the case observed in our dataset. Here, binary classification is the decision for either the presence or absence of a bio-interaction between a compound and a domain. Definitions below are used to calculate mapping scores for an example compound (C_1) and a domain (D_x):

- True positives (TP) represent the number of proteins that contain domain D_x, where the reported bioactivity against compound C_1 is within the actives portion (i.e., $xC_{50} < 10\ \mu M$),

- False positives (FP) represent the number of proteins that contain domain D_x, where the reported bioactivity against compound C_1 is within the inactives portion (i.e., $xC_{50} > 20\ \mu M$),

- False negatives (FN) represent the number of proteins that do not contain domain D_x, where the reported bioactivity against compound C_1 is within the actives portion (i.e., $xC_{50} < 10\ \mu M$),

- True negatives (TN) represent the number of proteins that do not contain domain D_x, where the reported bioactivity against compound C_1 is within the inactives portion (i.e., $xC_{50} > 20\ \mu M$).

Mapping score metrics are calculated using the above-defined TP, FP, FN, and TN; and their formulations are provided in Methods section 4.3. For all the compound – domain mappings, high scores indicate reliable mappings and a high probability that the region of interaction lies on the mapped domain. In Figure 1b, the mapping procedure is shown for 2 toy examples. Also, in Figure 1b, the number of TP, FP, FN, and TN for toy examples are given, together with the respective mapping scores (i.e., metrics). The first example corresponds to a case where there are 2 experimentally verified interacting (i.e., active) target proteins for compound C_1. Both of these proteins contain the blue domain (i.e., a structural unit responsible for the interaction with C_1.). C_1 also has 3 inactive proteins (i.e., targets with insufficient bioactivity), 2 of which contain the red domain and 1 contains the light green domain. With the selection of the domain with the maximum score, the blue domain is mapped to C_1. Another example mapping case is presented for compound C_2, where most of the known targets are multi-domain proteins. For C_2, many of the targets contain the green domain, red domain, or both of them. Association scores for single domains and domain pairs revealed that the best score is achieved when green and red domains exist together. It is observed that the real-world cases can be much more complicated compared to the toy examples provided in Figure 1b, as one protein can be the target of multiple compounds and one compound can target multiple proteins. To be able to separate reliable mappings from the non-reliable ones we determined and applied mapping score thresholds using the metrics provided in section 4.3. The test applied to determine these thresholds is described (together with its results) in the Results section 2.1.

With the purpose of increasing the reliability of the data in our verified bioactivity dataset, we directly eliminated the mappings to the compounds if the number of active and inactive targets is less than 3 (each). This filter was applied to eliminate the compounds with only a few data points, which could otherwise produce false high mapping scores. This application dramatically reduced the number of compounds in our source dataset from 1,033,581 to 51,750. To be able to incorporate more data points, we generated a second dataset by combining the active and inactive targets of the compounds in clusters, which were significantly similar to each other in terms of molecular structure, and treated each cluster as an individual compound while calculating the mapping scores. To distribute the compounds in clusters we used pairwise molecular similarities via Tanimoto coefficient (over ECFP4 fingerprints) with a threshold of 0.7, which was above the previously applied threshold to predict targets based on compound molecular similarities [51]. All compounds that were similar to each other with at least 0.7 Tanimoto similarity were placed in the same cluster. Clusters with less than 5 active and 5 inactive targets were directly eliminated to ensure reliability in terms of the number of data points. In this way, 202,238 clusters were generated with compound overlaps in-between. This procedure should not be confused with compound similarity-based propagation of target protein associations, which is explained in section 4.2.2. The mapping score calculation was carried out for all of the 51,750 individual compounds in our first dataset (i.e., single-compound-based mappings) and for 202,238 clusters in our second dataset (i.e., compound-cluster-based mappings) against domains of their respective target proteins. For the compound-cluster-based analysis, the score obtained for each domain mapping was propagated to all compounds in the corresponding cluster. This resulted in a total of 3,487,239 raw compound – domain mappings for the cluster-based bioactivity dataset (i.e., compound-cluster-based mappings) and 449,294 raw mappings for the individual compound-based dataset (i.e., single-compound-based mappings). Figure 7 displays the histograms composed of bins of the total number of targets, the number of active targets, and the number of inactive targets (X-axis), for individual compounds (Figure 7a, b, c) and for compound clusters (Figure 7d, e, f). Y-axis represents the number of compounds or compound clusters in the log scale. As observed, there was a steady decrease in the number of compounds/clusters when the number of targets per compound/cluster was increased. There was also a clear difference between active and inactive target bins, indeed no individual compound or cluster with higher than 80 inactive targets were identified. The most probable reason for this was that, negative results (i.e., non-interactions) are not usually reported in the literature. The gain from using compound clusters was highlighted especially for active targets and for all targets (i.e., a vs. d and b vs. e) with the increase in the height of the bars for more than 50 targets (notice the scaling difference in the X-axis between the individual compound histograms and the compound cluster histograms).

Figure 7. Log-scale histograms of the number of individual compounds and compound clusters (Y-axis) with the given number of target proteins (X-axis) in our source bioactivity dataset; for individual compounds: **(a)** all targets, **(b)** active targets, **(c)** inactive targets; and for compound clusters: **(d)** all targets, **(e)** active targets, **(f)** inactive targets.

A similar procedure was applied to map compounds to domain pairs. For this, all domain pair combinations were identified for each target protein in our source dataset, using the domain architecture information of the proteins extracted using the UniProt-DAAC method, which was described in our previous study [52]. All domain pairs were recorded as if they were single domains and the mapping procedure explained above was applied to obtain compound – domain pair mappings. This procedure yielded a total of 1,075,550 raw individual compound – domain pair mappings and 9,343,130 raw compound cluster – domain pair mappings. The high number (compared to single domain mappings) was due to the elevated number of domain pair combinations, especially for large proteins.

Once the mapping score threshold had been selected (as explained in Results section 2.1), all mappings below the threshold were discarded, and the remaining mappings constituted the finalized mapping dataset.

4.2.2 Propagation of associations

The second module starts with the detection of pairwise similarities between all compounds in our source dataset using molecular fingerprints. For this, Extended-Connectivity Fingerprints (ECFP4) [47] were generated for all compounds in our bioactivity dataset (i.e., 1,033,581). The pairwise similarities were measured using the Tanimoto coefficient with a threshold of 0.8 to signify significant similarities, which was even above the previously applied Tanimoto thresholds to safely transfer target annotations between small molecule compounds [51]. Briefly, domain associations that were produced in the previous step were transferred to new compounds that are similar to the mapped compound with a Tanimoto similarity value greater than equal to 0.8. The idea behind this application was that the structurally similar molecules tend to have similar interactions, as assumed in conventional ligand-based virtual screening [47].

Subsequently, all human protein records in the UniProtKB/Swiss-Prot database were searched for the mapped domains and domain pairs, using the InterPro domain annotation information. When a new protein was found to contain the domain in question, it was associated with the corresponding compound. In this way, new candidate ligands were predicted for both known targets and for new candidate target proteins that possess the mapped domains or domain pairs (Figure 1a).

4.3 Mapping Score and Performance Analysis Metrics

Precision, recall, accuracy, F1-score, and Matthew's correlation coefficient (MCC) metrics are used for both the calculation of mappings scores (Methods section 4.2.1) and calculation of the overall system performance (Results section 2.1). The formulation of these metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + FN} \quad (3)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (5)$$

Definitions for TP (i.e., true positives), FN (i.e., false negatives), FP (i.e., false positives) and TN (i.e., true negatives) are given in Method section 4.2.1.

4.4 Molecular Docking Experiments

For the molecular docking of predicted inhibitor compounds and their derivatives against kinase domains of LIMK1 and LIMK2 proteins, the crystal structure of LIMK1 kinase domain as a complex with staurosporine (PDB id: 3S95) and the crystal structure of LIMK2 kinase domain complex with bound 9D8 (PDB id: 5NXD) were retrieved from RCSB PDB database [53]. Then, the PDB files of both protein structures were loaded into AutoDockTools-1.5.6. For both proteins, which are in the form of 2-chain homodimer structures, only the A chain was kept for docking and preprocessed by deleting all heteroatoms, adding hydrogen atoms, computing Gasteiger charges, and merging non-polar hydrogens. The preprocessed protein structures were saved as pdbqt files. For flexible docking, contact residues of LIMK1 and LIMK2 proteins were selected and saved as flexible pdbqt files, while the remaining structures of the proteins were saved as rigid pdbqt files.

Full 3D structures of compounds were downloaded from ZINC (v15) database [54] in sdf file format and converted to PDB files by Open Babel file format converter [55]. Since the derivative compounds (i.e., LIMKi-2a, LIMKi-2b, LIMKi-2c, LIMKi-2d) could not be found in the ZINC database, compound 3D structures (in the form of PDB files) were generated from the SMILES representations of respective compounds using ChemAxon JChem software-based online tool at: http://pasilla.health.unm.edu/tomcat/biocomp/convert. Then, Gasteiger charges were added, rotatable bonds and the root for the identification of a central atom were detected for compound PDB structures, and they were saved as pdbqt files in AutoDockTools.

Grid map files for both rigid and flexible dockings were generated by AutoGrid4 program (AutoDock-4.2.6) [56] using protein and compound pdbqt files as inputs, and the x-y-z coordinates for the grid search were defined by calculating the mean coordinates of the reported interacting atoms of LIMK1 and LIMK2 proteins, which were retrieved from PDBsum [57]. Grid box parameters for grid search were set as shown in Table 3a. In the docking step, a genetic algorithm with default settings was used for parameter searching, and the docking analysis of each compound – protein pair was carried out by using AutoDock4 (v4.2.6) [56].

As a second docking validation, the same analysis was also performed by using MTiAutoDock [58] and SwissDock [59] web services. Protein pdb files were given as an input to the MTiAutoDock service together with the sdf formatted ligand structure files. List of residues mode was selected for grid calculation and the contact residues of each protein was given as input. MTiAutoDock service has automatically added the hydrogen atoms to the crystal structure and executed the docking procedure using AutoDock 4.2.6. For SwissDock, blind docking was implemented using protein PDB files and ligand mol2 files as input. For all docking analyses, different poses were evaluated via binding free energy calculations and the one with the lowest energy was selected as the finalized result (i.e., the best pose). UCSF Chimera software was used for the visualization of docking results.

4.5 Chemical Synthesis of the Predicted Inhibitors

DRUIDom predicted 4 compounds as inhibitors of LIMK1 and LIMK2 proteins, which have been selected as targets of the validation use-case study. Structures, database identifiers, and given names (by us) of these compounds (i.e., LIMKi-1, LIMKi-1a, LIMKi-2, LIMKi-3) are displayed in Figure 2. We synthesized these molecules to be used in the cell-based assays. Also, the structure of LIMKi-2 has been modified with the aim of building 4 new derivatives with a potentially higher biological activity (i.e., shown in Figure 2 as LIMKi-2a, LIMKi-2b, LIMKi-2c, LIMKi-2d), making a total of 8 molecules. Procedures used in the chemical synthesis of these molecules are given in the Supplementary Material document.

4.6 *In vitro* Experimental Assays

All LIMKi (LIM-Kinase Inhibitor) compounds were dissolved in DMSO and stored at -20 ^{0}C as 20 mM stocks.

4.6.1 Cell Culture

Human hepatocellular carcinoma cell lines (Huh7, Mahlavu), colon carcinoma cell line (HCT116), breast cancer cell line (MCF-7) were maintained in Dulbecco's Modified Eagle Media (DMEM) (Gibco, Cat:31885-023): together with 10% FBS (Gibco, Cat:10270), 1% Non-essential Amino Acid (MEM-NEAA) (Gibco, Cat:11140-050) and 1% Penicillin-Streptomycin (Gibco, Cat:15140-122); whereas human embryonic kidney cell line (HEK-293) was maintained in same reagents described above together with 100 μg/ml Hygromycin B (Invitrogen, Cat: 10687-010) at 37^{0}C under 5% $CO_{4.}$ All cells used in this study are STR authenticated and regularly tested for contamination with the mycoplasma test kit (MycoAlert™, Lonza, Cat:LT07-118).

4.6.2 SRB (Sulforhodamine B) Assay

Cells were collected with trypsinization after washed with PBS once. Collected cells seeded in 96-well cell culture plate, adjusted with 150 ul/well as followed; Huh-7 (2500 cells/well), Mahlavu (1500 cells/well), HCT-116 (2000 cells/well), MCF-7 (2000 cells/well) and Hek-293 (3000 cells/well). LIMKi compounds were administered in the range of concentration from 40 μM to 2,5 μM, 24 hours later from the initial seeding step. After 72 hours of treatment, cells were fixed with 10% trichloroacetic acid (TCA;Sigma, Cat:27242) and proteins were stained with 0,4% sulforhodamine B sodium salt (SRB; Sigma, Cat: S1402) solution, dissolved in 1% acetic acid (Sigma, Cat: 27225) [60]. Plates were read on BMG SpectroStar Nano Spectrophotometer at 515nm. IC_{50} values were calculated based on the normalization according to DMSO-treated (Sigma, Cat: D2650) groups.

4.6.3 Western Blotting

500.000 cells of Huh7 and 250.000 cells of Mahlavu were seeded in 150 mm cell culture dishes (Sarstedt, Cat: 83.3903). After 24 hours, the old media was removed and fresh media containing 20 μM of each LIMK inhibitor were added. All treatments were performed as duplicates for 48 hours. At the end of the treatment, cells were scraped and protein extraction was performed. Protein Electrophoresis (Bio-Rad, Mini-PROTEAN® Tetra Cell Systems and TGX precast gels) and transfer system (Bio-Rad, Trans-Blot Turbo Transfer System) were used according to the manufacturer's protocol. Proteins were transferred to a PVDF-LF membrane (Bio-Rad, Cat:1620260) Following antibodies were used as described within western blotting protocol. phospho-Cofilin (CST, Cat: 3313) (1:200 v/v), Total Cofilin (CST, Cat:5175) (1:200 v/v), and IRDye® 800CW Goat-anti-Rabbit IgG Secondary Antibody (LI-COR, Cat:926-32211) (1:20000 v/v). For normalization, REVERT™ 700 Total Protein Stain Kit (LI-COR, Cat:926-11016) was used according to the manufacturer's protocol. Images were taken with LI-COR Odyssey Clx Imaging Device. Signal normalization was performed based on the REVERT™ Total Protein Stain Normalization protocol by LI-COR Biosciences and imaging analysis was performed by LI-COR, Image Studio Lite software. For efficiency testing for LIMKi compounds with IC_{100} dosages; anti-rabbit IgG (Sigma, Cat: A6154) was used as a secondary antibody (1:5000 v/v), and for imaging; SuperSignal West Femto (Thermo Scientific; Cat: 34095) was used. Imaging was acquired by using LI-COR C-DiGit ® Blot Scanner. Signal intensity analysis was performed by LI-COR, Image Studio Lite software.

4.6.4 Scratch Assay

Huh7 (150.000 cells) and Mahlavu (100.000 cells) cells were seeded to 35 mm cell culture dishes (Corning, Cat:430165) and incubated for at least 24h until cells attached and became confluent. The wound was created in confluent (nearly 100%) monolayer cells by using p30 pipet tip followed by washing with PBS (Gibco, Cat: 14190-169) three times before adding the serum-free medium (1% FBS) that includes LIMK inhibitors or vehicle DMSO. The migration rate of LIMK inhibitor-treated cells was analyzed by comparing samples with the migration of control cells treated with DMSO controls. Gap closure was analyzed by capturing images with time-lapse Nikon ECLIPSE Ti-S inverted microscopy for 10 min intervals for 10 hours (high-quality images of the treated cells are given in the data repository of the study). Upon 10 hours the distance of the same reference point measured at the first and last frame were compared by using NIS-Elements software.

4.6.5 Real-Time Cell invasion Analysis

Cells were seeded on CIM-Plate 16^{TM} (ACEA, Cat: 05 665 817 001), (20.000 cells/well for Mahlavu and 50.000 cells/well for Huh7 as triplicates) and monitored their invasion capacity on xCELLigence DP RTCA System, in the presence of 20 μM LIMKi compounds. The lower chamber of CIM-Plate was filled with 160 μl DMEM containing 10% FBS. Cells were resuspended with LIMKi compounds in serum-free DMEM (1% FBS, 1% NEAA, and 1% Penicillin / Streptomycin) and inoculated into the upper chamber in 150 ul as final volume. After the inoculation, CIM-Plate was incubated at room temperature for 30 min to allow the cells to settle; then the system was initiated to record CI data for 48 hours with 15-minute intervals. CI values were used to represent time-dependent invasion patterns of cells.

4.6.6 Statistical Analysis

All SRB and migration data in this study were obtained from three independent experiments with n ≥ 3 biological replicates. Western Blot experiments were performed as duplicates with 3 independent experiments. The statistical analysis for Western Blot was performed using Welch's *t-test* (Prism, GraphPad) and for the migration assay, Two-way ANOVA (Prism, GraphPad) was performed. Standard deviations of IC_{50} results from SRB Assay and from real-time cell proliferation data were calculated on Microsoft Excel. Statistically significant results were represented as follows: *: p-value <0.05; **: p-value <0.01; ***: p-value <0.001; and ****: p-value <0.0001.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery. 2010 Mar;9(3):203-14.
2. Hopkins AL. Predicting promiscuity. Nature. 2009 Nov;462(7270):167-8.
3. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nature Reviews Genetics. 2004 Apr;5(4):262-75.
4. Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Briefings in bioinformatics. 2019 Sep;20(5):1878-912.
5. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK. Recognizing pitfalls in virtual screening: a critical review. Journal of chemical information and modeling. 2012 Apr 23;52(4):867-81.
6. Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. Molecules. 2020 Jan;25(6):1375.
7. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004 Dec;432(7019):862-5.
8. Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics. 2008 Oct 1;24(19):2149-56.
9. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Current opinion in chemical biology. 2006 Jun 1;10(3):194-202.
10. Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Doğan T. DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chemical science. 2020;11(9):2531-57.
11. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nature biotechnology. 2007 Feb;25(2):197-206.
12. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, Li X, Zhou W, Wang W, Wang Y. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PloS one. 2012 May 30;7(5):e37608.

13. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic acids research. 2014 Jul 1;42(W1):W32-8.

14. Zhou W, Wang Y, Lu A, Zhang G. Systems pharmacology in small molecular drug discovery. International journal of molecular sciences. 2016 Feb;17(2):246.

15. Talele TT, Khedkar SA, Rigby AC. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Current topics in medicinal chemistry. 2010 Jan 1;10(1):127-41.

16. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L. PubChem 2019 update: improved access to chemical data. Nucleic acids research. 2019 Jan 8;47(D1):D1102-9.

17. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M. ChEMBL: towards direct deposition of bioassay data. Nucleic acids research. 2019 Jan 8;47(D1):D930-40.

18. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proceedings of the National Academy of Sciences. 1973 Mar 1;70(3):697-701.

19. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL. The Pfam protein families database in 2019. Nucleic acids research. 2019 Jan 8;47(D1):D427-32.

20. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic acids research. 2019 Jan 8;47(D1):D351-60.

21. Li Q, Cheng T, Wang Y, Bryant SH. Characterizing protein domain associations by Small-molecule ligand binding. Journal of proteome science and computational biology. 2012 Dec 3;1.

22. Kruger FA, Rostom R, Overington JP. Mapping small molecule binding data to structural domains. InBMC bioinformatics 2012 Dec (Vol. 13, No. 17, pp. 1-13). BioMed Central.

23. Kruger FA, Gaulton A, Nowotka M, Overington JP. PPDMs—a resource for mapping small molecule bioactivities from ChEMBL to Pfam-A protein domains. Bioinformatics. 2015 Mar 1;31(5):776-8.

24. Kobren SN, Singh M. Systematic domain-based aggregation of protein structures highlights DNA-, RNA-and other ligand-binding positions. Nucleic acids research. 2019 Jan 25;47(2):582-93.

25. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic acids research. 2019 Jan 8;47(D1):D506-15.

26. Ersahin T, Tuncbag N, Cetin-Atalay R. The PI3K/AKT/mTOR interactive pathway. Molecular BioSystems. 2015;11(7):1946-54.

27. Scott RW, Olson MF. LIM kinases: function, regulation and association with human disease. Journal of molecular medicine. 2007 Jun 1;85(6):555-68.

28. Mardilovich K, Baugh M, Crighton D, Kowalczyk D, Gabrielsen M, Munro J, Croft DR, Lourenco F, James D, Kalna G, McGarry L. LIM kinase inhibitors disrupt mitotic microtubule organization and impair tumor cell proliferation. Oncotarget. 2015 Nov 17;6(36):38469.

29. Scott RW, Hooper S, Crighton D, Li A, König I, Munro J, Trivier E, Wickman G, Morin P, Croft DR, Dawson J. LIM kinases are required for invasive path generation by tumor and tumor-associated stromal cells. Journal of Cell Biology. 2010 Oct 4;191(1):169-85.

30. Lagoutte E, Villeneuve C, Lafanechère L, Wells CM, Jones GE, Chavrier P, Rossé C. LIMK regulates tumor-cell invasion and matrix degradation through tyrosine phosphorylation of MT1-MMP. Scientific reports. 2016 Apr 27;6(1):1-2.

31. Yoshioka K, Foletta V, Bernard O, Itoh K. A role for LIM kinase in cancer invasion. Proceedings of the National Academy of Sciences. 2003 Jun 10;100(12):7247-52.

32. Bu Z, Callaway DJ. Proteins move! Protein dynamics and long-range allostery in cell signaling. Advances in protein chemistry and structural biology. 2011 Jan 1;83:163-221.

33. England KS, Tumber A, Krojer T, Scozzafava G, Ng SS, Daniel M, Szykowska A, Che K, von Delft F, Burgess-Brown NA, Kawamura A. Optimisation of a triazolopyridine based histone demethylase inhibitor yields a potent and selective KDM2A (FBXL11) inhibitor. MedChemComm. 2014;5(12):1879-86.

34. Tang ZY. Hepatocellular carcinoma-cause, treatment and metastasis. World journal of gastroenterology. 2001 Aug 15;7(4):445.

35. Buontempo F, Ersahin T, Missiroli S, Senturk S, Etro D, Ozturk M, Capitani S, Cetin-Atalay R, Neri ML. Inhibition of Akt signaling in hepatoma cells induces apoptotic cell death independent of Akt activation status. Investigational new drugs. 2011 Dec;29(6):1303-13.

36. Mizuno K. Signaling mechanisms and functional roles of cofilin phosphorylation and dephosphorylation. Cellular signalling. 2013 Feb 1;25(2):457-69.

37. Meng Y, Zhang Y, Tregoubov V, Janus C, Cruz L, Jackson M, Lu WY, MacDonald JF, Wang JY, Falls DL, Jia Z. Abnormal spine morphology and enhanced LTP in LIMK-1 knockout mice. Neuron. 2002 Jul 3;35(1):121-33.

38. Croft DR, Crighton D, Samuel MS, Lourenco FC, Munro J, Wood J, Bensaad K, Vousden KH, Sansom OJ, Ryan KM, Olson MF. p53-mediated transcriptional regulation and activation of the actin cytoskeleton regulatory RhoC to LIMK2 signaling pathway promotes cell survival. Cell research. 2011 Apr;21(4):666-82.

39. Dan S, Tsunoda T, Kitahara O, Yanagawa R, Zembutsu H, Katagiri T, Yamazaki K, Nakamura Y, Yamori T. An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. Cancer Research. 2002 Feb 15;62(4):1139-47.

40. Po'Uha ST, Shum MS, Goebel A, Bernard O, Kavallaris M. LIM-kinase 2, a regulator of actin dynamics, is involved in mitotic spindle integrity and sensitivity to microtubule-destabilizing drugs. Oncogene. 2010 Jan;29(4):597-607.

41. Gamell C, Schofield AV, Suryadinata R, Sarcevic B, Bernard O. LIMK2 mediates resistance to chemotherapeutic drugs in neuroblastoma cells through regulation of drug-induced cell cycle arrest. PLoS One. 2013 Aug 21;8(8):e72850.

42. Llovet JM, Fuster J, Bruix J. The Barcelona approach: diagnosis, staging, and treatment of hepatocellular carcinoma. Liver transplantation. 2004 Feb;10(S2):S115-20.

43. Kahraman DC, Hanquet G, Jeanmart L, Lanners S, Šramel P, Boháč A, Cetin-Atalay R. Quinoides and VEGFR2 TKIs influence the fate of hepatocellular carcinoma and its cancer stem cells. MedChemComm. 2017;8(1):81-7.

44. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences. 2020 Jan 21;117(3):1496-503.

45. Sun J, Jeliazkova N, Chupakhin V, Golib-Dzib JF, Engkvist O, Carlsson L, Wegner J, Ceulemans H, Georgiev I, Jeliazkov V, Kochev N. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. Journal of cheminformatics. 2017 Dec;9(1):1-9.

46. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences. 1988 Feb 1;28(1):31-6.

47. Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling. 2010 May 24;50(5):742-54.

48. Landrum G. RDKit: Open-source cheminformatics.

49. Dalke A. The chemfp project. Journal of Cheminformatics. 2019 Dec;11(1):1-21.

50. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061. 2020 Oct 11.

51. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry: miniperspective. Journal of medicinal chemistry. 2014 Apr 24;57(8):3186-204.

52. Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, Martin MJ. UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. Bioinformatics. 2016 Aug 1;32(15):2264-71.

53. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic acids research. 2019 Jan 8;47(D1):D464-74.

54. Sterling T, Irwin JJ. ZINC 15–ligand discovery for everyone. Journal of chemical information and modeling. 2015 Nov 23;55(11):2324-37.

55. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of cheminformatics. 2011 Dec;3(1):1-4.

56. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of computational chemistry. 2009 Dec;30(16):2785-91.

57. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein science. 2018 Jan;27(1):129-34.

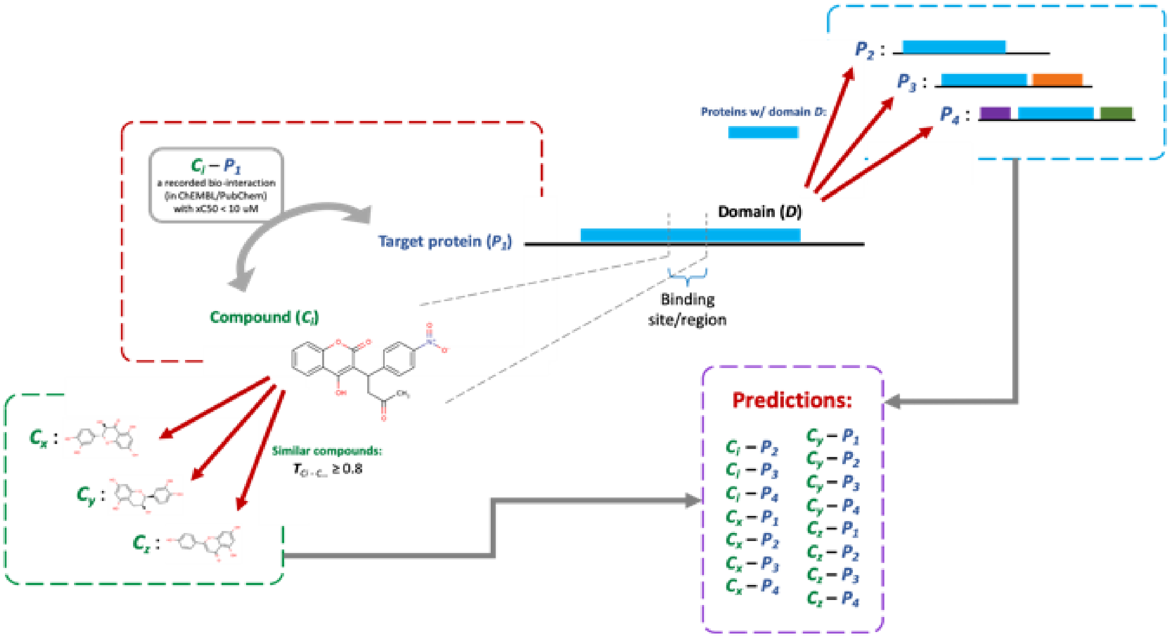
58. Labbé CM, Rey J, Lagorce D, Vavruša M, Becot J, Sperandio O, Villoutreix BO, Tufféry P, Miteva MA. MTiOpenScreen: a web server for structure-based virtual screening. Nucleic acids research. 2015 Jul 1;43(W1):W448-54.

59. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic acids research. 2011 May 28;39(suppl_2):W270-7.

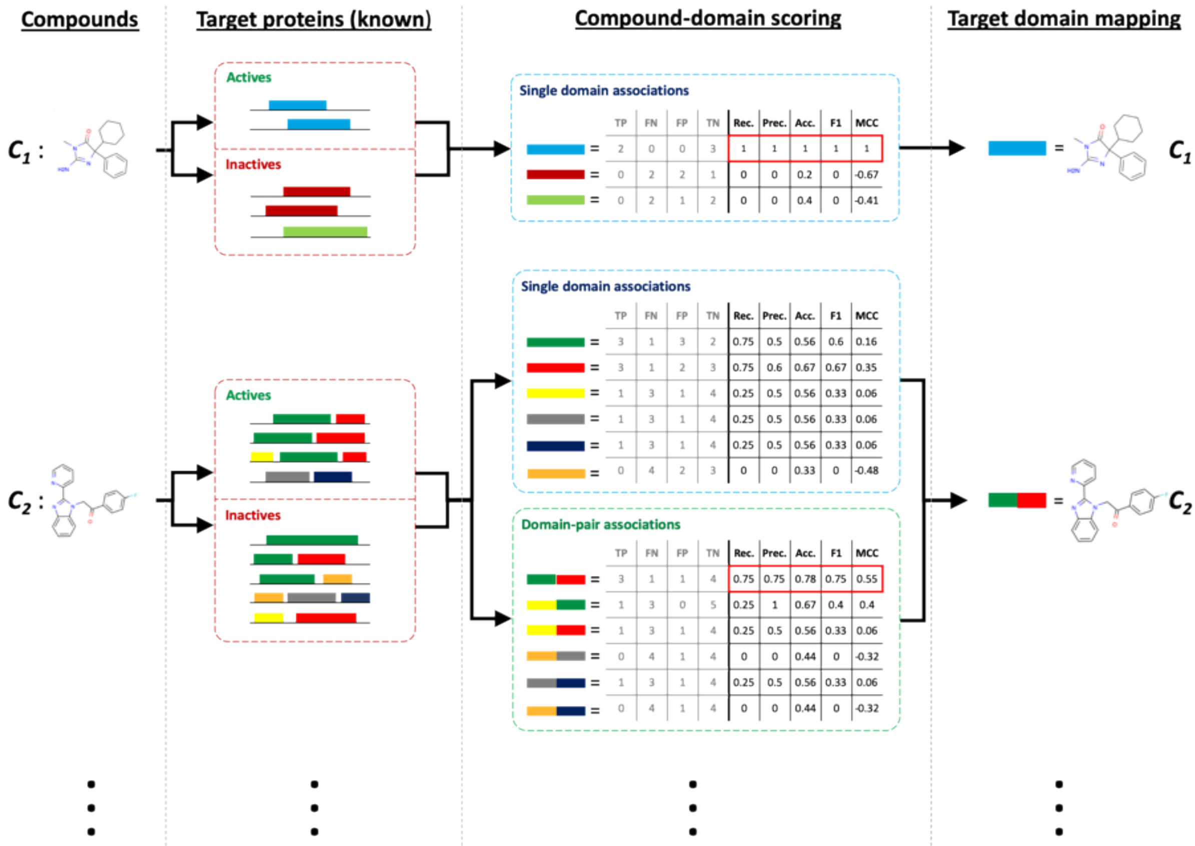
60. Kahraman DC, Kahraman T, Cetin-Atalay R. Targeting PI3K/Akt/mTOR pathway identifies differential expression and functional role of IL8 in liver cancer stem cell enrichment. Molecular cancer therapeutics. 2019 Nov 1;18(11):2146-57.

Figures

(a)



(b)



Single domain associations

	TP	FN	FP	TN	Rec.	Prec.	Acc.	F1	MCC
🟦 =	2	0	0	3	1	1	1	1	1
🟥 =	0	2	2	1	0	0	0.2	0	-0.67
🟩 =	0	2	1	2	0	0	0.4	0	-0.41

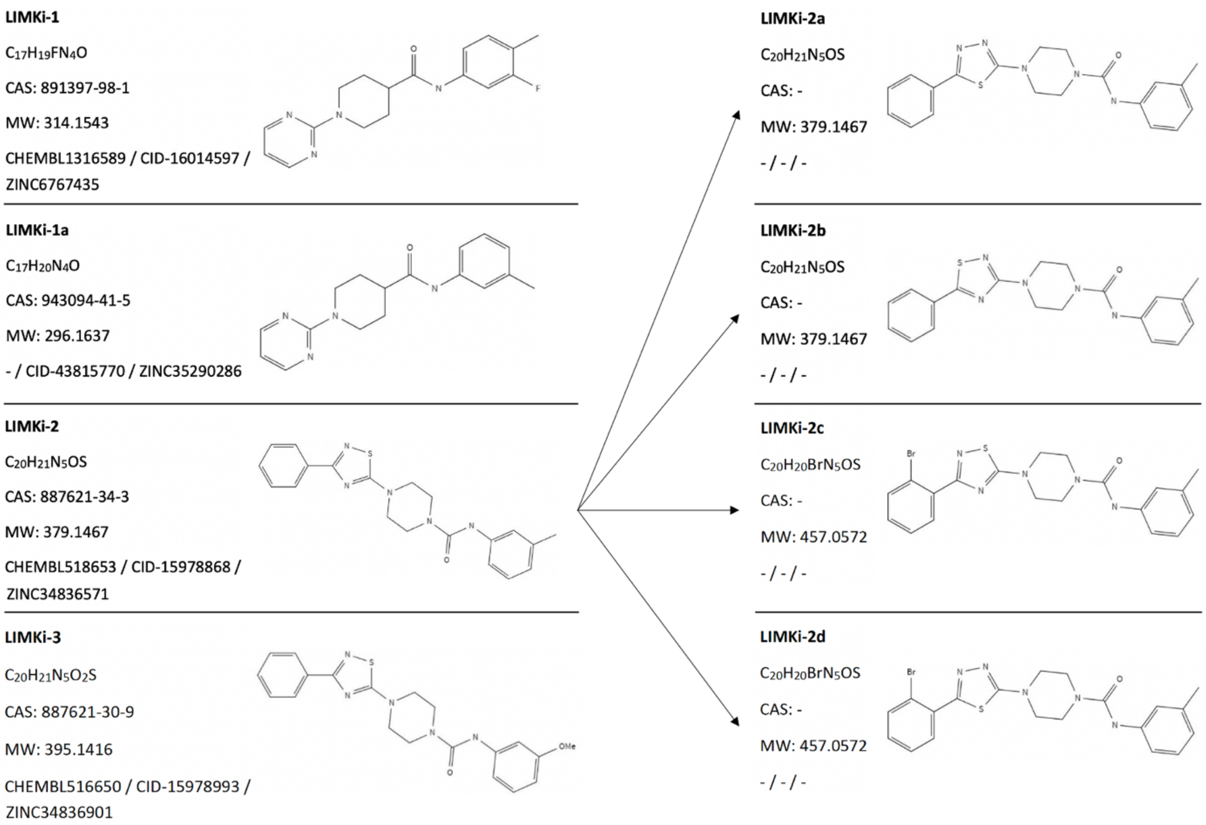
Single domain associations

	TP	FN	FP	TN	Rec.	Prec.	Acc.	F1	MCC
🟩 =	3	1	3	2	0.75	0.5	0.56	0.6	0.16
🟥 =	3	1	2	3	0.75	0.6	0.67	0.67	0.35
🟨 =	1	3	1	4	0.25	0.5	0.56	0.33	0.06
▨ =	1	3	1	4	0.25	0.5	0.56	0.33	0.06
⬛ =	1	3	1	4	0.25	0.5	0.56	0.33	0.06
🟧 =	0	4	2	3	0	0	0.33	0	-0.48

Domain-pair associations

	TP	FN	FP	TN	Rec.	Prec.	Acc.	F1	MCC
🟩🟥 =	3	1	1	4	0.75	0.75	0.78	0.75	0.55
🟨🟩 =	1	3	0	5	0.25	1	0.67	0.4	0.4
🟨🟥 =	1	3	1	4	0.25	0.5	0.56	0.33	0.06
🟧▨ =	0	4	1	4	0	0	0.44	0	-0.32
▨⬛ =	1	3	1	4	0.25	0.5	0.56	0.33	0.06
🟧⬛ =	0	4	1	4	0	0	0.44	0	-0.32

Figure 1. (a) The overall representation of the drug/compound – target protein interaction prediction approach used in DRUIDom (the diagram only depicts the relationship in terms of physical binding; however, DRUIDom also covers functional relationships between domains and compounds); **(b)** drug/compound – domain mapping procedure and its scoring over two representative (c_1, c_2) toy examples.



LIMKi-1

$C_{17}H_{19}FN_4O$

CAS: 891397-98-1

MW: 314.1543

CHEMBL1316589 / CID-16014597 / ZINC6767435

LIMKi-1a

$C_{17}H_{20}N_4O$

CAS: 943094-41-5

MW: 296.1637

- / CID-43815770 / ZINC35290286

LIMKi-2

$C_{20}H_{21}N_5OS$

CAS: 887621-34-3

MW: 379.1467

CHEMBL518653 / CID-15978868 / ZINC34836571

LIMKi-3

$C_{20}H_{21}N_5O_2S$

CAS: 887621-30-9

MW: 395.1416

CHEMBL516650 / CID-15978993 / ZINC34836901

LIMKi-2a

$C_{20}H_{21}N_5OS$

CAS: -

MW: 379.1467

- / - / -

LIMKi-2b

$C_{20}H_{21}N_5OS$

CAS: -

MW: 379.1467

- / - / -

LIMKi-2c

$C_{20}H_{20}BrN_5OS$

CAS: -

MW: 457.0572

- / - / -

LIMKi-2d

$C_{20}H_{20}BrN_5OS$

CAS: -

MW: 457.0572

- / - / -

Figure 2. Structures, database identifiers, and 2-D representations of predicted LIMK inhibitory compounds (LIMKi-1, 1a, 2, and 3) and derivatives (LIMKi-2a, b, c, and d).

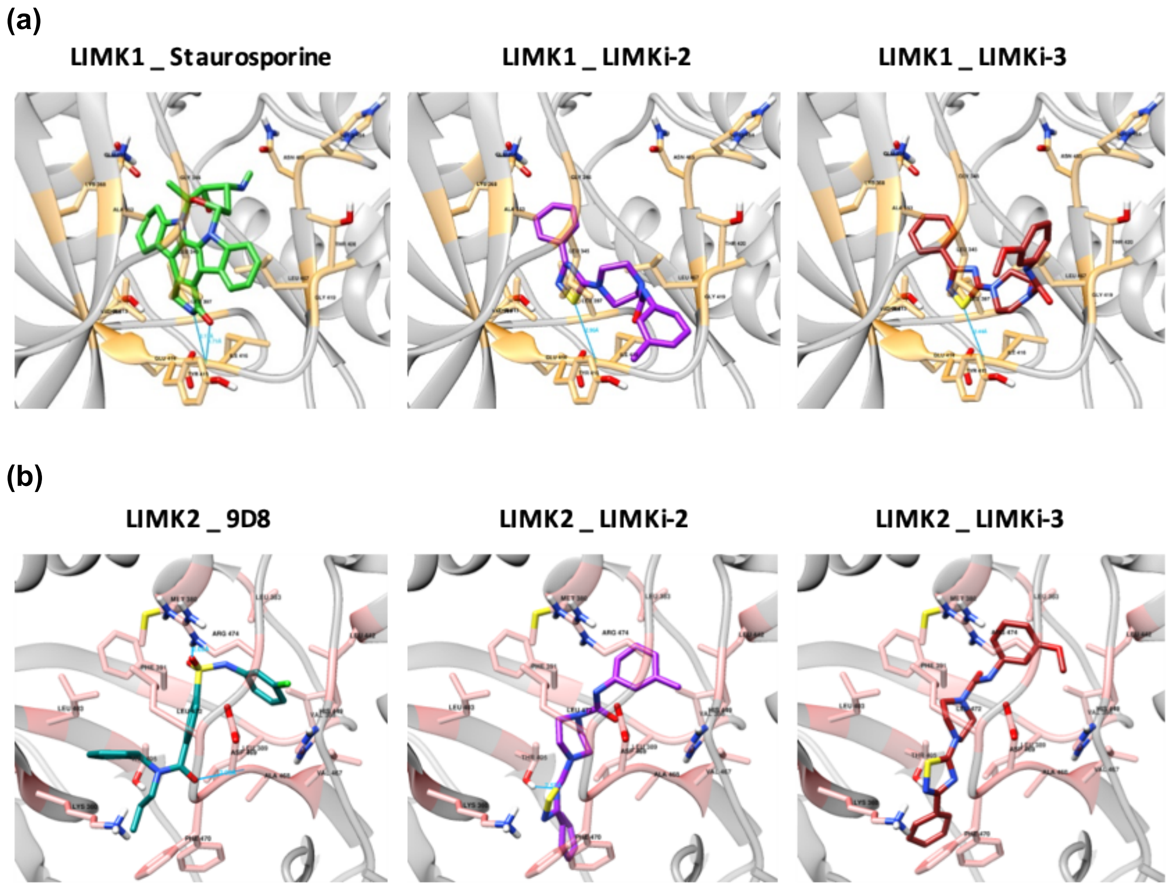


Figure 3. Visualization of the docked complex structures of **(a)** LIMK1 kinase domain in complex with the reference molecule staurosporine (green), LIMKi-2 (violet), and LIMKi-3 (red), and **(b)** LIMK2 kinase domain in complex with the reference molecule 9D8 (dark cyan), LIMKi-2 (violet), and LIMKi-3 (red) at the best poses. Hydrogen bonds are displayed with dark blue lines. Gold and pink colors represent LIMK1 and LIMK2 protein residues interacting with the corresponding compounds.

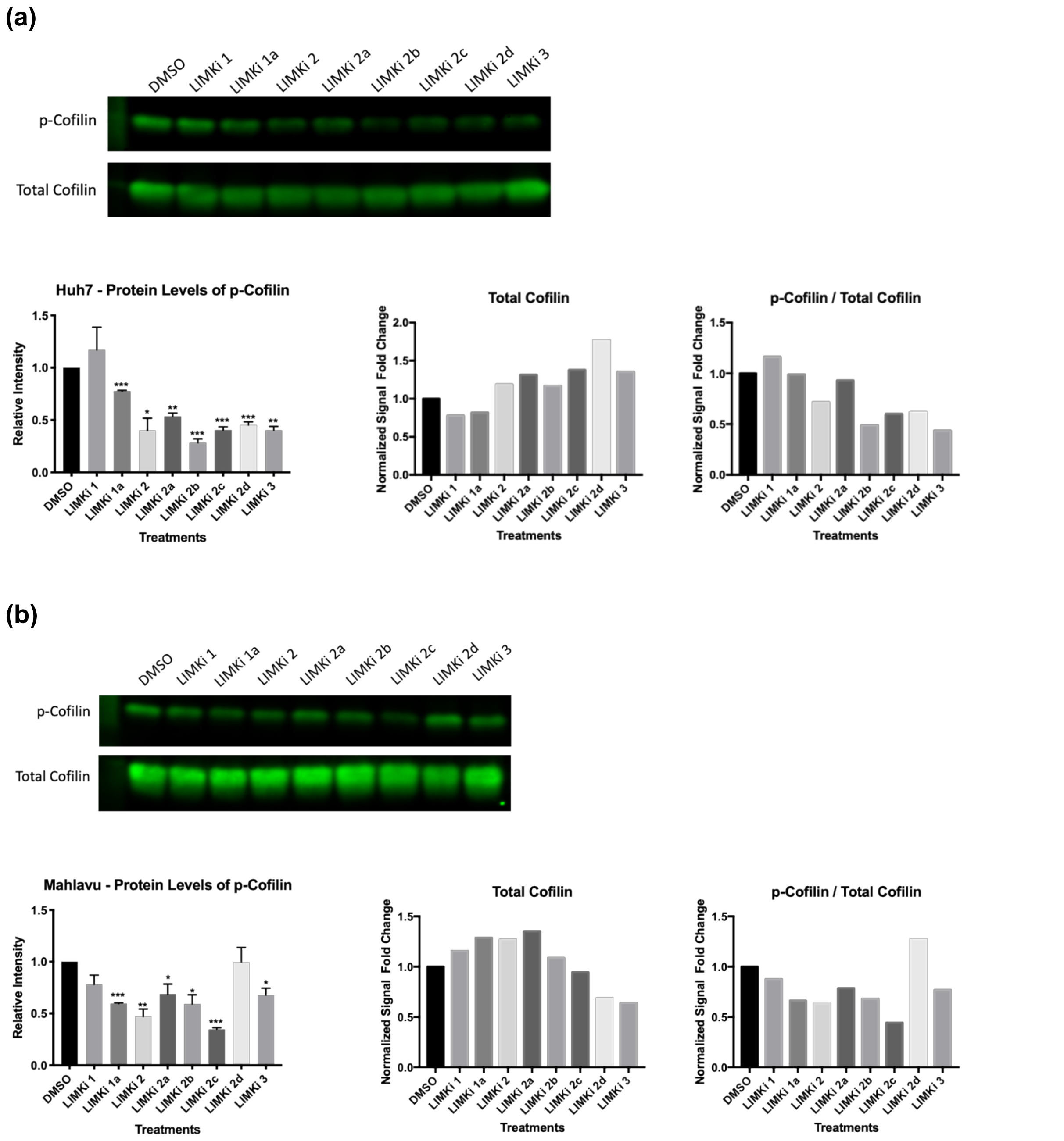


Figure 4. Phospho-Cofilin protein expression. **(a)** Huh7 and **(b)** Mahlavu cells were cultured with LIMK inhibitors (20 μM) for 48 hours and expression of active p-Cofilin and total Cofilin levels were assessed with western blot analysis. Bar graph indicates the relative intensity of p-Cofilin levels compared to untreated DMSO controls. The equal loading control was analyzed based on the total protein staining normalization protocol. The ratio of phospho- and total Cofilin levels for both Mahlavu and Huh7 cell lines were calculated.

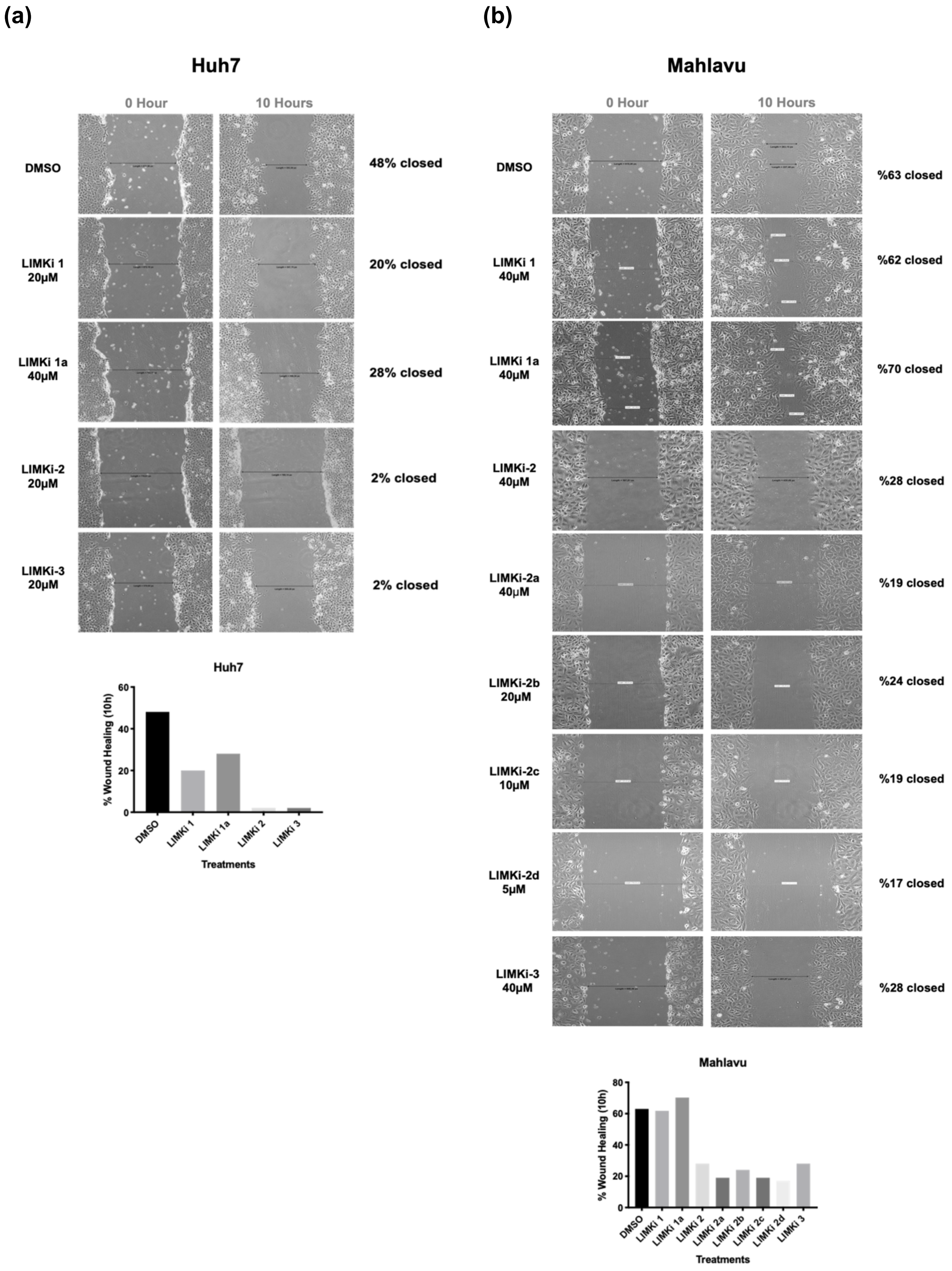


Figure 5: Wound healing assay. *In vitro* “wound” was created by a straight-line scratch across the monolayer **(a)** Huh7, **(b)** Mahlavu cells. Then cells were treated with indicated concentrations of LIMKi compounds for 10 hours and percent-based wound gap closures were calculated. Bar graphs represent percent-based wound healing for Huh7 and Mahlavu cell lines.

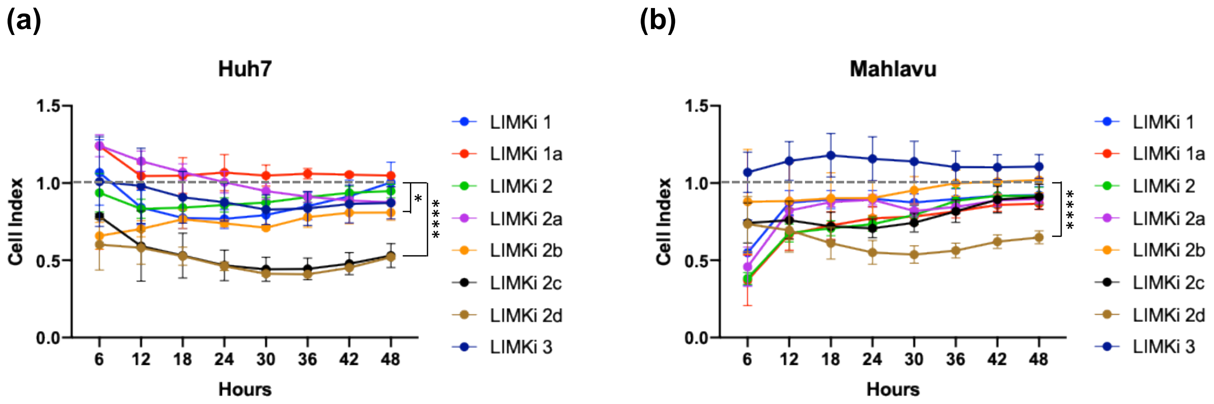


Figure 6: Cell invasion assay. Average cell index values are normalized according to DMSO, which is represented by the horizontal gray dashed line; **(a)** Huh7, and **(b)** Mahlavu cell lines, in the presence of LIMK inhibitors. The serum-free media containing 20 µM of each LIMKi compound were used and invasion progress of cells was monitored via xCelligence DP RTCA System (*: p-value < 0.05, ****: p-value < 0.0001, p-values were calculated in comparison to DMSO before the normalization).

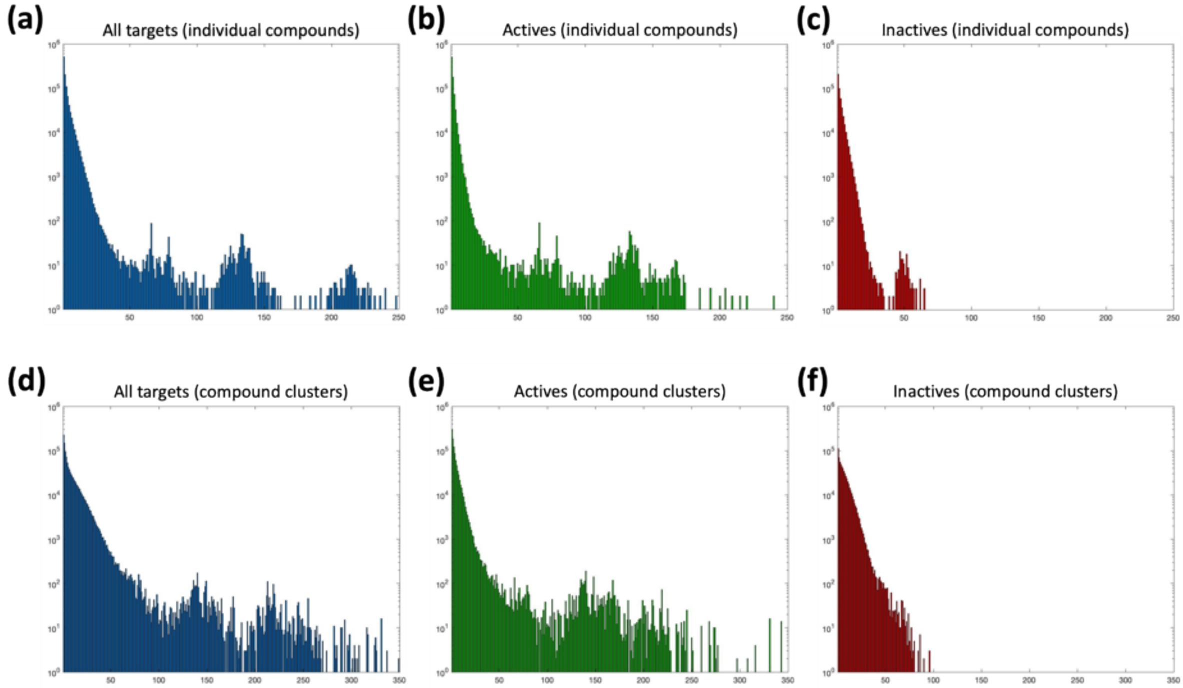


Figure 7. Log-scale histograms of the number of individual compounds and compound clusters (Y-axis) with the given number of target proteins (X-axis) in our source bioactivity dataset; for individual compounds: **(a)** all targets, **(b)** active targets, **(c)** inactive targets; for compound clusters: **(d)** all targets, **(e)** active targets, **(f)** inactive targets.

Supplementary Material

1. Chemical Synthesis of Inhibitor Molecules

1.1. Synthesis of pyrimidine-based structures **1** and **2** (LIMKi-1 and LIMKi-1a)

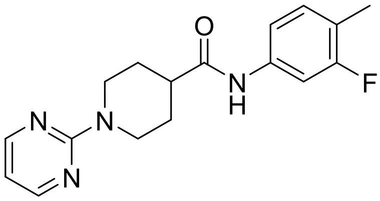
Procedure A:

To a solution of 2-chloropyrimidine (10 mmol) and ethyl isonipecotate (10 mmol) in MeCN (5 mL) was added solid potassium carbonate (11 mmol). The resulting reaction mixture was heated at 80 °C for 16 hours. After cooling to ambient temperature and evaporation of acetonitrile the residue was redissolved in ethyl acetate (25 mL) and extracted with water (3 x 10 mL). The organic extract was dried over anhydrous sodium sulfate, filtered and evaporated to dryness to yield the crude ester product as brown liquid (quantitative yield).

The ester intermediate was dissolved in a mixture of water and methanol (50 mL, 1:1 ratio by volume) and treated with solid sodium hydroxide (1.0 g). After heating this mixture at 60 °C for 3 hours, the reaction mixture was allowed to cool to room temperature. The mixture was extracted twice with dichloromethane (2 x 10 mL), the aqueous layer was acidified (1 M HCl) and extracted with dichloromethane (2 x 10 mL). The combined layers of this last extraction were dried over anhydrous sodium sulfate, filtered and evaporated to dryness yielding the corresponding carboxylic acid as colorless oil (92% yield – two steps).

A sample of the carboxylic acid (4 mmol) was dissolved in dry MeCN (1 M solution) and 1,1'-carbonyldiimidazole (5 mmol) was added. After heating for 2 hours at 50 °C the mixture was split into two equal volumes and treated separately with either 3-methylaniline (2.2 mmol) or 3-fluoro-4-methylaniline (2.2 mmol). Each sample was heated at 50 °C for a further 3 hours and the mixtures then allowed to cool to room temperature leading to precipitation of the desired products. Filtration of these solids followed by recrystallization from dichloromethane furnished the desired products (LIMKi-1 and LIMKi-1a) in high yield and purity as white solids.

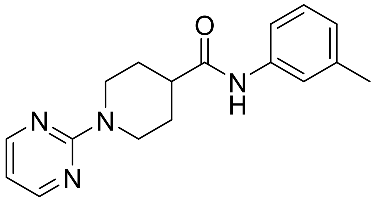
***N*-(3-Fluoro-4-methylphenyl)-1-(pyrimidin-2-yl)piperidine-4-carboxylate, 1 (LIMKi-1):**



White solid, 83% yield. ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.28 (d, *J* = 4.7 Hz, 2H), 7.66 (s, 1H), 7.43 – 7.35 (m, 1H), 7.10 – 7.00 (m, 2H), 6.46 (t, *J* = 4.7, 4.7 Hz, 1H), 4.85 – 4.75 (m, 2H), 2.89 (ddd, *J* = 13.4, 12.1, 2.8 Hz, 2H), 2.48 (tt, *J* = 11.6, 3.8 Hz, 1H), 2.19 (d, *J* = 2.0 Hz, 3H), 1.99 – 1.90 (m, 2H), 1.85 – 1.68 (m, 2H). ^{13}C NMR (101 MHz, Chloroform-*d*) δ 173.1 (C), 161.5 (C), 161.0 (CF, d, *J* = 245 Hz), 157.7 (2CH), 136.9 (C, d, *J* = 11 Hz), 131.3 (CH, d, *J* = 6 Hz), 120.6 (C, d, *J* = 18

Hz), 115.1 (CH, d, *J* = 3 Hz), 109.8 (CH), 107.4 (CH, d, *J* = 27 Hz), 44.6 (CH), 43.3 (2 x CH_2), 28.5 (2 x CH_2), 14.1 (CH_3, d, *J* = 3 Hz). ^{19}F NMR (376 MHz, Chloroform-*d*) δ -115.4. HRMS (TOF ES+) calculated for $C_{17}H_{20}N_4OF$ 315.1621, found 315.1625 (Δ = 1.3 ppm).

1-(Pyrimidin-2-yl)-*N*-(*m*-tolyl)piperidine-4-carboxamide, 2 (LIMKi-1a):



White solid, 79% yield. ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.28 (d, *J* = 4.7 Hz, 2H), 7.68 (s, 1H), 7.38 (s, 1H), 7.27 (d, *J* = 7.8 Hz, 1H), 7.15 (t, *J* = 7.8 Hz, 1H), 6.94 – 6.84 (m, 1H), 6.45 (t, *J* = 4.8 Hz, 1H), 4.80 (dt, *J* = 13.4, 2.7 Hz, 2H), 2.88 (ddd, *J* = 13.4, 12.1, 2.8 Hz, 2H), 2.48 (tt, *J* = 11.5, 3.8 Hz, 1H), 2.27 (s, 3H), 1.98 – 1.88 (m, 2H), 1.86 – 1.70 (m, 2H). ^{13}C NMR (101 MHz, Chloroform-*d*) δ 173.1 (C), 161.5 (C), 157.7 (2CH), 138.9 (C), 137.8 (C), 128.8 (CH), 125.2 (CH), 120.7 (CH), 117.1 (CH), 109.8 (CH), 44.6 (CH), 43.3 (2 x CH_2), 28.5 (2 x CH_2), 21.5 (CH_3). HRMS (TOF ES+) calculated for $C_{17}H_{21}N_4O$ 297.1715, found 297.1720 (Δ = 1.7 ppm).

1.2. Synthesis of thiadiazole-based structures **3** and **4** (LIMKi-2 and LIMKi-3)

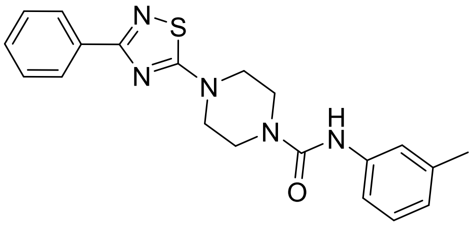
Procedure B:

To a suspension of the desired benzamidine hydrochloride hydrate (9 mmol) in dichloromethane (15 mL, 0 °C) was added trichloromethyl sulfenylchloride (10 mmol) and aqueous sodium hydroxide solution (9 mL, 6 N). After stirring this mixture for 1 hour at 0 °C the aqueous layer was separated and piperazine (20 mmol) was added to the organic layer. The resulting mixture was stirred at ambient temperature for 12 hours after which water (20 mL) was added. Extraction of the mixture was performed with dichloromethane (3 x 10 mL) and the combined organic layers were dried over anhydrous sodium sulfate, filtered and evaporated to yield the desired piperazine adduct as an off-white solid (75% yield).

Solutions of the above piperazine adduct were prepared in two separate vials (2 mmol each) in dichloromethane (3 mL each). To each vial was added the corresponding isocyanate (e.g., 3-methylphenylisocyanate or 3-methoxyphenylisocyanate; 2.2 mmol). After stirring this mixture for 5 hours at ambient temperature a white precipitate formed that was isolated by filtration. Recrystallisation from dichloromethane/hexane (1:1) furnished the desired adducts (LIMKi-2 and LIMKi-3) as white solids.

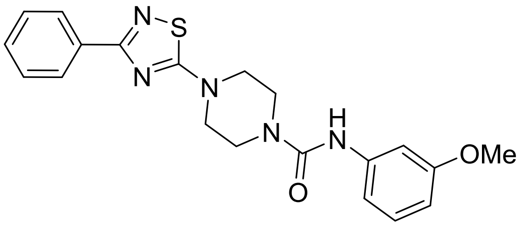
Further members of this small library (e.g. LIMKi-2a-d) were prepared in an analogous fashion and used after appropriate purifications.

4-(3-Phenyl-1,2,4-thiadiazol-5-yl)-*N*-(m-tolyl)piperazine-1-carboxamide, 3 (LIMKi-2):



White solid, 60% yield. ^{1}H NMR (700 MHz, DMSO-*d*$_6$) δ 8.63 (s, 1H), 8.13 – 8.07 (m, 2H), 7.45 (m, 3H), 7.28 (d, *J* = 2.0 Hz, 1H), 7.25 (dd, *J* = 8.1, 2.2 Hz, 1H), 7.10 (t, *J* = 7.8 Hz, 1H), 6.76 – 6.72 (m, 1H), 3.65 – 3.55 (m, 8H), 2.23 (s, 3H). ^{13}C NMR (176 MHz, DMSO-*d*$_6$) δ 185.2 (C), 169.5 (C), 155.3 (C), 140.7 (C), 137.8 (C), 133.3 (C), 130.5 (CH), 129.1 (2 x CH), 128.6 (CH), 128.0 (2 x CH), 123.1 (CH), 120.7 (CH), 117.3 (CH), 48.8 (2 x CH$_2$), 43.4 (2 x CH$_2$), 21.6 (CH$_3$). HRMS (TOF ES+) calculated for $C_{20}H_{22}N_5OS$ 380.1545, found 380.1532 (Δ = 3.4 ppm).

***N*-(3-Methoxyphenyl)-4-(3-phenyl)-1,2,4-thiadiazol-5-yl)piperazine-1-carboxamide, 4 (LIMKi-3):**



White solid, 66% yield. ^{1}H NMR (700 MHz, Chloroform-*d*) δ 8.20 – 8.15 (m, 2H), 7.46 – 7.38 (m, 3H), 7.18 (t, *J* = 8.1 Hz, 1H), 7.07 (t, *J* = 2.3 Hz, 1H), 6.87 (ddd, *J* = 8.0, 2.1, 0.9 Hz, 1H), 6.78 (d, *J* = 3.5 Hz, 1H), 6.61 (ddd, *J* = 8.3, 2.5, 0.9 Hz, 1H), 3.76 (s, 3H), 3.63 (dd, *J* = 7.2, 3.8 Hz, 4H), 3.62 – 3.58 (m, 4H). ^{13}C NMR (176 MHz, Chloroform-*d*) δ 185.1 (C), 170.4 (C), 160.2 (C), 154.9 (C), 139.9 (C), 133.2 (C), 130.0 (CH), 129.6 (CH), 128.5 (2 x CH), 128.0 (2 x CH), 112.5 (CH), 109.2 (CH), 106.3 (CH), 55.3 (CH$_3$), 48.3 (2 x CH$_2$), 43.3 (2 x CH$_2$). HRMS (TOF ES+) calculated for $C_{20}H_{22}N_5O_2S$ 396.1494, found 396.1490 (Δ = 1.0 ppm).