1 **Protein Domain-Based Prediction of Drug/Compound–Target Interactions and**

2 **Experimental Validation on LIM Kinases**

3 Tunca Doğan[1,2,3,*], Ece Akhan Güzelcan[3,4], Marcus Baumann[5], Altay Koyas[3], Heval Atas[3], Ian R.

4 Baxendale[6], Maria Martin[7] and Rengul Cetin-Atalay[3,8,*]

5 [1] Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

6 [2] Institute of Informatics, Hacettepe University, 06800 Ankara, Turkey

7 [3] CanSyL, Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey

8 [4] Center for Genomics and Rare Diseases & Biobank for Rare Diseases, Hacettepe University, 06230

9   Ankara, Turkey

10 [5] School of Chemistry, University College Dublin, D04 N2E2. Dublin, Ireland

11 [6] Department of Chemistry, University of Durham, DH1 3LE. Durham, UK

12 [7] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome

13   Trust Genome Campus, CB10 1SD. Hinxton, Cambridge, UK

14 [8] Section of Pulmonary and Critical Care Medicine, University of Chicago, Chicago IL, 60637, USA

15 * To whom correspondence should be addressed.

16   E-mail: tuncadogan@gmail.com & rengul@uchicago.edu

17 **Abstract**

18 Predictive approaches such as virtual screening have been used in drug discovery with the

19 objective of reducing developmental time and costs. Current machine learning and network-

20 based approaches have issues related to generalization, usability, or model interpretability,

21 especially due to the complexity of target proteins' structure/function, and bias in system

22 training datasets. Here, we propose a new method "DRUIDom" (DRUg Interacting Domain

23 prediction) to identify bio-interactions between drug candidate compounds and targets by

24 utilizing the domain modularity of proteins, to overcome problems associated with current

25 approaches. DRUIDom is composed of two methodological steps. First, ligands/compounds

26 are statistically mapped to structural domains of their target proteins, with the aim of

27 identifying their interactions. As such, other proteins containing the same mapped domain or

28   domain pair become new candidate targets for the corresponding compounds. Next, a

29   million-scale dataset of small molecule compounds, including those mapped to domains in

30   the previous step, are clustered based on their molecular similarities, and their domain

31   associations are propagated to other compounds within the same clusters. Experimentally

32   verified bioactivity data points, obtained from public databases, are meticulously filtered to

33   construct datasets of active/interacting and inactive/non-interacting drug/compound – target

34   pairs (~2.9M data points), and used as training data for calculating parameters of

35   compound–domain mappings, which led to 27,032 high-confidence associations between

36   250 domains and 8,165 compounds, and a finalized output of ~5 million new compound–

37   protein interactions. DRUIDom is experimentally validated by syntheses and bioactivity

38   analyses of compounds predicted to target LIM-kinase proteins, which play critical roles in

39   the regulation of cell motility, cell cycle progression, and differentiation through actin filament

40   dynamics. We showed that LIMK-inhibitor-2 and its derivatives significantly block the cancer

41   cell migration through inhibition of LIMK phosphorylation and the downstream protein cofilin.

42   One of the derivative compounds (LIMKi-2d) was identified as a promising candidate due to

43   its action on resistant Mahlavu liver cancer cells. The results demonstrated that DRUIDom

44   can be exploited to identify drug candidate compounds for intended targets and to predict

45   new target proteins based on the defined compound–domain relationships. Datasets,

46   results, and the source code of DRUIDom are fully-available at:

47   https://github.com/cansyl/DRUIDom.


48   **Author Summary**

49   Drug development comprises several interlinked steps from designing drug candidate

50   molecules to running clinical trials, with the aim to bring a new drug to market. A critical yet

51   costly and labor-intensive stage is drug discovery, in which drug candidate molecules that

52   specifically interact with the intended biomolecular target (mostly proteins) are identified.

53   Lately, data-centric computational methods have been proposed to aid experimental

54   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. Herein, we have developed a new method called DRUIDom (DRUg Interacting Domain prediction) to identify 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 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 takes on average 13 years [1]. One of the major cost influences 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,

81    considering the combinations between millions of small molecule drug candidate compounds

82    and thousands of potential protein targets, the combinatorial number of experiments is

83    extremely high, which is not possible to experimentally evaluate.

84    Over the last two decades, computational approaches have been developed with the

85    objective of aiding experimental studies in drug discovery, defining a new field entitled

86    "virtual screening" or "drug/compound – target protein interaction (DTI) prediction" [4-6].

87    Here, the aim is to predict unknown compound – target interactions with the construction

88    and application of statistical models, using various types of molecular descriptors [7]. There

89    are two distinct approaches to virtual screening. In the ligand-based approach, new chemical

90    substances are predicted as binders of the intended target biomolecules. This is usually

91    performed by calculating molecular similarities between the drug/compound that is known to

92    interact with the intended protein and other chemical substances in the library, thus,

93    returning the most similar ones as predictions via "guilt by association" [8]. Since the

94    predicted ligands of a target are usually limited to the compounds that are highly similar to its

95    known ligands, discovering new scaffolds is difficult through this approach. In structure-

96    based virtual screening methods, 3-D structural information of known ligand – receptor

97    complexes are used to model the interactions and predict new DTIs with similar interactive

98    properties [9]. Structure-based virtual screening is a costly process due to both highly

99    intensive computational processes and challenges associated with obtaining 3-D structures

100    of both protein and receptor-ligand complexes [2]. As a result, they are mostly limited to the

101    well-characterized portion of the target protein space. New computational approaches have

102    emerged to address these issues by adopting machine learning and/or network analysis

103    techniques [10-14]. There are cases where the drug candidate compounds, first discovered

104    by virtual screening, or via computer-aided drug discovery in general, became approved

105    drugs [4,15].

106    DTI prediction methods usually require large training datasets (i.e., experimentally verified

107    interaction information between compounds and proteins) to build accurate models.

108     Bioactivity databases such as PubChem [16] and ChEMBL [17] curate and publish *in vitro*

109     and *in vivo* bioassays, in the form of compound – target bioactivity measurements, which are

110     used by DTI predictors as training data. The open-access data presented in these resources

111     are extremely valuable for the research community; however, it is still difficult to find data

112     concerning less-studied targets, which prevents building predictive models for these less

113     common targets. Besides, the information in these databases is typically incomplete,

114     meaning that there are many unknown/undiscovered interactions for the compounds and the

115     targets presented in these resources, an aspect that is especially critical for estimating the

116     off-target effects of the drug candidate compounds. Nevertheless, computational predictions

117     concerning both under-studied targets and never-before-targeted proteins is an important

118     topic that may help researchers to assess the druggability of these proteins and develop

119     new therapeutic approaches.

120     Modelling interactions between compounds and proteins is a difficult task especially due to

121     the fact that molecular interactions between proteins and compounds are complex, also,

122     many proteins expressed by the human genome are yet to be structurally characterized. In

123     this sense, it is critical to reduce the complexity to a level where the modelling is feasible, the

124     required data is available at large scale, and the results produced are biologically relevant.

125     Proteins have modular structures made up of functional building blocks called domains.

126     Domains can fold, function, and evolve independently from the rest of the protein [18].

127     Protein regions that correspond to domains are evolutionarily highly conserved since

128     mutations in these functionally critical regions may lead to adverse consequences for the

129     organism. Once identified on the structures of characterized proteins, domains can be

130     detected (i.e., predicted) on structurally uncharacterized proteins by constructing domain

131     sequence profiles and by searching for these profiles on the amino acid sequences of

132     uncharacterized proteins [19,20]. Thanks to this application, domain/family annotation

133     coverage is considerably high on the documented protein sequence space in the UniProt

134     Knowledgebase (UniProtKB), i.e., 96.7% for UniProtKB/Swiss-Prot and 81.3% for

135    UniProtKB/TrEMBL. A few literature studies have investigated the relationship between

136    domains and small molecules within the perspective of drug discovery and repositioning. For

137    instance, Li *et al.* characterized the experimentally known binding interactions between

138    domains and small molecules using data from Protein Data Bank (PDB). Consequently, they

139    constructed a drug-domain network and used this to interpret modules of similar ligands and

140    domains [21]. Kruger *et al.* proposed a simple heuristic to map Pfam domains to small

141    molecules using ChEMBL bioactivity data as the source. The authors investigated the

142    structural relevance of the idea of mapping domains to Pfam profiles with statistical tests and

143    concluded that their heuristic produced accurate results [22,23]. Similar approaches have

144    been used to predict gene – phenotype associations [24] and domain – domain interactions

145    [25,26]. In a recent study, Kobren and Singh identified interactions between Pfam

146    family/domain entries and various types of ligands using PDB co-complex structures. Their

147    system InteracDome, employs the positional correspondence between Pfam HMMs and

148    amino acid sequences of the protein chains in PDB structures, together with known ligand-

149    binding regions on the same protein chains, to predict the interacting receptor-ligand pairs

150    [27]. Despite generating highly accurate mappings, InteracDome's coverage is limited

151    considering small molecule ligands due to its reliance on PDB co-complex structures. These

152    studies laid the foundation for the idea of associating small molecule binding to protein

153    domains but they have neither proposed a complete end-to-end prediction pipeline, nor

154    leveraged the advantage of using large-scale experimental bioactivity data accumulated in

155    public databases such as PubChem and ChEMBL. Consequently, there is a clear

156    requirement for new computational DTI prediction methods/tools, capable of producing

157    reliable and consistent results by using all available data in data resources to aid

158    experimental procedures in the field of drug discovery and repositioning.

159    In this study, we propose a new computational method called DRUIDom (DRUg Interacting

160    Domain prediction) for the comprehensive prediction of interactions between drugs/drug-like

161    compounds and target proteins to aid experimental and computational research in drug

162    discovery and repositioning. DRUIDom is based on associating compounds (i.e., small

163    molecule ligands) with complementary protein domains. The assumption behind the

164    mapping between domains and compounds is that, the binding region of the ligand is on the

165    mapped structural domain(s). Consequently, it is highly probable that other proteins

166    containing the mapped domain (or combination of domains) will possess the required

167    structural properties to interact with the compound of interest. DRUIDom employs a

168    supervised modelling approach, where the manually curated DTI information in ChEMBL

169    and PubChem databases are used in combination with the protein sequence and annotation

170    information in the UniProtKB [28] and the InterPro databases [20], for the construction of the

171    predictive model. The resulting predictions cover compound and human target protein

172    spaces recorded in the above-listed databases. In DRUIDom, we also evaluated compound

173    to domain pair mappings, in order to account for the cases where multiple domains are

174    required for the indented ligand interaction.

175    Our principal focus here was developing a complete chemogenomics-based drug/compound

176    – target protein interaction prediction system with a global perspective without focusing on

177    specific target families. For this, we constructed a large source bioactivity dataset and

178    applied a scoring-based heuristic to generate the compound – domain associations, which

179    are then propagated to other drug-like compounds and potential target proteins in the

180    massive chemogenomics space to produce DTI predictions at large scale. We believe this

181    study will provide valuable information for estimating both novel on-target and off-target

182    effects of drugs and drug candidate compounds.

183    With the aim of validating DRUIDom, we selected the PI3K/AKT/mTOR signalling pathway

184    for our experimental use-case study. PI3K/AKT/mTOR pathway is altered during the

185    progression of various cancer types [29]. Therefore, it is therapeutically relevant to target

186    this pathway. As such we analyzed interacting compound predictions of DRUIDom for

187    PI3K/AKT/mTOR pathway proteins, out of which, we focused on LIMK1 and LIMK2 proteins

188    and their new ligand predictions, as these proteins have been associated with several types

189    of cancer especially in terms of progression and metastasis [30-33]. To be used in the

190    experimental study, we synthesized the original 4 compounds predicted to inhibit LIMKs,

191    together with their novel derivatives. Bioactivities of the predicted small molecule

192    compounds were analyzed on transformed normal cells and cancer cell lines. The results of

193    these experimental assays, which are described in the following sections, validated the

194    computational predictions and indicate potential novel inhibitors for LIMK1 and LIMK2

195    proteins that can be further investigated for their anti-migratory effects.

196

## 197    2. Results

198    Our source/training dataset is composed of 2,869,943 drug/compound – target protein pair

199    data points (1,637,599 actives and 1,232,344 inactives) between 1,033,581 compounds and

200    3,644 target proteins. Using drug/compound – target associations contained in this dataset,

201    we first mapped compounds to domains, then, we produced DTI predictions by propagating

202    mappings to new compounds and new proteins (Figure 1). Detailed information about the

203    methodological procedure is given under 4.2.1 of the Methods section. Below, we first

204    explain the conducted data analysis together with its results (2.1), serving both as a guide to

205    determine the mapping parameters/thresholds and as a predictive performance evaluation of

206    DRUIDom. This is followed by the detailed analysis of compound – domain pair mappings in

207    comparison to single domain mappings (2.2), large-scale production of new drug/compound

208    – target protein interaction predictions (2.3), a validation use-case study on the

209    hepatocellular carcinoma (HCC) disease (2.4) with the selection of relevant targets (i.e.,

210    LIMK kinases) via literature review (2.4.1), molecular docking of selected novel inhibitor

211    predictions against LIMK proteins as an *in silico* validation (2.4.2), and the wet-lab *in vitro*

212    analysis of LIMK inhibition with the treatment of predicted inhibitors via chemical syntheses

213    and cell-based assays (2.4.3).

**Figure 1. (a)** The overall representation of the drug/compound – target protein interaction

prediction approach used in DRUIDom; and **(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 the InteracDome (the non-

redundant representable list - v0.3) mappings [27] 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.

The main reason behind using InteracDome as the reference dataset for the performance

analysis of DRUIDom was that their samples are reliable, as all of them are cases of

physical binding obtained from PDB.

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, a collection of 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 a total of

1,522 compounds in the non-redundant representable InteracDome dataset, a group of

1,144 were included in our mappings, and thus, constitute the intersecting compounds set.

The main reason for the numerical difference 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 steadily 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

268    structures and measured assay-based bioactivities, respectively). Since the performance is

269    calculated considering the intersecting compounds and domains at each score threshold,

270    the performance gradually increases with the increasing threshold, in terms of all metrics.

271    Both the ligand and domain coverage, at the score threshold (0.9) that yielded the highest

272    performance, was around 1% of the InteracDome.

273    Here, InteracDome represents an incomplete reference dataset, as a result, DRUIDom's

274    compound – domain mappings, which are not presented in the InteracDome dataset, are not

275    necessarily false positives. In cases like this, it is important to focus on performance in terms

276    of detecting known/true (positive) associations. Inspired from a few domain – domain

277    interaction prediction studies [25, 26], we calculated the enrichment of reference compound

278    – domain associations in highest-ranking DRUIDom mappings (accompanied with its

279    statistical significance value) with respect to random mappings on the exact same domain

280    and compound datasets. For this, we ranked DRUIDom mappings according to pre-

281    calculated MCC-based mapping scores. At each arbitrarily selected mapping score

282    threshold, we calculated the number of InteracDome associations found in DRUIDom

283    mappings, in comparison to the expected number of InteracDome associations to be

284    retrieved by randomly selecting the same number of pairs. For calculating the enrichment

285    scores and their statistical significance, we used the hypergeometric test, as described in

286    section "4.3 Mapping Score and Performance Analysis Metrics", equations 6 and 7.

287    According to results, the highest enrichment scores (i.e., > 6) are achieved when the

288    mapping score threshold is selected between 0.1 and 0.6, all of which are found to be

289    statistically significant with p-values $< 10^{-16}$ (Table 1). It is also indicated from enrichment

290    results that, at very low and high mapping thresholds enrichment is low, thus, selecting

291    these points may not be ideal.

292    Considering the trade-off between coverage and performance, we selected the confidence

293    threshold of 0.5, which provided an acceptable performance (i.e., accuracy: 0.95 and MCC:

294    0.78) and an InteracDome coverage of compounds: ~5% and domains: ~6%. At this

295     mapping score threshold, our approach produced 27,032 mappings between 250 domains

296     and 8,165 compounds/ligands.

297     We also calculated coverage extension values at each mapping score threshold to indicate

298     the numbers of new domains and compounds that have been included in DRUIDom

299     mappings (which are not presented in the InteracDome mappings at all). Coverage

300     extensions are given in terms of rates (percentages) calculated with respect to the total

301     number of domains (i.e., 1,043) and compounds (i.e., 1,144) intersecting between DRUIDom

302     and InteracDome mappings (e.g., in the case of score threshold: 0.5, DRUIDom has

303     mappings for 250 unique domains, 50 of which is shared with InteracDome, and it has been

304     previously calculated that the total number of intersecting domains between DRUIDom and

305     InteracDome is 1,043, as a result, the domain coverage extension at this point is calculated

306     as (250-50)/1043 = 0.192, in other words: 19.2%). Considering coverage extension values,

307     our mappings enriched the InteracDome dataset by ~19% for domains and ~707% for

308     compounds (at the selected score threshold of 0.5), which indicates the added value of our

309     approach, especially in terms of mapping new ligands. In this study, all of the steps followed

310     after this point were carried out using the mapping set generated with the score threshold of

311     0.5. However, in order to allow users to select other threshold values, we have also shared a

312     file in our data repository (https://github.com/cansyl/DRUIDom) that includes raw/non-filtered

313     compound – domain mappings together with their mapping scores.

314     **Table 1.** Compound – domain mapping performance analysis results.

| Mapping score threshold | # of retrieved: | | | Domain coverage* | Compound coverage* | Domain coverage extension* | Compound coverage extension* | Performance analysis results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Classification | | | | | | | | | Enrichment | |
| | Mappings | Domains | Compounds | | | | | TP | FP | FN | TN | Recall | Precision | Accuracy | F1-Score | MCC | Score | p-value |
| 0 | 3,245,943 | 1,018 | 215,432 | 31.0 | 16.5 | 66.6 | 18,814.9 | 163 | 3,235 | 116 | 9,414 | 0.58 | 0.05 | 0.74 | 0.09 | 0.11 | 2.195 | 6.05E-14 |
| 0.1 | 1,872,420 | 894 | 193,538 | 23.8 | 15.9 | 61.9 | 16,901.7 | 120 | 453 | 68 | 5,362 | 0.64 | 0.21 | 0.91 | 0.32 | 0.33 | 6.533 | 5.11E-43 |
| 0.2 | 548,679 | 759 | 95,934 | 15.7 | 13.2 | 57.0 | 8,372.6 | 96 | 170 | 36 | 2,328 | 0.73 | 0.36 | 0.92 | 0.48 | 0.48 | 6.979 | 2.18E-34 |
| 0.3 | 143,332 | 590 | 36,887 | 10.5 | 9.9 | 46.1 | 3,214.5 | 87 | 82 | 10 | 1,127 | 0.9 | 0.51 | 0.93 | 0.65 | 0.65 | 6.758 | 7.32E-28 |
| 0.4 | 36,112 | 299 | 13,408 | 6.5 | 7.8 | 22.1 | 1,164.2 | 80 | 54 | 4 | 787 | 0.95 | 0.6 | 0.94 | 0.73 | 0.73 | 6.451 | 1.02E-23 |
| *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 | 6.443 | 5.65E-21 |
| 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 | 6.111 | 1.28E-17 |

12

| 0.7 | 17,207 | 115 | 2,476 | 2.2 | 3.2 | 8.8 | 213.2 | 55 | 9 | 0 | 215 | 1 | 0.86 | 0.97 | 0.92 | 0.91 | 4.359 | 7.50E-10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.8 | 6,846 | 93 | 1,155 | 1.3 | 1.8 | 7.6 | 99.1 | 36 | 3 | 0 | 81 | 1 | 0.92 | 0.98 | 0.96 | 0.94 | 3.077 | 0.00024 |
| 0.9 | 2,783 | 70 | 372 | 1.2 | 1.0 | 5.6 | 31.5 | 21 | 1 | 0 | 38 | 1 | 0.95 | 0.98 | 0.98 | 0.96 | 2.727 | 0.01617 |
| 1 | 174 | 54 | 119 | 0.8 | 0.0 | 4.4 | 10.4 | 0 | 0 | 0 | 0 | - | - | - | - | - | - | 1 |

315 \* Given as % of InteracDome.

316 \*\*The selected mapping score threshold (shown in bold font).

317

318 **2.2 Domain pair to compound mappings**

319 Here, our aim was to observe if it would be possible to identify the cases where the

320 presence of a single domain is not sufficient for the occurrence of the interaction with the

321 intended compound, instead, an interface composed of multiple domains are required. Other

322 possible explanations for the requirement of multiple domains would be the allosteric

323 binding/regulation phenomenon [34]. To analyze this process, we generated compound –

324 domain pair mappings using the procedure explained at the end of Methods section 4.2.1.

325 For this procedure, we used the "bag of domains" approach where the order of the domains

326 on the protein sequence was not taken into account and all possible pair combinations were

327 then generated and tested. The reason for this evaluation is that domains that are quite far

328 away from each other on the linear protein sequence can be located very close to each

329 other upon folding of the protein.

330 Following the procedure described in the Methods section 4.2.1 and the thresholding/filtering

331 of mappings with the selected parameter values described in the Results section 2.1, 3,721

332 mappings were obtained between 1,456 compounds and 270 domain pairs. Next, these

333 pairs were compared with single domain pairings of the same compounds, in terms of the

334 mapping performance scores (e.g., $C_1 – D_xD_y$ is compared to $C_1$-$D_x$ and $C_1$-$D_y$ where C1

335 represents a compound and $D_xD_y$ represents a domain pair composed of the domains: $D_x$

336 and $D_y$), to observe if there is any performance improvement by mapping a pair instead of a

337 single domain (which is expected to provide more specific/defined interaction properties). In

338 most of the cases, the performance of the domain pair mapping was the same as the

339 mapping of the same compound to one of the single domains presented in the

340    corresponding domain pair, which indicates that only a single domain is sufficient for the

341    binding, and the other domain in the domain pair is just an extra (i.e., the second domain

342    does not play a detectable role in the binding). We called these domain pair mappings

343    "neutral domain pair associations". However, there were a few cases that domain pair

344    mapping actually increased the association performance, namely "positive domain pair

345    associations". To prepare the finalized compound – domain pair mapping set, all of the

346    neutral associations were discarded, yielding only 22 positive associations between 10

347    compounds and 12 domain pairs. Below, we investigated one example from positive domain

348    pair associations as a case study. The experimental bioactivity results of the case study

349    were obtained from the ChEMBL database (document link:

350    https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3621091), which was

351    previously curated from the study by England *et al.* where the authors investigated potent

352    inhibitors for KDM protein subfamilies [35].

353    The compound with the ChEMBL id "CHEMBL3621867" (link:

354    https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL3621867) was mapped to a

355    single InterPro domain record named: "JmjN domain" (id: IPR003349, description: domains

356    frequently found in the jumonji family of transcription factors, link:

357    https://www.ebi.ac.uk/interpro/entry/IPR003349) with the confusion matrix values TP:3,

358    FN:0, FP:1 and TN:2 (recall:1.00, precision:0.75, accuracy:0.83, F1-core:0.86, and

359    MCC:0.71), the false positive hit indicates that there is one protein that contains IPR003349

360    (gene: KDM4E, protein: "Lysine-specific demethylase 4E" in human, UniProt protein

361    accession: B2RXH2, link: https://www.uniprot.org/uniprot/B2RXH2), which was recorded to

362    be inactive against CHEMBL3621867 in ChEMBL database with a bioactivity value of $IC_{50}$ =

363    79.4 $\mu$M (and thus reported as a false positive in our analysis since the above mentioned

364    single domain mapping predicted B2RXH2 as a target of CHEMBL3621867). Similarly, the

365    same compound (CHEMBL3621867) was mapped to another single InterPro domain record

366    named: "Zinc finger, PHD-type" (id: IPR001965, description: a C4HC3 zinc-finger-like motif

367  found in nuclear proteins thought to be involved in chromatin-mediated transcriptional

368  regulation, link: https://www.ebi.ac.uk/interpro/entry/IPR001965) with values TP:3, FN:0,

369  FP:1 and TN:2 (recall:1.00, precision:0.75, accuracy:0.83, F1-core:0.86 and MCC:0.71),

370  indicating that, again, there is one protein that contains IPR001965 (gene: KDM2A, protein:

371  "Lysine-specific demethylase 2A" in human, UniProt protein accession: Q9Y2K7, link:

372  https://www.uniprot.org/uniprot/Q9Y2K7), which was recorded to be inactive against

373  CHEMBL3621867 in ChEMBL database with a bioactivity value of $IC_{50}$ = 50.1 $\mu$M (and thus

374  reported as a false positive in our analysis since the above mentioned single domain

375  mapping would predict Q9Y2K7 as a target of CHEMBL3621867). However, the mapping

376  between CHEMBL3621867 and the domain pair IPR003349-IPR001965 yielded an excellent

377  mapping performance with metrics TP:3, FN:0, FP:0 and TN:3 (recall:1.00, precision:1.00,

378  accuracy: 1.00, F1-core: 1.00 and MCC: 1.00), by eliminating the false positive target

379  predictions of B2RXH2 and Q9Y2K7 for CHEMBL3621867. The domain pair IPR003349-

380  IPR001965 is presented in 3 reviewed human protein entries among 6 proteins with

381  measured activities against CHEMBL3621867 (i.e., Lysine-specific demethylases 4C, 5C

382  and 4A, genes: KDM4C, KDM5C, and KDM4A, UniProt protein accessions: Q9H3R0,

383  P41229, and O75164), all of which were targets of the corresponding compound verified in

384  their respective binding assays with bioactivities of $IC_{50}$ = 7.9, 6.3 and 5.0 $\mu$M, respectively.

385  The protein that was accurately predicted as inactive by both single domain and domain pair

386  mappings (i.e., as a true negative) was "Lysine-specific demethylase 6B" (gene: KDM6B,

387  UniProt protein accession: O15054), which neither possessed IPR003349 nor IPR001965.

388  This target also received a bioactivity measurement of $IC_{50}$ = 63.1 $\mu$M against

389  CHEMBL3621867. The IPR003349 domain is annotated to 10 reviewed human protein

390  entries in the UniProtKB/Swiss-Prot database, also, IPR001965 is annotated to 88 reviewed

391  human protein entries. Whereas together, IPR003349 and IPR001965 are annotated to 7

392  reviewed human protein entries. Due to sequence differences between KDM subfamily

393  proteins (i.e., only 6 identical positions and 39 similar positions out of more than 1500

394   positions in the multiple sequence alignment of 6 KDM subfamily proteins), their domain

395   annotations are different from each other, which is possibly reflected in their 3-D structure

396   (although it is not possible to be sure without a crystal structure), and thus, the interaction

397   with the corresponding compound (i.e., CHEMBL3621867).

398   It is important to note that, proteins annotated with only one of the domains listed above (i.e.,

399   IPR003349 or IPR001965) are also targeted by CHEMBL3621867; however, corresponding

400   IC50s are way beyond plausible bioactivity values accepted for potential drug candidates

401   (i.e., < 10 $\mu$M). On the other hand, the presence of both domains on the target protein

402   yielded IC50 values that are within the acceptable range. This predicted domain pair –

403   compound mapping does not directly state a true physical binding between the mapped

404   domain pair and the compound, it rather suggests a relationship between the two entities

405   where the interaction is stronger in the cases with the presence of both domains. Thus,

406   targeting KDM subfamily proteins containing both IPR003349 and IPR001965 with

407   CHEMBL3621867 would have a higher chance of success in a drug discovery study.

408   It is probable for Q9Y2K7 (KDM2A) protein to partially possess the IPR003349 domain at

409   the N-terminal side. If this is the case, the InterProScan tool might not report the hit due to

410   obtaining a low score under the default statistical cut-off value. To analyze the case, we

411   locally aligned (using Smith-Waterman with default parameters of gap open:10, gap

412   extend:0.5, and scoring matrix:BLOSUM62) the first 100 N-terminal residues of Q9Y2K7

413   (KDM2A) and O75164 (KDM4A), which is reported to possess IPR003349 between the

414   positions 13 and 56 according to InterPro

415   (https://www.ebi.ac.uk/interpro/protein/UniProt/O75164/). The output alignment reported a

416   statistically significant hit (with 53.6% similarity between two sequences along the alignment

417   length of 28 residues) between KDM4A sequence positions 11 and 38, which roughly spans

418   the half of the IPR003349 domain, indicating the partial existence of the domain on Q9Y2K7

419   (KDM2A). Nevertheless, the partial existence of the domain may be the reason behind

420   observing interaction with a rather high bioactivity value (i.e., $IC_{50}$ = 50.1 $\mu$M).  It is not

16

421 possible for us to further comment on the physical binding as there is no co-crystal structure

422 of a KDM subfamily protein with CHEMBL3621867.

423 Besides single domains and domain pairs, it is also possible for some of the drug/compound

424 – target interactions to require three or even more domains to be presented at the target

425 protein. We could not account for these cases in DRUIDom since they dramatically increase

426 the complexity of the analysis, as a result, we chose to omit the cases requiring more than 2

427 domains.

428 **2.3 Predicting New Drug/Compound – Target Protein Interactions**

429 Drug/compound – target protein interaction predictions were generated by propagating the

430 drug/compound – single domain (or domain pair) mappings to proteins and other

431 compounds, using the procedure explained in Methods section 4.2.2. The crossing of new

432 compounds and targets for each mapping led to a geometric increase in the number of

433 associations/predictions. Finally, a simple post-processing filter was applied to predictions to

434 remove the known/recorded drug/compound – target protein interactions from the prediction

435 set.

436 First, 3,672,076 novel interactions (between 8,158 compounds and 5,563 proteins) were

437 generated with the propagation of single domains to proteins (i.e., 250 domains to 5,563

438 proteins). Also, 631 novel interactions (between 9 compounds and 286 proteins) were

439 produced with the propagation of domain pairs to proteins (i.e., 12 domain pairs to 286

440 proteins). The low number of predictions with domain pairs was due to the elimination of the

441 domain pair mappings that did not display a performance increase over the single domain

442 mappings of the same compound. At this point, the merged prediction dataset contained

443 3,672,220 novel interactions between 8,163 compounds and 5,563 proteins, after the

444 removal of duplicates. The finalized prediction dataset was obtained following the

445 propagation of the compounds in the previous prediction set to significantly similar

446 compounds according to molecular similarity-based compound clusters, which yielded

17

447    5,050,841 novel interactions between 10,944 compounds and 5,461 proteins in the finalized

448    prediction dataset, following the removal of known interactions. One observation here is that

449    there was only a slight increase in the number of compounds (from 8,163 to 10,944) after

450    the pairwise molecular similarity-based propagation, which can be explained by the strict

451    Tanimoto threshold of 0.8, which only passes the most reliable predictions.

452    With the aim of making this long list of predictions more accessible, we grouped them based

453    on signaling and metabolic pathways, in which the corresponding target proteins take roles.

454    We believe this pathway-based classification will be useful to researchers, especially for the

455    cases where specific biological processes are to be targeted (rather than a predefined single

456    protein). Out of 5,050,841 interaction predictions, 3,686,558 of them contain target proteins

457    that are annotated to one or more than one of 212 KEGG [36] signaling and metabolic

458    pathways. We extracted pathway-based prediction statistics, including the number of

459    interactions, and the number of unique compounds and proteins, for each pathway (Table

460    S1), which showed that olfactory transduction, neuroactive ligand-receptor interaction, and

461    calcium signaling pathways have the highest number of predicted ligand interactions.

462    These predictions can be potentially used in future drug discovery/repurposing studies. Both

463    the whole and pathway-based grouped compound – target interaction prediction lists are

464    made available in the GitHub repository of the study (https://github.com/cansyl/DRUIDom).

465    As explained in the dataset construction section (4.1), we discarded bioactivity data points

466    between 10 and 20 $\mu$M from our training dataset as these are neither considered to be active

467    nor inactive with any certainty. Thus, with the aim of observing how many of the compound –

468    target pairs that fall into this bioactivity range (in the current version -v29- of ChEMBL) are

469    predicted to be active/interacting by DRUIDom, we searched for these pairs among our

470    finalized compound – target interaction predictions. Out of the 178,089 unique compound –

471    target pairs with reported bioactivities between 10 and 20 $\mu$M, 263 of them are predicted to

472    be active/interacting by our method (i.e., 0.15% of them), which are given in Table S2. This

18

473  low number can be attributed to the fact that most of the compounds in bioactivity databases

474  are presented in only one (or just a few) compound – target pairs, and since pairs that fall

475  into the 10-to-20 $\mu$M range have been omitted from the training dataset of DRUIDom in the

476  first place, their compounds are missing from predictions (if they are not re-included due to

477  ligand similarity-based extension of mappings). This is supported by the following statistics;

478  among these 178,089 pairs, there are 124,189 unique compounds, and only 1483 of them

479  are presented in all DRUIDom predictions. It is also important to state that this is a favorable

480  finding since it is not desirable to produce active/interacting predictions to pairs with

481  ambiguous relationships in reality. In the dataset of compound – target pairs with 10-to-20

482  $\mu$M bioactivities, pairs that are predicted to be active by DRUIDom have a slightly lower

483  median activity value (14.1 $\mu$M), i.e., more bioactive with lower xC50 values, compared to

484  pairs that are not predicted to be active (mean activity: 14.7 $\mu$M).

485  **2.4 Validation of Predicted Molecular Interactions**

486  <u>2.4.1 Selection of Target Proteins</u>

487  For *in vitro* and *in silico* experimental validation, we focused on the hepatocellular carcinoma

488  (HCC) (i.e., a sub-type of liver cancer), which is the fourth most deadly cancer in the world

489  [37], and on the PI3K/AKT/mTOR signalling pathway, due to its critical role in various types

490  of cancer and cancer cell stemness [29]. To select inhibitory compound predictions, we first

491  checked our large-scale drug/compound – target interaction prediction dataset and found

492  116 inhibitor predictions (Table S3) for 4 PI3K/AKT/mTOR signalling pathway genes/proteins

493  (i.e., VEGFA, MDM2, LIMK1 and LIMK2). Out of these 4 genes, VEGFA and MDM2 are

494  relatively well-studied effectors of liver cancer, and there are several drugs and drug

495  candidates that are being studied in the context of targeting these proteins as reported in the

496  literature (please see https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL1783 and

497  https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL5023, for VEGFA and MDM2,

498  respectively). However, it is not possible to state the same for LIM kinases.

499    Metastatic potential and the invasiveness of cancer cells is dependent on the regulation of

500    cytoskeletal remodeling and cell migration. LIMK proteins (i.e., serine/threonine-protein

501    kinases) play important roles in metastasis by phosphorylating cofilin proteins which are

502    involved in the dynamic remodeling of actin filaments [30]. LIMKs are required for the

503    collective invasion by taking roles in invadopodium formation and extracellular matrix

504    degradation in cancer cells [38,39]. Guo *et al.* reported a critical increase in LIMK1 levels in

505    HCC, compared to the normal liver tissue. They also showed that the proliferation and

506    migration capacity of liver cancer cells are suppressed with the down regulation of LIMK1

507    [40]. Another study reported that the inhibition of LIMK/cofilin pathway via PAK1 inhibition

508    (i.e., an activator of LIMK), suppresses the growth of several HCC cell lines, which is

509    accompanied by decreased tumor value *in vivo*, due to the enhancement of apoptosis in

510    relation to the blocked NF-kB activation [41].

511    Recent studies have shown that LIMK inhibition is effective in terms of decreasing

512    proliferative and metastatic features of various tumor cells [31]. It has also been reported

513    that motility and invasion capacity of breast and prostate cancer cells (with overexpressed

514    LIMK1) was attenuated when inhibitors of upstream LIMK regulators are administered [42].

515    In the light of this information, LIM kinases can be considered as suitable candidates to

516    target cancer progression and metastasis.

517    With the aim of evaluating the prognostic effects of LIMK1 and LIMK2, we performed a

518    Kaplan-Meier survival analysis on liver cancer (HCC) RNA-seq data, collected from 364

519    patients [43]. It is observed from the results that the high expression of LIMK1 is significantly

520    associated with a lower survival rate in HCC patients (Figure S1a). The median survival rate

521    was calculated as 71 months in patients with low LIMK1 expression; whereas, the mean was

522    37.8 months in patients with high LIMK1 expression. On the other hand, we did not observe

523    a clear association between the survival rate and LIMK2 gene expression levels (Figure

524    S1b). To examine the documented functional and sequence-based similarities/differences

525    between LIMK1 and LIMK2, we checked their domain annotations from the InterPro

526     database, and observed that these proteins have the exact same domain architecture (i.e.,

527     from N to C terminal: "IPR001781: Zinc finger, LIM-type" – "IPR001781: Zinc finger, LIM-

528     type" – "IPR001478: PDZ domain" – "IPR001245: Serine-threonine/tyrosine-protein kinase,

529     catalytic domain", detailed information can be obtained from:

530     https://www.ebi.ac.uk/interpro/protein/reviewed/P53667 and

531     https://www.ebi.ac.uk/interpro/protein/reviewed/P53671). A Needleman–Wunsch pairwise

532     global sequence alignment between LIMK1 and LIMK2 (with default parameters; the scoring

533     matrix:BLOSUM62, gap open:10, gap extend:0.5) showed that these proteins are 51.9%

534     identical and 66.8% similar to each other, and non-similar positions are mainly concentrated

535     around the N and C terminal ends (the full alignment output is given in Supplementary

536     Material section 3). Based on these results, it can be stated that further research is required

537     to analyze different aspects of LIMK1 and LIMK2 in various types of cancer.

538     In the light of this information, we decided to target LIM kinases in the experimental

539     validation part of this study, with the main focus on LIMK1. Among the large-scale

540     drug/compound – target interaction predictions provided by DRUIDom, 4 compounds have

541     been predicted as inhibitors of both LIMK1 and LIMK2 proteins. Structures of these

542     compounds are given in Figure 2 together with their ChEMBL database identifiers and short

543     names as used in this study. These compounds are associated with LIMKs over their

544     "Serine-threonine/tyrosine-protein kinase, catalytic domain" (InterPro domain id:

545     IPR001245). In addition, we designed, synthesized, and tested 4 novel derivatives of the

546     compound: "LIMKi-2" (derivatives: LIMKi-2 to d in Figure 2), which is found to be the most

547     active one among the originally predicted inhibitors (explained below).

548     **Figure 2.** Structures, database identifiers, and 2-D representations of predicted LIMK

549     inhibitory compounds (LIMKi-1, 1a, 2, and 3) and derivatives (LIMKi-2a, b, c, and d).

2.4.2 Molecular Docking of Novel LIMK Inhibitors

551 For *in silico* validation of computationally predicted LIMK inhibitors, molecular docking

552 analyses were conducted. LIMK proteins (LIMK1 and LIMK2) are serine/threonine kinases

553 with multidomain structures including 2 LIM zinc-binding domains, 1 PDZ domain, and 1

554 protein kinase domain. Multi-kinase inhibitor staurosporine and previously described LIMK

555 inhibitor 9D8 have published crystal structures with the kinase domains of LIMK1 and LIMK2

556 proteins. These molecules were used as reference for docking, i.e., docking-based binding

557 free energy (ΔG) output of computationally predicted LIMK inhibitors are evaluated in

558 comparison to the docking output of these native ligands. In addition to computationally

559 predicted compounds (i.e., LIMKi-1, LIMKi-1a, LIMKi-2, and LIMKi-3), novel derivatives of

560 LIMKi-2 (i.e., LIMKi-2a, LIMKi-2b, LIMKi-2c, and LIMKi-2d) were also docked against kinase

561 domains of LIMK1 and LIMK2 proteins. AutoDock grid box parameters used in these

562 analyses are displayed in Table 2a, and the docking results of each LIMK protein –

563 compound combination are shown in Table 2b, which displays the lowest of the binding free

564 energies calculated from several poses obtained either from rigid or flexible docking in

565 AutoDock. All files and results of the docking analysis, including the ones for online

566 MTiAutoDock and SwissDock docking runs, are available in the data repository of the study

567 (https://github.com/cansyl/DRUIDom) and in Table S4, respectively. Docking results

568 obtained from different tools are consistent with each other. Based on the results in Table

569 2b; LIMKi-2, LIMKi-2d, and LIMKi-3 have binding free energy values close to that of the

570 reference ligand staurosporine ("staurosporine" ΔG=-10.55 kcal/mol, Ki=18.47 nM; "9D8"

571 ΔG=-12.38 kcal/mol, Ki=0.837 nM) for the LIMK1 protein, where the lower values indicate

572 stronger interactions. As for the LIMK2 protein, binding free energy values for all ligands,

573 except LIMKi-1 and LIMKi-1a, were around the generally accepted thresholds to assume a

574 potential activity (i.e., -10 to -12 kcal/mol), which were close to the value of reference ligand

575 9D8 (i.e., -12.38 kcal/mol). In Figure 3, the best poses of LIMKi-2 and LIMKi-3 dockings

576 against kinase domain binding sites of LIMK proteins are visualized along with the docking

577 of reference molecules. The results indicate computationally predicted LIMK inhibitors,

578    especially LIMKi-2 (including its derivatives) and LIMKi-3, could be promising candidate

579    molecules for targeting LIM kinases.

580    **Table 2. (a)** Grid box parameters for AutoDock in the molecular docking analysis; **(b)**

581    molecular docking results of computationally predicted LIMK inhibitors and their derivatives

582    against kinase domains of LIMK proteins in terms of binding free energy ($\Delta G$) and inhibition

583    constant (*Ki*) estimations at the best poses (selected with respect to lowest $\Delta G$).

584    **(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 |

585    **(b)**

|  | $\Delta G$ (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 |

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

588 **Figure 3.** Visualization of the docked complex structures of **(a)** LIMK1 kinase domain in

589 complex with the reference molecule staurosporine (green), LIMKi-2 (violet), and LIMKi-3

590 (red), and **(b)** LIMK2 kinase domain in complex with the reference molecule 9D8 (dark

591 cyan), LIMKi-2 (violet), and LIMKi-3 (red) at the selected best poses with lowest binding free

592 energy ($\Delta$G). Hydrogen bonds are displayed with dark blue lines. Gold and pink colors

593 represent LIMK1 and LIMK2 protein residues interacting with the corresponding compounds.

594 2.4.3 *In vitro* Experimental Analysis of LIMK Inhibition

595 *LIMKi Compounds have inhibitory effects on human cancer cells*

596 To address whether predicted inhibitors have cytotoxic effects on transformed normal

597 human (HEK-238) and various epithelial cancer cell lines (e.g., MCF-7, HCT116, Huh7, and

598 Mahlavu), cells were treated with LIMKi compounds with a concentration gradient of 40 $\mu$M

599 to 2.5 $\mu$M for 72 hours. The resulting cytotoxic $IC_{50}$ values are given in Table 3a. While there

600 is no cytotoxicity observed on normal cells, LIMKi-2 and LIMKi-3 compounds display

601 cytotoxic activities between 5.5-17.3 $\mu$M on cancer cells.  Since LIMKi-2 showed the most

602 potential bioactivity, we synthesized four novel derivatives of LIMKi-2 and assessed their

603 bioactivities on Huh7 and Mahlavu liver cancer cells. LIMKi-2 derivatives; 2c, 2d displayed

604 cytotoxic activities on Huh7 and Mahlavu cells (~8$\mu$M and <20$\mu$M, respectively), while

605 LIMKi-2a had no effect (Table 3b).

606 **Table 3.** Cytotoxic bioactivities of LIMKi molecules on human cells: **(a)** LIMKi-1,3

607 compounds **(b)** LIMKi-2 derivatives.

608 **(a)**

| LIMKi molecules | $IC_{50}$ Values ($\mu$M) | | | |
| --- | --- | --- | --- | --- |
| | 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 |

609 **(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 |

610

611 As stated above, phosphorylated LIMK proteins are involved in actin cytoskeleton dynamics

612 through cofilin phosphorylation, hence we performed experiments on the migration and

613 invasion properties of liver cancer cells in the presence of LIMK inhibitors. We focused on

614 Huh7 and Mahlavu liver cancer cells for the rest of the study, because primary liver cancer

615 (hepatocellular cancer, HCC) usually presents with multiple tumors within the liver and

616 intrahepatic metastatic spread is a major problem for this cancer [44].

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

618 Cofilin is a downstream molecule and its function is regulated by LIMK. Hence, we assessed

619 phospho-Cofilin protein levels in Huh7 and Mahlavu cells in the presence of LIMK inhibitors.

620 Phosphorylation of cofilin by LIMKs is significantly reduced upon treatment with LIMK

621 inhibitors in both Huh7 and Mahlavu cells except for LIMKi-1 and LIMKi-2d, respectively

622 (Figure 4a, b). Mahlavu cells are reported to have a resistant phenotype due to PTEN tumor-

623 suppressive protein deficiency for migration [45]. Therefore, the differential response against

624 LIMK inhibitors by well-differentiated Huh7 cells and poorly differentiated drug-resistant

625 Mahlavu cells are as expected and allows us to better assess the dose-response of LIMK

626 inhibitors.

25

627   The ratio of phosphorylated to non-phosphorylated Cofilin protein levels, together with LIMK

628   protein phosphorylation was previously reported as an indication of the metastatic potential

629   of a cell [30]. Therefore, we also checked the ratio of phospho- to total Cofilin levels for both

630   Huh7 and Mahlavu cells (Figure 4a, b) and found that LIMK inhibitors decreased the

631   phospho-Cofilin ratio significantly. These results may lead to the discovery of novel

632   therapeutic agents against the metastatic capacity of hepatocellular carcinoma cancer cells.

633   **Figure 4**. Phospho-Cofilin protein expression; **(a)** Huh7 and **(b)** Mahlavu cells were cultured

634   with LIMK inhibitors (20 μM) for 48 hours and expression of active p-Cofilin and total Cofilin

635   levels were assessed with western blot analysis. The bar graph indicates the relative

636   intensity of p-Cofilin levels compared to untreated DMSO controls. The equal loading control

637   was analyzed based on the total protein staining normalization protocol. The ratios of

638   phospho- and total Cofilin levels for both Mahlavu and Huh7 cell lines were calculated.

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

640   LIMK/Cofilin/ADF cascade has been described as one of the major regulators for actin

641   cytoskeleton dynamics and reorganization [46]. Bioactivities of LIMKi compounds were

642   tested for their effects on the migration and invasion capacity of HCC cell lines by wound

643   healing and real-time cell invasion Transwell assays, respectively. First, Huh7 cell migration

644   was analyzed in the presence of predicted LIMK inhibitors 1, 1a, 2, and 3. Huh7 cells have

645   less migration capability compared to Mahlavu cells, so Huh7 migration was only tested with

646   the originally predicted molecules. LIMKi-2 and LIMKi-3 strongly reduced the migration (2%

647   gap closure) of Huh7 cells when compared to DMSO controls (48% gap closure) within 10

648   hours (Figure 5a).  Then LIMKi-1, LIMKi-1a, LIMKi-2, LIMKi-3 and LIMKi-2 derivatives were

649   tested on the migration of Mahlavu cells. LIMKi-2 derivatives reduced the resistant Mahlavu

650   cell migration by 2.6-3.7 folds when compared to DMSO controls (Figure 5b).

651   We also tested the bioactivities of predicted compounds and their derivatives by real-time

652   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 $\mu$M 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. 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 drug/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 (https://github.com/cansyl/DRUIDom) and can be used by

678    researchers working in the fields of drug discovery and repurposing, both as a training and

679    benchmark dataset for the construction of new computational predictive models.

680    The idea behind DRUIDom's methodology is to identify the protein domains that are required

681    for successful interaction, and propagating these associations to proteins that possess those

682    same domains. Thus, it was critical to successfully separate compound – domain mappings

683    that indicate a true relationship from incidences observed by chance. For this, we

684    incorporated known/verified compound – target protein relations with undesired bioactivity

685    levels (i.e., high $xC_{50}$ values: > 20 $\mu$M) as "inactives" even though they also are interactors,

686    along with "actives" (compound – target protein pars with the desired levels of bioactivity:

687    $xC_{50}$ < 10 $\mu$M), as two different datasets. This approach enabled us to score compound –

688    domain mappings in terms of potential true-false positives and true-false negatives (as

689    explained in the Methods section 4.2.1), and to identify interacting pairs with a potential to

690    ultimately become new treatment options.

691    One limitation of our data-centric methodological approach is penalizing a compound –

692    domain mapping with a false negative count if one of the known active target proteins does

693    not contain the mapped domain. It is known that a small molecule (or fragment) can be the

694    ligand of different proteins and different domains, especially when the structural features of

695    the corresponding binding sites are similar to each other. In such cases penalizing a

696    mapping leads to the underestimation of its mapping score. In order to minimize this effect,

697    we took the InterPro domain hierarchy into account while calculating the mapping scores.

698    InterPro combines domains from the same functional family under distinct hierarchical trees.

699    There are also significant similarities between the sequence profiles of domains from the

700    same hierarchy. In DRUIDom, while scoring a mapping, we checked whether the known

701    active and inactive target proteins of the intended compound possess domains from the

702    same hierarchy. Therefore, we counted an active target protein containing a domain from

703    the same hierarchy (but not the actual mapped domain) as a true positive (instead of false

704    negative) and counted an inactive target protein containing a domain from the same

705 hierarchy as a false positive (instead of true negative). In this way, domain similarity has

706 been incorporated in DRUIDom. However, there are also cases where a single compound

707 binds to domains from completely different hierarchies. Our approach currently does not

708 take these cases into account.

709 During the parameter optimization and performance analyses of DRUIDom, it was important

710 to make sure that there was no data leak from the benchmark test dataset to our training set.

711 This condition has been automatically satisfied since the source of the mappings in the

712 InteracDome benchmark dataset (i.e., PDB co-complex structures) and the source of the

713 mappings in our training dataset (i.e., assay-based biological activity measurements

714 obtained from ChEMBL and PubChem databases) are completely independent from each

715 other.

716 In our analysis, we observed that only a small portion of the InterPro domain entries appear

717 in the finalized compound – domain mappings, with the total number of 250 domains, as

718 opposed to 8,165 compounds, at the selected mapping score threshold (i.e., 0.5). The main

719 reason behind this observation could arise from the data distribution in the source bioactivity

720 dataset; i.e., members from the same protein superfamilies have been targeted in most of

721 the experimental bioassays (e.g., kinases, GPCRs). The distribution of the number of

722 compounds mapped to each domain reveals that the top ten domains constitute 56.7% of

723 27,032 mappings in total (i.e. "IPR000719 - Protein kinase domain", "IPR001245 - Serine-

724 threonine/tyrosine-protein kinase, catalytic domain", "IPR017452 - GPCR, rhodopsin-like,

725 7TM", "IPR020635 - Tyrosine-protein kinase, catalytic domain", "IPR028174 - Fibroblast

726 growth factor receptor 1, catalytic domain", "IPR030611 - Aurora kinase A", "IPR034670 -

727 Checkpoint kinase 1, catalytic domain", "IPR035588 - Janus kinase 2, pseudokinase

728 domain", "IPR035589 - Janus kinase 2, catalytic domain", "IPR039192 - Glycogen synthase

729 kinase 3, catalytic domain"). Overall, eight out of ten of these domains belong to kinases.

730 We examined the difference in target proteins between our source bioactivity dataset and

731 the resulting predicted DTI dataset, to observe if it was possible to produce predictions for

732 under-studied proteins with the approach outlined in this study. The unique number of target

733 proteins in our source bioactivity dataset is 3,644, whereas, this number is 5,563 for our

734 finalized DTI prediction dataset, which indicates that there is a 52.7% increase in target

735 proteins due to the domain-based association approach. We also checked the protein

736 superfamily distribution of the targets in the original and the predicted interaction datasets,

737 considering five main classes of proteins as enzymes, membrane receptors, ion channels,

738 transcription factors, and others (i.e., a combination of transporters, epigenetic regulators,

739 secreted proteins, other cytosolic proteins, other nuclear proteins, and other categories),

740 according to the first level (L1) of ChEMBL protein classification

741 (https://www.ebi.ac.uk/chembl/g/#browse/targets). For this, we compared the target protein

742 family distribution in the original bioactivity dataset (i.e., 64% enzymes, 11% membrane

743 receptors, 5% ion channels, 4% transcription factors, and 16% others) with our DTI

744 prediction dataset (i.e., 50% enzymes, 25% membrane receptors, 7% ion channels, 8%

745 transcription factors, and 10% others). Although dominating families in the source bioactivity

746 dataset prevail in the predicted DTIs dataset, we were able to produce interacting compound

747 predictions for a critically higher number of proteins from membrane receptor, ion channel,

748 and transcription factor superfamilies with a 248%, 114%, and 238% increase, respectively.

749 These results, again, demonstrate the effectiveness of the domain-based approach in

750 predicting new target proteins.

751 In this study, we aimed to validate our drug/compound – target protein interaction prediction

752 method by targeting the PI3K/Akt/mTOR pathway by focusing on the predicted LIM kinase

753 inhibitors. The importance of selecting LIMKs as targets comes from their unique kinase

754 domains which have longer activation loops compared to many kinases, allowing the design

755 of specific inhibitors against cancer invasion and metastasis [42]. Furthermore, LIMK1

756 knockout was not embryonically lethal in mice making this protein a good candidate for drug

757 design [47]. Another study showed that LIMK activity is beneficial for cancer cells in terms of

758 coping with chemotherapeutics and ionizing radiation, which renders cells resistant to these

759   treatments [48-51]. Therefore, LIMKs are promising candidates due to their essential role in

760   cytoskeletal remodeling leading to cell migration and invasion. Hence, the lack of cytotoxicity

761   of our predicted compounds on normal transformed HEK-238 cells is in parallel with the

762   above-mentioned cellular LIMK activities, which is prominent in cancer cells.

763   For the validation study, we initially examined the binding properties of 4 originally predicted

764   compounds (i.e., LIMKi-1, 1a, 2, and 3) by computational docking and comparing with the

765   crystal structures of multi-kinase inhibitor staurosporine and previously identified LIMK ligand

766   9D8 in complex with LIMK1 and LIMK2 proteins, respectively. LIMKi-2, its derivatives, and

767   LIMKi-3 had the most significant binding energies. During the *in vitro* validation stage of the

768   study, we performed bioactivity experiments on liver cancer cells because intrahepatic

769   metastatic migration/invasion is a major problem for patient survival and the specific

770   selection of treatment is dependent on the number of distinct cancer nodules within the

771   organ [52]. Our observations from the docking analysis were further supported by

772   cytotoxicity and migration/invasion experiments where LIMKi-2 was the most significant

773   compound regarding its action on cancer cells. Our promising results with LIMKi-2 directed

774   us to synthesize 4 novel derivatives of this compound (i.e., LIMKi-2a, b, c, and d). Among

775   these derivative compounds, LIMKi-2c and LIMKi-2d displayed highly significant anti-

776   migratory and anti-invasive properties on liver cancer cells, together with strong docking

777   binding affinities. The increased activity for LIMKi-2c and 2d is interesting and seems to

778   indicate a favorable change in conformation due to the bromide substituent that twists the

779   benzene ring against the thiadiazole and causes loss of co-planarity. Finally, our evaluation

780   singled out the novel LIMKi-2d compound as a promising candidate therapeutic agent due to

781   its action on mesenchymal Mahlavu cells which are highly aggressive in terms of drug

782   resistance for cytotoxicity, motility, and migration [53].

783   As future work, we plan to further develop our predictive approach by identifying

784   associations between ligands and experimentally characterized protein structures (from

785   Protein Data Bank) and high-quality structure models generated by cutting-edge structure

786    prediction methods [54,55]. Furthermore, we plan to extend our mappings to

787    uncharacterized protein sequence signatures using sources such as Pfam's domains of

788    unknown function (DUFs) [56], and potentially functional regions detected by different

789    computational approaches [57]. Additionally, we are going to integrate DRUIDom's

790    compound – domain and compound – target interaction predictions to our large-scale

791    biological and biomedical data integration and representation system CROssBAR [58] with

792    the aim of enriching the biological relationship-based information provided in this service

793    (https://crossbar.kansil.org/). This way, users can easily browse pre-computed DRUIDom

794    associations/predictions for their proteins of interest, on the fly, together with other types of

795    biomolecular relationships provided in this system (i.e., genes/proteins to diseases,

796    phenotypes, pathways/functions, drugs, in addition to PPIs). Finally, we plan to extend the

797    work on LIMK inhibition with additional *in vitro* experiments and *in vivo* studies, with the

798    ultimate aim of contributing to the development of new anti-cancer drugs.

799    The computational drug/compound – target protein interaction prediction approach proposed

800    in this study led to the identification of novel interactions, a selected subset of which were

801    then validated by both *in silico* and *in vitro* experiments. Results of the cell-based validation

802    experiments indicate DRUIDom has the ability to generate generalized predictions that are

803    well-translated into higher organizational levels such as the cell. Also based on these

804    results, it is possible to state that the approach proposed here is producing biologically

805    relevant results that can be utilized in drug discovery and repurposing studies beyond

806    PI3K/Akt/mTOR pathway and cancer, especially for pathological conditions where specific

807    domain-based targeting may be critical, such as metabolic disorders.

808

809    **4. Methods**

810    In this section, we first explain employed procedures for dataset construction and data

811    processing (4.1), then we provided details of the proposed DTI prediction system, DRUIDom

812  (4.2), which is followed by the definition of scoring functions (4.3), particulars of docking

813  analyses (4.4), and finally, a short summary of chemical synthesis and *in vitro* experiment

814  procedures (4.5), details of which are given in Supplementary Material sections 1 and 2.

815  **4.1 Data**

816  4.1.1 The bioactivity dataset

817  Bioactivity data points, each of which indicates the experimentally verified interaction

818  between a compound and a target biomolecule (i.e., protein), were downloaded from open-

819  access bioassay databases and divided into 2 classes as active (i.e., interacting) and

820  inactive (i.e., non-interacting, or more precisely: "non-interacting at the desired level") pairs.

821  For the selection of active data points, we used a bioactivity value threshold of < 10 $\mu$M x$C_{50}$

822  (i.e., $IC_{50}$ or equivalent). For inactives, we used a bioactivity value threshold of > 20 $\mu$M x$C_{50}$.

823  The data points between 10 and 20 $\mu$M were discarded, since their classification to either

824  class was considered to be ambiguous.

825  ChEMBL bioactivity database [17] and PubChem bioassay database [16] were used as the

826  bioactivity data source. The bioactivity data was acquired from the ChEMBL database (v23)

827  via SQL queries with specified parameters (i.e., assay type: binding, target type: single

828  protein, taxon: metazoa, standard value:  < 10 $\mu$M for active/interacting pairs and > 20 $\mu$M

829  for inactive/non-interacting pairs). We only selected the data points with a pChEMBL value,

830  which corresponds to a calculated activity measure of half-maximal response

831  concentration/potency/affinity (e.g., $IC_{50}$, $EC_{50}$, $AC_{50}$, Ki, Kd, and potency) in the negative

832  logarithmic scale. pChEMBL value of 5 is equal to an x$C_{50}$ measurement of 10 $\mu$M. The

833  presence of a pChEMBL value indicates that the data point has been checked by a curator.

834  Following these filtering operations, there were still cases where multiple bioactivity values

835  are reported between a particular compound and target (i.e., duplicates). We thus take the

836  median bioactivity value into account to treat these cases, similar to previous studies [10].

837  After the elimination of duplicates, the final ChEMBL set contained 718,102 bioactivity data

838     points (627,353 actives and 90,749 inactives) between 3,533 target proteins and 467,658

839     compounds.

840     Due to the structural organization of the PubChem bioassay database, it was not

841     straightforward to obtain a bioactivity dataset with desired properties. However, the

842     developers of ExCAPE-DB solved this problem by extensively filtering and organizing

843     PubChem bioactivity data (together with ChEMBL bioactivity data) and presented the results

844     in a database [59]. ChEMBL v20 and the PubChem bioassay database (January 2016) are

845     incorporated in ExCAPE. In our study, we incorporated PubChem bioactivities directly using

846     the ExCAPE-DB. We discarded the PubChem data points where the actual bioactivity values

847     were missing. These points could have been included using the assay outcome field, where

848     each data point is already marked as either "active" or "inactive"; however, the test

849     concentrations for these data points are not available, and it is probable that many of them

850     do not obey the thresholds we determined. Following the elimination of data points with

851     activity values between 10 and 20 $\mu$M, the final ExCAPE bioactivity dataset contained

852     2,514,439 bioactivity values between 1,648 target proteins and 856,216 compounds. The

853     reason behind the low number of target proteins compared to the ChEMBL dataset was that,

854     in ExCAPE, only three organisms (i.e., human, mouse and rat) were included. Finally,

855     ChEMBL v23 and ExCAPE datasets were merged to obtain the finalized bioactivity training

856     dataset of the study. Since ExCAPE-DB incorporates ChEMBL data (from v20, which is an

857     older version compared to the one we used) along with PubChem, many duplicates were

858     added to our dataset following merging, which were eliminated by simply deleting repeat

859     data points. Our finalized source bioactivity dataset contains 2,869,943 data points between

860     3,644 target proteins and 1,033,581 compounds. 1,637,599 of these data points are in the

861     actives class, and the remaining 1,232,344 are in the inactives class. The contradictions

862     between active and inactive classes (i.e., compound – protein pairs that are listed both as

863     active and inactive) are low, with only 1,574 cases (< 0.06%).

864    <u>4.1.2 Target proteins and domains</u>

865    UniProt Knowledgebase -UniProtKB- v2019_01 [25] and InterPro v72 database [20] were

866    employed as the source for target protein sequences and their domain annotations,

867    respectively. InterPro integrates sequence signatures with functional significance from 13

868    different manually curated and automated databases presenting functional and structural

869    protein information. In InterPro, domain content, order and positions are pre-computed for

870    each UniProtKB protein sequence using the InterProScan tool and the sequence

871    profiles/HMMs and presented within a public dataset. We downloaded InterPro annotations

872    for all of the target proteins in our dataset (i.e., 3,644) and eliminated the InterPro hits for

873    non-domain type entries such as families and sites. This resulted in a total of 3,118 target

874    proteins that had at least one InterPro domain hit, and thus, could be further used in our

875    study. The average number of domains in these target proteins was 2.44. We also

876    generated domain architectures, which can be defined as the linear arrangement of the

877    domain hits on the protein sequence, for each multi-domain protein in our dataset. The

878    domain architecture information is later used for mapping compounds to domain pairs, to

879    account for the cases where multiple domains are required to be presented in the protein to

880    have an interaction with the corresponding compound (the detailed procedure is described

881    below).

882    <u>4.1.3 Compound representation and analysis</u>

883    Canonical SMILES notations were employed to represent the compounds. SMILES is a

884    widely used system that defines the structures of chemical species as line notations [60].

885    SMILES representations of all compounds in our dataset were directly downloaded from

886    ChEMBL and PubChem databases. Extended-Connectivity Fingerprints (ECFP4) [61] were

887    generated for all compounds in our bioactivity dataset (i.e., 1,033,581), using SMILES as the

888    input. Pairwise molecular similarities were measured between all compound pair

889    combinations using the Tanimoto coefficient. Python RDKit module [62] and ChemFP library

890    [63] were employed to generate the fingerprints and to calculate the pairwise molecular

891    similarities.

**4.2 DTI Prediction System**

893    The proposed prediction system contains two modules: compound – domain mapping

894    (section 4.2.1) and the propagation of associations to other proteins and compounds

895    (section 4.2.2). In the mapping module, small molecule drugs/compounds are

896    probabilistically associated to single domains (or domain pairs) on target proteins, using

897    experimentally verified compound – target interaction data in bioactivity data resources. In

898    the second module, for each compound – domain pair, all proteins that contain the mapped

899    domain and all compounds that are significantly similar to the mapped compound (in terms

900    of molecular similarity) are crossed with each other to produce new drug/compound – target

901    protein predictions.

4.2.1 Compound – domain mapping

903    Figure 1a displays the overall methodology within a schematic representation. In this

904    example, a compound ($C_i$) and its target protein ($P_1$) is reported to be interacting/bioactive

905    (i.e., according to our definition of active; $xC_{50} < 10\ \mu M$) in ChEMBL and/or PubChem. In this

906    toy example, it has been identified from the InterPro database that $P_1$ has one domain

907    annotation (i.e., blue domain), on which the binding site/region of $C_i$ (with the desired

908    bioactivity) is assumed to reside. This makes other human proteins containing the blue

909    domain (i.e., $P_2$, $P_3$, and $P_4$) candidate targets for $C_i$ and for other drug-like compounds that

910    are significantly similar to $C_i$ with Tanimoto similarity greater than or equal to 0.8 (i.e., $C_x$, $C_y$,

911    and $C_z$).

912    To quantize the association between a compound and a domain, we calculated mapping

913    scores for each compound – domain combination, using verified active and inactive

914    compound – target protein data points in our source ChEMBL + PubChem bioactivity

915    dataset. For this, precision, recall, accuracy, F1-score, and Matthew's correlation coefficient

916   (MCC) metrics are employed. MCC successfully measures the quality of binary

917   classifications when there is a class imbalance [64], such as the case observed in our

918   dataset. Here, binary classification is the decision for either the presence or absence of a

919   bio-interaction between a compound and a domain. Definitions below are used to calculate

920   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).

933   Mapping score metrics were calculated using the above-defined TP, FP, FN, and TN; with

934   their formulations being provided in Methods section 4.3. For all the compound – domain

935   mappings, high scores indicate reliable mappings and a high probability that the region of

936   interaction lies on the mapped domain. In Figure 1b, the mapping procedure is shown for 2

937   toy examples. Also, in Figure 1b, the number of TP, FP, FN, and TN for toy examples are

938   given, together with the respective mapping scores (i.e., metrics). The first example

939   corresponds to a case where there are 2 experimentally verified interacting (i.e., active)

940   target proteins for compound $C_1$. Both of these proteins contain the blue domain (i.e., a

941   structural unit responsible for the interaction with $C_1$.). $C_1$ also has 3 inactive proteins (i.e.,

942    targets with insufficient bioactivity), 2 of which contain the red domain and 1 contains the

943    light green domain. With the selection of the domain with the maximum score, the blue

944    domain is mapped to $C_1$. A further example mapping case is presented for compound $C_2$,

945    where most of the known targets are multi-domain proteins. For $C_2$, many of the targets

946    contain the green domain, red domain, or both of them. Association scores for single

947    domains and domain pairs revealed that the best score is achieved when green and red

948    domains exist together. It is observed that the real-world cases can be much more

949    complicated compared to the toy examples provided in Figure 1b, as one protein can be the

950    target of multiple compounds and one compound can target multiple proteins. To be able to

951    separate reliable mappings from the non-reliable ones we determined and applied mapping

952    score thresholds using the metrics provided in section 4.3. The test applied to determine

953    these thresholds is described (together with its results) in the Results section 2.1.

954    With the purpose of increasing the reliability of the data in our verified bioactivity dataset, we

955    directly eliminated the mappings to the compounds if the number of active and inactive

956    targets is less than 3 (each). This filter was applied to eliminate the compounds with only a

957    few data points, which could otherwise produce false high mapping scores. This application

958    dramatically reduced the number of compounds in our source dataset from 1,033,581 to

959    51,750. To be able to incorporate more data points, we generated a second dataset by

960    combining the active and inactive targets of the compounds in clusters, which were

961    significantly similar to each other in terms of molecular structure, and treated each cluster as

962    an individual compound while calculating the mapping scores. To distribute the compounds

963    in clusters we used pairwise molecular similarities via Tanimoto coefficient (over ECFP4

964    fingerprints) with a threshold of 0.7, which was above the previously applied threshold to

965    predict targets based on compound molecular similarities [65]. All compounds that were

966    similar to each other with at least 0.7 Tanimoto similarity were placed in the same cluster.

967    Clusters with less than 5 active and 5 inactive targets were directly eliminated to ensure

968    reliability in terms of the number of data points. In this way, 202,238 clusters were generated

969 with compound overlaps in-between. This procedure should not be confused with compound

970 similarity-based propagation of target protein associations, which is explained in section

971 4.2.2 below. The mapping score calculation was carried out for all of the 51,750 individual

972 compounds in our first dataset (i.e., single-compound-based mappings) and for 202,238

973 clusters in our second dataset (i.e., compound-cluster-based mappings) against domains of

974 their respective target proteins. For the compound-cluster-based analysis, the score

975 obtained for each domain mapping was propagated to all compounds in the corresponding

976 cluster. This resulted in a total of 3,487,239 raw compound – domain mappings for the

977 cluster-based bioactivity dataset (i.e., compound-cluster-based mappings) and 449,294 raw

978 mappings for the individual compound-based dataset (i.e., single-compound-based

979 mappings).

980 Figure 7 displays the histograms composed of bins of the total number of targets, the

981 number of active targets, and the number of inactive targets (X-axis), for individual

982 compounds (Figure 7a, b, c) and for compound clusters (Figure 7d, e, f). Y-axis represents

983 the number of compounds or compound clusters in the log scale. As observed, there was a

984 steady decrease in the number of compounds/clusters when the number of targets per

985 compound/cluster was increased. There was also a clear difference between active and

986 inactive target bins. Indeed, no individual compound or cluster with higher than 80 inactive

987 targets was identified. The most probable reason for this is that negative results (i.e., non-

988 interactions) are not usually reported in the literature. The gain from using compound

989 clusters was highlighted especially for active targets and for all targets (i.e., a vs. d and b vs.

990 e) with the increase in the height of the bars for more than 50 targets (notice the scaling

991 difference in the X-axis between the individual compound histograms and the compound

992 cluster histograms).

993 **Figure 7.** Log-scale histograms of the number of individual compounds and compound

994 clusters (Y-axis) with the given number of target proteins (X-axis) in our source bioactivity

995 dataset; for individual compounds: **(a)** all targets, **(b)** active targets, **(c)** inactive targets; and

996 for compound clusters: **(d)** all targets, **(e)** active targets, **(f)** inactive targets.

997 A similar procedure was applied to map compounds to domain pairs. For this, all domain

998 pair combinations were identified for each target protein in our source dataset, using the

999 domain architecture information of the proteins extracted using the UniProt-DAAC method,

1000 which was described in our previous study [66]. All domain pairs were recorded as if they

1001 were single domains and the mapping procedure explained above was applied to obtain

1002 compound – domain pair mappings. This procedure yielded a total of 1,075,550 raw

1003 individual compound – domain pair mappings and 9,343,130 raw compound cluster –

1004 domain pair mappings. The high number (compared to single domain mappings) was due to

1005 the elevated number of domain pair combinations, especially for large proteins.

1006 Once the mapping score threshold had been selected (as explained in the Results section

1007 2.1), all mappings below the threshold were discarded, and the remaining mappings

1008 constituted the finalized mapping dataset.

1009 <u>4.2.2 Propagation of associations</u>

1010 The second module starts with the detection of pairwise similarities between all compounds

1011 in our source dataset using molecular fingerprints. For this, Extended-Connectivity

1012 Fingerprints (ECFP4) [61] were generated for all compounds in our bioactivity dataset (i.e.,

1013 1,033,581). The pairwise similarities were measured using the Tanimoto coefficient with a

1014 threshold of 0.8 to signify significant similarities, which was even above the previously

1015 applied Tanimoto thresholds to safely transfer target annotations between small molecule

1016 compounds [65]. Briefly, domain associations that were produced in the previous step were

1017 transferred to new compounds that are similar to the mapped compound with a Tanimoto

1018 similarity value greater than or equal to 0.8. The idea behind this application was that

1019 structurally similar molecules tend to have similar interactions, as assumed in conventional

1020 ligand-based virtual screening [61].

1021  Subsequently, all human protein records in the UniProtKB/Swiss-Prot database were

1022  searched for the mapped domains and domain pairs, using the InterPro domain annotation

1023  information. When a new protein was found to contain the domain in question, it was

1024  associated with the corresponding compound. In this way, new candidate ligands were

1025  predicted for both known targets and for new candidate target proteins that possess the

1026  mapped domains or domain pairs (Figure 1a).

1027  **4.3 Mapping Score and Performance Analysis Metrics**

1028  Precision, recall, accuracy, F1-score, and Matthew's correlation coefficient (MCC) metrics

1029  are used for both the calculation of mappings scores (Methods section 4.2.1) and calculation

1030  of the overall system performance (Results section 2.1). The formulation of these metrics

1031  are as follows:

1032
$$Precision = \frac{TP}{TP+FP} \tag{1}$$

1033
$$Recall = \frac{TP}{TP+FN} \tag{2}$$

1034
$$Accuracy = \frac{TP+TN}{TP+FN+FP+FN} \tag{3}$$

1035
$$F1-Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{4}$$

1036
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{5}$$

1037  Definitions for TP (i.e., true positives), FN (i.e., false negatives), FP (i.e., false positives) and

1038  TN (i.e., true negatives) in the context of mappings scores are given in the Method section

1039  4.2.1. We also calculated the enrichment of reference/known compound – domain

1040  associations in DRUIDom mappings at different mapping score cut-offs, as a means of

41

1041 performance evaluation. For the calculation of enrichment score and its statistical

1042 significance, we used hypergeometric test for overrepresentation, as shown in equations 6

1043 and 7, respectively:

1044
$$E_t = \frac{m_t/n_t}{M_t/N_t} \tag{6}$$

1045
1046

1047
$$P_t = \sum_{i=m_t}^{n_t} \frac{\binom{M_t}{m_t}\binom{N_t-M_t}{n_t-m_t}}{\binom{N_t}{n_t}} \tag{7}$$

1048

1049 where $E_t$ is the enrichment score calculated at the mapping threshold score of $t$; $m_t$

1050 represents the number of compound – domain mappings retrieved by DRUIDom (at the

1051 mapping threshold score of $t$) that are also found in reference/true associations list; $n_t$

1052 represents the total number of compound – domain mappings retrieved by DRUIDom (at the

1053 mapping threshold score of $t$); $M_t$ is the total number of reference/true compound – domain

1054 associations when we limit our domain and compound lists to the ones left in the DRUIDom

1055 mappings list at the mapping threshold score of $t$; and $N_t$ represents the total number of

1056 compound – domain pair combinations (i.e., all random combinations) when we limit our

1057 domain and compound lists to the ones left in the DRUIDom mappings list at the mapping

1058 threshold score of $t$. $P_t$ represents the significance (p-value) of the enrichment at the mapping

1059 threshold score of $t$.

1060 **4.4 Molecular Docking Experiments**

1061 For the molecular docking of predicted inhibitor compounds and their derivatives against

1062 kinase domains of LIMK1 and LIMK2 proteins, the crystal structure of LIMK1 kinase domain

1063 as a complex with staurosporine (PDB id: 3S95) and the crystal structure of LIMK2 kinase

1064 domain complex with bound 9D8 (PDB id: 5NXD) were retrieved from RCSB PDB database

1065 [67]. Then, the PDB files of both protein structures were loaded into AutoDockTools-1.5.6.

1066 For both proteins, which are in the form of 2-chain homodimer structures, only the A chain

1067 was retained for docking and preprocessed by deleting all heteroatoms, adding hydrogen

1068    atoms, computing Gasteiger charges, and merging non-polar hydrogens. The preprocessed

1069    protein structures were saved as pdbqt files. For flexible docking, contact residues of LIMK1

1070    and LIMK2 proteins were selected and saved as flexible pdbqt files, while the remaining

1071    structures of the proteins were saved as rigid pdbqt files.

1072    Full 3-D structures of compounds were downloaded from ZINC (v15) database [68] in sdf file

1073    format and converted to PDB files by Open Babel file format converter [69]. Since the

1074    derivative compounds (i.e., LIMKi-2a, LIMKi-2b, LIMKi-2c, LIMKi-2d) could not be found in

1075    the ZINC database, compound 3-D structures (in the form of PDB files) were generated from

1076    the SMILES representations of respective compounds using ChemAxon JChem software-

1077    based online tool at: http://pasilla.health.unm.edu/tomcat/biocomp/convert. Then, Gasteiger

1078    charges were added, rotatable bonds and the root for the identification of a central atom

1079    were detected for compound PDB structures, and they were saved as pdbqt files in

1080    AutoDockTools.

1081    Grid map files for both rigid and flexible dockings were generated by AutoGrid4 program

1082    (AutoDock-4.2.6) [70] using protein and compound pdbqt files as inputs, and the X-Y-Z

1083    coordinates for the grid search were defined by calculating the mean coordinates of the

1084    reported interacting atoms of LIMK1 and LIMK2 proteins, which were retrieved from

1085    PDBsum [71]. Grid box parameters for grid search were set as shown in Table 2a. In the

1086    docking step, a genetic algorithm with default settings was used for parameter searching,

1087    and the docking analysis of each compound – protein pair was carried out by using

1088    AutoDock4 (v4.2.6) [70].

1089    As a second docking validation, the same analysis was also performed using MTiAutoDock

1090    [72] and SwissDock [73] web services. Protein pdb files were given as an input to the

1091    MTiAutoDock service together with the sdf formatted ligand structure files. List of residues

1092    mode was selected for grid calculation and the contact residues of each protein were given

1093    as input. MTiAutoDock service has automatically added the hydrogen atoms to the crystal

1094    structure and executed the docking procedure using AutoDock 4.2.6. For SwissDock, blind

1095     docking was implemented using protein PDB files and ligand mol2 files as input. For all

1096     docking analyses, different poses were evaluated via binding free energy calculations and

1097     the one with the lowest energy was selected as the finalized result (i.e., the best pose).

1098     UCSF Chimera software was used for the visualization of docking results.

1099     **4.5 Chemical Synthesis and *in vitro* Validation of the Predicted Inhibitors**

1100     DRUIDom predicted 4 compounds as inhibitors of LIMK1 and LIMK2 proteins, which have

1101     been selected as targets of the validation use-case study. Structures, database identifiers,

1102     and given names (by us) of these compounds (i.e., LIMKi-1, LIMKi-1a, LIMKi-2, LIMKi-3) are

1103     displayed in Figure 2. We synthesized these molecules to be used in the cell-based assays.

1104     Also, the structure of LIMKi-2 has been modified with the aim of building 4 new derivatives

1105     with a potentially higher biological activity (i.e., shown in Figure 2 as LIMKi-2a, LIMKi-2b,

1106     LIMKi-2c, LIMKi-2d), making a total of 8 molecules. Procedures used in the chemical

1107     synthesis of these molecules and the methodological details of *in vitro* experimental

1108     analyses are given in the Supplementary Material sections 1 and 2.

## 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.

1138    13.  Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction:
1139         a web server for target prediction of bioactive small molecules. Nucleic acids research.
1140         2014 Jul 1;42(W1):W32-8.

1141    14.  Zhou W, Wang Y, Lu A, Zhang G. Systems pharmacology in small molecular drug
1142         discovery. International journal of molecular sciences. 2016 Feb;17(2):246.

1143    15.  Talele TT, Khedkar SA, Rigby AC. Successful applications of computer aided drug
1144         discovery: moving drugs from concept to the clinic. Current topics in medicinal
1145         chemistry. 2010 Jan 1;10(1):127-41.

1146    16.  Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA,
1147         Yu B, Zaslavsky L. PubChem 2019 update: improved access to chemical data. Nucleic
1148         acids research. 2019 Jan 8;47(D1):D1102-9.

1149    17.  Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP,
1150         Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M. ChEMBL: towards direct
1151         deposition of bioassay data. Nucleic acids research. 2019 Jan 8;47(D1):D930-40.

1152    18.  Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins.
1153         Proceedings of the National Academy of Sciences. 1973 Mar 1;70(3):697-701.

1154    19.  El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M,
1155         Richardson LJ, Salazar GA, Smart A, Sonnhammer EL. The Pfam protein families
1156         database in 2019. Nucleic acids research. 2019 Jan 8;47(D1):D427-32.

1157    20.  Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY,
1158         El-Gebali S, Fraser MI, Gough J. InterPro in 2019: improving coverage, classification
1159         and access to protein sequence annotations. Nucleic acids research. 2019 Jan
1160         8;47(D1):D351-60.

1161    21.  Li Q, Cheng T, Wang Y, Bryant SH. Characterizing protein domain associations by
1162         Small-molecule ligand binding. Journal of proteome science and computational biology.
1163         2012 Dec 3;1.

1164    22.  Kruger FA, Rostom R, Overington JP. Mapping small molecule binding data to structural
1165         domains. InBMC bioinformatics 2012 Dec (Vol. 13, No. 17, pp. 1-13). BioMed Central.

1166   23. Kruger FA, Gaulton A, Nowotka M, Overington JP. PPDMs—a resource for mapping
1167       small molecule bioactivities from ChEMBL to Pfam-A protein domains. Bioinformatics.
1168       2015 Mar 1;31(5):776-8.

1169   24. Doğan T. HPO2GO: prediction of human phenotype ontology term associations for
1170       proteins using cross ontology annotation co-occurrences. PeerJ. 2018 Aug 2;6:e5298.

1171   25. Riley R, Lee C, Sabatti C, Eisenberg D. Inferring protein domain interactions from
1172       databases of interacting proteins. Genome biology. 2005 Oct;6(10):1-7.

1173   26. Lee H, Deng M, Sun F, Chen T. An integrated approach to the prediction of domain-
1174       domain interactions. BMC bioinformatics. 2006 Dec;7(1):1-5.

1175   27. Kobren SN, Singh M. Systematic domain-based aggregation of protein structures
1176       highlights DNA-, RNA-and other ligand-binding positions. Nucleic acids research. 2019
1177       Jan 25;47(2):582-93.

1178   28. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic acids
1179       research. 2019 Jan 8;47(D1):D506-15.

1180   29. Ersahin T, Tuncbag N, Cetin-Atalay R. The PI3K/AKT/mTOR interactive pathway.
1181       Molecular BioSystems. 2015;11(7):1946-54.

1182   30. Scott RW, Olson MF. LIM kinases: function, regulation and association with human
1183       disease. Journal of molecular medicine. 2007 Jun 1;85(6):555-68.

1184   31. Mardilovich K, Baugh M, Crighton D, Kowalczyk D, Gabrielsen M, Munro J, Croft DR,
1185       Lourenco F, James D, Kalna G, McGarry L. LIM kinase inhibitors disrupt mitotic
1186       microtubule organization and impair tumor cell proliferation. Oncotarget. 2015 Nov
1187       17;6(36):38469.

1188   32. Bagheri-Yarmand R, Mazumdar A, Sahin AA, Kumar R. LIM kinase 1 increases tumor
1189       metastasis of human breast cancer cells via regulation of the urokinase-type
1190       plasminogen activator system. International journal of cancer. 2006 Jun 1;118(11):2703-
1191       10.

1192   33. Pan Z, Liu C, Zhi Y, Xie Z, Wu L, Jiang M, Zhang Y, Zhou R, Zhao L. LIMK1 nuclear
1193       translocation promotes hepatocellular carcinoma progression by increasing p-ERK
1194       nuclear shuttling and by activating c-Myc signalling upon EGF stimulation. Oncogene.
1195       2021 Apr;40(14):2581-95.

1196    34.   Bu Z, Callaway DJ. Proteins move! Protein dynamics and long-range allostery in cell
1197        signaling. Advances in protein chemistry and structural biology. 2011 Jan 1;83:163-221.

1198    35.   England KS, Tumber A, Krojer T, Scozzafava G, Ng SS, Daniel M, Szykowska A, Che
1199        K, von Delft F, Burgess-Brown NA, Kawamura A. Optimisation of a triazolopyridine
1200        based histone demethylase inhibitor yields a potent and selective KDM2A (FBXL11)
1201        inhibitor. MedChemComm. 2014;5(12):1879-86.

1202    36.   Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima
1203        S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the
1204        environment. Nucleic acids research. 2007 Dec 12;36(suppl_1):D480-4.

1205    37.   Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, Dicker DJ,
1206        Chimed-Orchir O, Dandona R, Dandona L, Fleming T. Global, regional, and national
1207        cancer incidence, mortality, years of life lost, years lived with disability, and disability-
1208        adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the
1209        global burden of disease study. JAMA oncology. 2017 Apr 1;3(4):524-48.

1210    38.   Scott RW, Hooper S, Crighton D, Li A, König I, Munro J, Trivier E, Wickman G, Morin P,
1211        Croft DR, Dawson J. LIM kinases are required for invasive path generation by tumor
1212        and tumor-associated stromal cells. Journal of Cell Biology. 2010 Oct 4;191(1):169-85.

1213    39.   Lagoutte E, Villeneuve C, Lafanechère L, Wells CM, Jones GE, Chavrier P, Rossé C.
1214        LIMK regulates tumor-cell invasion and matrix degradation through tyrosine
1215        phosphorylation of MT1-MMP. Scientific reports. 2016 Apr 27;6(1):1-2.

1216    40.   Guo D, Li YR, Chen DF, Wang RH, Zhang D, Zhu M, He SX, Lu XL. Regulatory effects
1217        of LIM kinase 1 on the proliferation and metastasis of hepatocellular carcinoma cells.
1218        Chinese Journal of Hepatology. 2021 May 1;29(5):427-32.

1219    41.   Wong LL, Lam IP, Wong TY, Lai WL, Liu HF, Yeung LL, Ching YP. IPA-3 inhibits the
1220        growth of liver cancer cells by suppressing PAK1 and NF-κB activation. PLoS One.
1221        2013 Jul 19;8(7):e68843.

1222    42.   Yoshioka K, Foletta V, Bernard O, Itoh K. A role for LIM kinase in cancer invasion.
1223        Proceedings of the National Academy of Sciences. 2003 Jun 10;100(12):7247-52.

1224   43. Menyhárt O, Nagy Á, Győrffy B. Determining consistent prognostic biomarkers of overall
1225       survival and vascular invasion in hepatocellular carcinoma. Royal Society open science.
1226       2018 Dec 5;5(12):181006

1227   44. Tang ZY. Hepatocellular carcinoma-cause, treatment and metastasis. World journal of
1228       gastroenterology. 2001 Aug 15;7(4):445.

1229   45. Buontempo F, Ersahin T, Missiroli S, Senturk S, Etro D, Ozturk M, Capitani S, Cetin-
1230       Atalay R, Neri ML. Inhibition of Akt signaling in hepatoma cells induces apoptotic cell
1231       death independent of Akt activation status. Investigational new drugs. 2011
1232       Dec;29(6):1303-13.

1233   46. Mizuno K. Signaling mechanisms and functional roles of cofilin phosphorylation and
1234       dephosphorylation. Cellular signalling. 2013 Feb 1;25(2):457-69.

1235   47. Meng Y, Zhang Y, Tregoubov V, Janus C, Cruz L, Jackson M, Lu WY, MacDonald JF,
1236       Wang JY, Falls DL, Jia Z. Abnormal spine morphology and enhanced LTP in LIMK-1
1237       knockout mice. Neuron. 2002 Jul 3;35(1):121-33.

1238   48. Croft DR, Crighton D, Samuel MS, Lourenco FC, Munro J, Wood J, Bensaad K,
1239       Vousden KH, Sansom OJ, Ryan KM, Olson MF. p53-mediated transcriptional regulation
1240       and activation of the actin cytoskeleton regulatory RhoC to LIMK2 signaling pathway
1241       promotes cell survival. Cell research. 2011 Apr;21(4):666-82.

1242   49. Dan S, Tsunoda T, Kitahara O, Yanagawa R, Zembutsu H, Katagiri T, Yamazaki K,
1243       Nakamura Y, Yamori T. An integrated database of chemosensitivity to 55 anticancer
1244       drugs and gene expression profiles of 39 human cancer cell lines. Cancer Research.
1245       2002 Feb 15;62(4):1139-47.

1246   50. Po'Uha ST, Shum MS, Goebel A, Bernard O, Kavallaris M. LIM-kinase 2, a regulator of
1247       actin dynamics, is involved in mitotic spindle integrity and sensitivity to microtubule-
1248       destabilizing drugs. Oncogene. 2010 Jan;29(4):597-607.

1249   51. Gamell C, Schofield AV, Suryadinata R, Sarcevic B, Bernard O. LIMK2 mediates
1250       resistance to chemotherapeutic drugs in neuroblastoma cells through regulation of drug-
1251       induced cell cycle arrest. PLoS One. 2013 Aug 21;8(8):e72850.

1252   52. Llovet JM, Fuster J, Bruix J. The Barcelona approach: diagnosis, staging, and treatment
1253       of hepatocellular carcinoma. Liver transplantation. 2004 Feb;10(S2):S115-20.

1254    53. Kahraman DC, Hanquet G, Jeanmart L, Lanners S, Šramel P, Boháč A, Cetin-Atalay R.
1255        Quinoides and VEGFR2 TKIs influence the fate of hepatocellular carcinoma and its
1256        cancer stem cells. MedChemComm. 2017;8(1):81-7.

1257    54. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool
1258        K, Bates R, Žídek A, Potapenko A, Bridgland A. Highly accurate protein structure
1259        prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-9.

1260    55. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein
1261        structure prediction using predicted interresidue orientations. Proceedings of the
1262        National Academy of Sciences. 2020 Jan 21;117(3):1496-503.

1263    56. Bateman A, Coggill P, Finn RD. DUFs: families in search of function. Acta
1264        Crystallographica Section F: Structural Biology and Crystallization Communications.
1265        2010 Oct 1;66(10):1148-52.

1266    57. Doğan T, Karaçalı B. Automatic identification of highly conserved family regions and
1267        relationships in genome wide datasets including remote protein sequences. PloS one.
1268        2013 Sep 12;8(9):e75458.

1269    58. Doğan T, Atas H, Joshi V, Atakan A, Rifaioglu AS, Nalbat E, Nightingale A, Saidi R,
1270        Volynkin V, Zellner H, Cetin-Atalay R. CROssBAR: comprehensive resource of
1271        biomedical relations with knowledge graph representations. Nucleic Acids Research.
1272        2021 Sep 20;49(16):e96-.

1273    59. Sun J, Jeliazkova N, Chupakhin V, Golib-Dzib JF, Engkvist O, Carlsson L, Wegner J,
1274        Ceulemans H, Georgiev I, Jeliazkov V, Kochev N. ExCAPE-DB: an integrated large
1275        scale dataset facilitating Big Data analysis in chemogenomics. Journal of
1276        cheminformatics. 2017 Dec;9(1):1-9.

1277    60. Weininger D. SMILES, a chemical language and information system. 1. Introduction to
1278        methodology and encoding rules. Journal of chemical information and computer
1279        sciences. 1988 Feb 1;28(1):31-6.

1280    61. Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information
1281        and modeling. 2010 May 24;50(5):742-54.

1282    62. Landrum G. RDKit: Open-source cheminformatics.

1283    63. Dalke A. The chemfp project. Journal of Cheminformatics. 2019 Dec;11(1):1-21.

1284    64. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness,
1285        markedness and correlation. arXiv preprint arXiv:2010.16061. 2020 Oct 11.

1286    65. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry:
1287        miniperspective. Journal of medicinal chemistry. 2014 Apr 24;57(8):3186-204.

1288    66. Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, Martin MJ.
1289        UniProt-DAAC: domain architecture alignment and classification, a new method for
1290        automatic functional annotation in UniProtKB. Bioinformatics. 2016 Aug 1;32(15):2264-
1291        71.

1292    67. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C,
1293        Dalenberg K, Duarte JM, Dutta S, Feng Z. RCSB Protein Data Bank: biological
1294        macromolecular structures enabling research and education in fundamental biology,
1295        biomedicine, biotechnology and energy. Nucleic acids research. 2019 Jan
1296        8;47(D1):D464-74.

1297    68. Sterling T, Irwin JJ. ZINC 15–ligand discovery for everyone. Journal of chemical
1298        information and modeling. 2015 Nov 23;55(11):2324-37.

1299    69. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open
1300        Babel: An open chemical toolbox. Journal of cheminformatics. 2011 Dec;3(1):1-4.

1301    70. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ.
1302        AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.
1303        Journal of computational chemistry. 2009 Dec;30(16):2785-91.

1304    71. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural
1305        summaries of PDB entries. Protein science. 2018 Jan;27(1):129-34.

1306    72. Labbé CM, Rey J, Lagorce D, Vavruša M, Becot J, Sperandio O, Villoutreix BO, Tufféry
1307        P, Miteva MA. MTiOpenScreen: a web server for structure-based virtual screening.
1308        Nucleic acids research. 2015 Jul 1;43(W1):W448-54.

1309    73. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web
1310        service based on EADock DSS. Nucleic acids research. 2011 May
1311        28;39(suppl_2):W270-7.
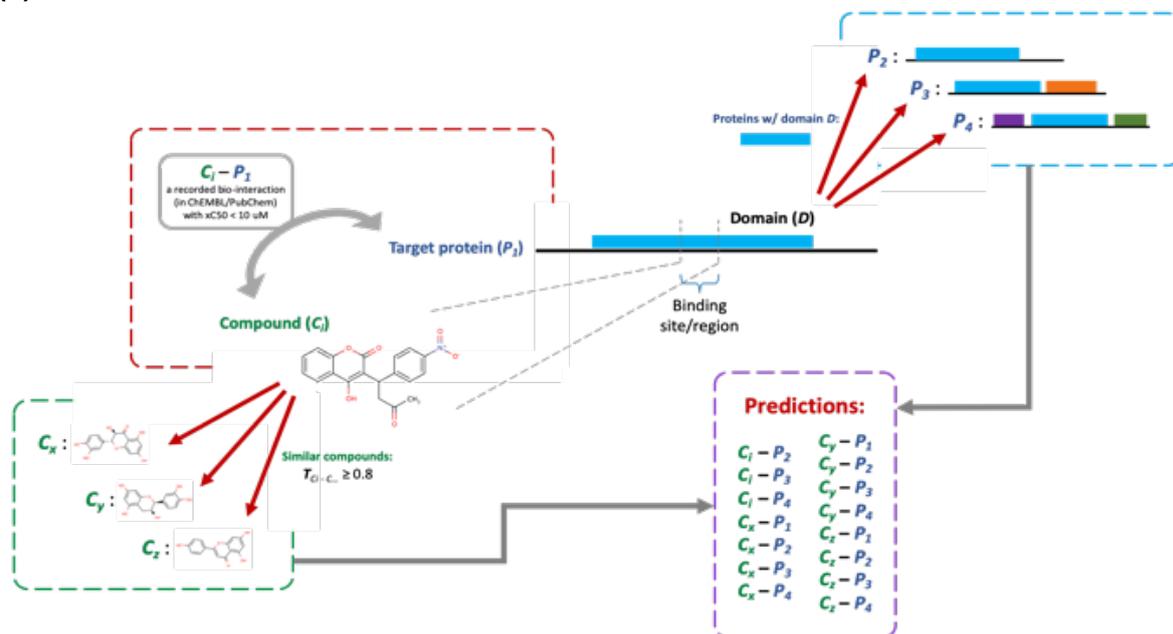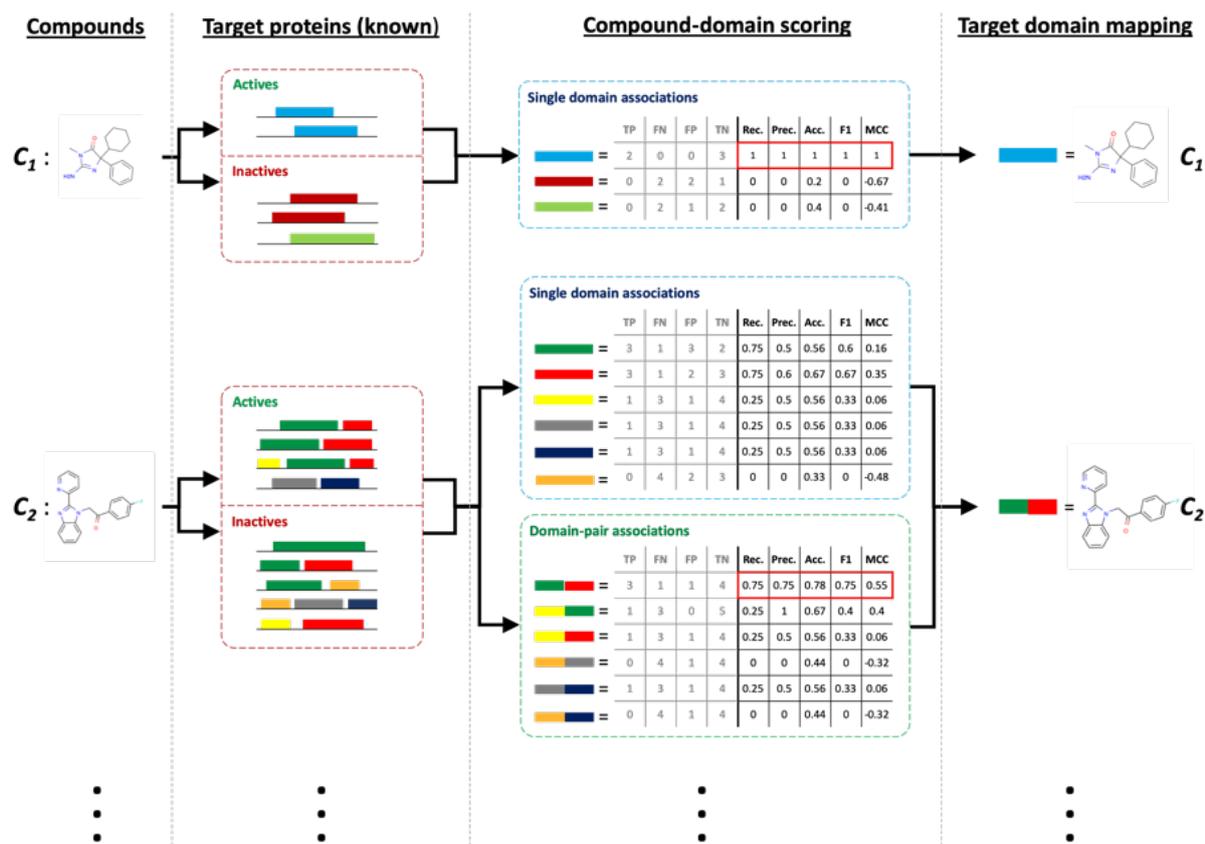
**Figures**

**(a)**

**(b)**

**Figure 1. (a)** The overall representation of the drug/compound – target protein interaction
prediction approach used in DRUIDom; and **(b)** drug/compound – domain mapping
procedure and its scoring over two representative ($c_1$, $c_2$) toy examples.

52

**LIMKi-1**
C17H19FN4O
CAS: 891397-98-1
MW: 314.1543
CHEMBL1316589 / CID-16014597 /
ZINC6767435
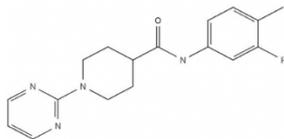
**LIMKi-1a**
C17H20N4O
CAS: 943094-41-5
MW: 296.1637
- / CID-43815770 / ZINC35290286
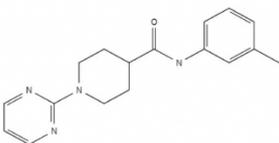
**LIMKi-2**
C20H21N5OS
CAS: 887621-34-3
MW: 379.1467
CHEMBL518653 / CID-15978868 /
ZINC34836571
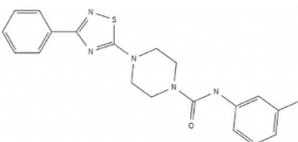
**LIMKi-3**
C20H21N5O2S
CAS: 887621-30-9
MW: 395.1416
CHEMBL516650 / CID-15978993 /
ZINC34836901
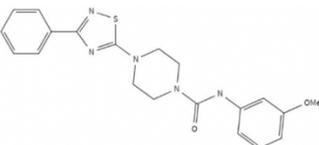
**LIMKi-2a**
C20H21N5OS
CAS: -
MW: 379.1467
- / - / -
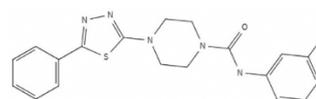
**LIMKi-2b**
C20H21N5OS
CAS: -
MW: 379.1467
- / - / -
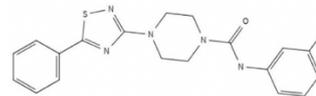
**LIMKi-2c**
C20H20BrN5OS
CAS: -
MW: 457.0572
- / - / -
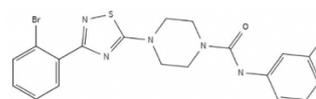
**LIMKi-2d**
C20H20BrN5OS
CAS: -
MW: 457.0572
- / - / -
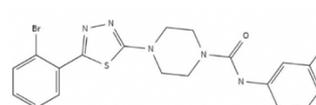
1320

1321 **Figure 2.** Structures, database identifiers, and 2-D representations of predicted LIMK
1322 inhibitory compounds (LIMKi-1, 1a, 2, and 3) and derivatives (LIMKi-2a, b, c, and d).
1323
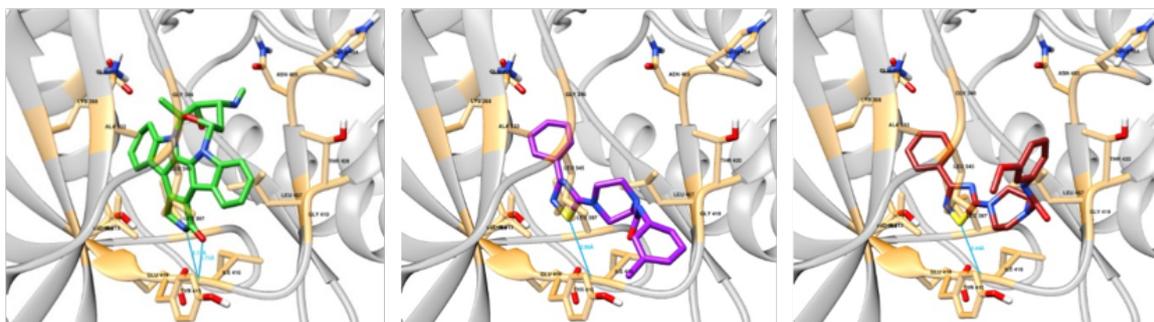
**(a)**

**(b)**
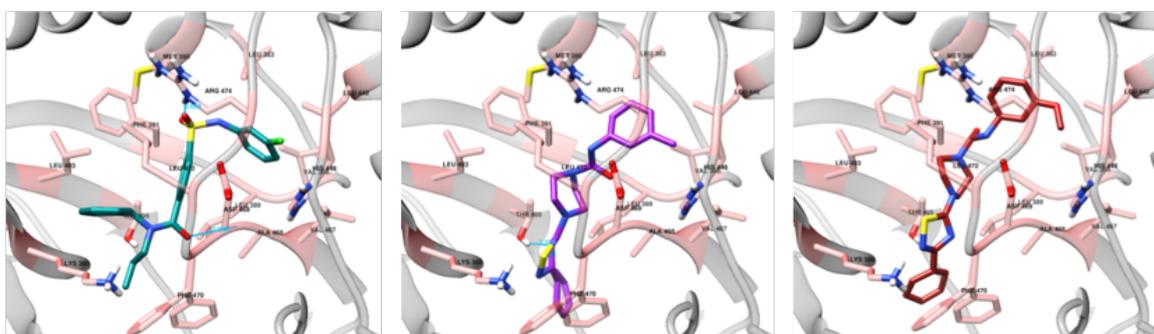
**Figure 3.** Visualization of the docked complex structures of **(a)** LIMK1 kinase domain in complex with the reference molecule staurosporine (green), LIMKi-2 (violet), and LIMKi-3 (red), and **(b)** LIMK2 kinase domain in complex with the reference molecule 9D8 (dark cyan), LIMKi-2 (violet), and 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.

1335  **(a)**
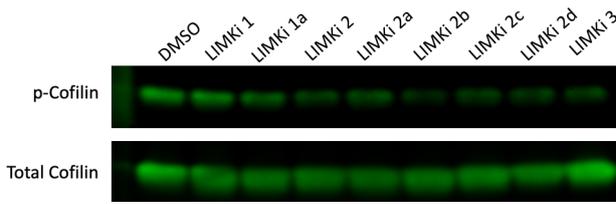


1336
1337
1338
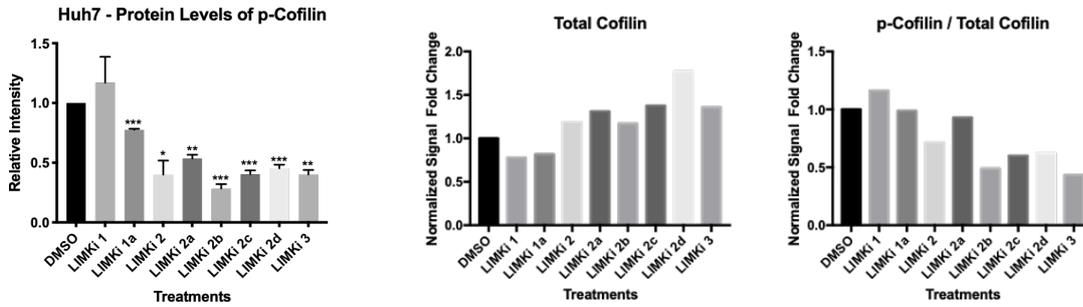


1339

1340  **(b)**
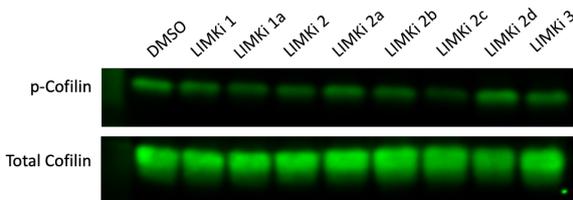


1341
1342
1343
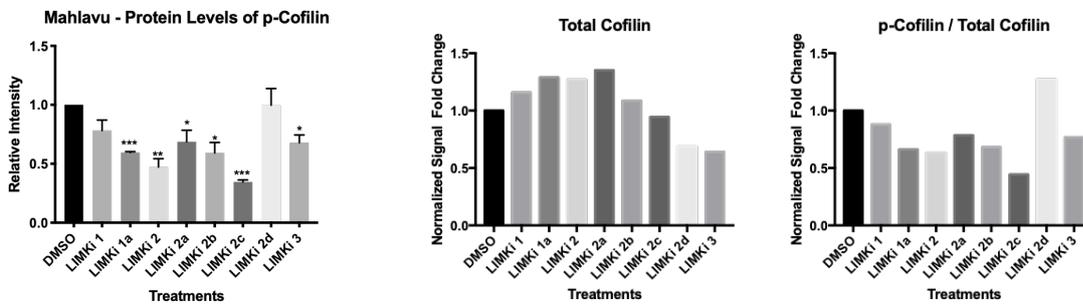


1344

1345  **Figure 4**. Phospho-Cofilin protein expression; **(a)** Huh7 and **(b)** Mahlavu cells were cultured
1346  with LIMK inhibitors (20 μM) for 48 hours and expression of active p-Cofilin and total Cofilin
1347  levels were assessed with western blot analysis.  Bar graph indicates the relative intensity of
1348  p-Cofilin levels compared to untreated DMSO controls. The equal loading control was
1349  analyzed based on the total protein staining normalization protocol. The ratios of phospho-
1350  and total Cofilin levels for both Mahlavu and Huh7 cell lines were calculated.
1351

1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
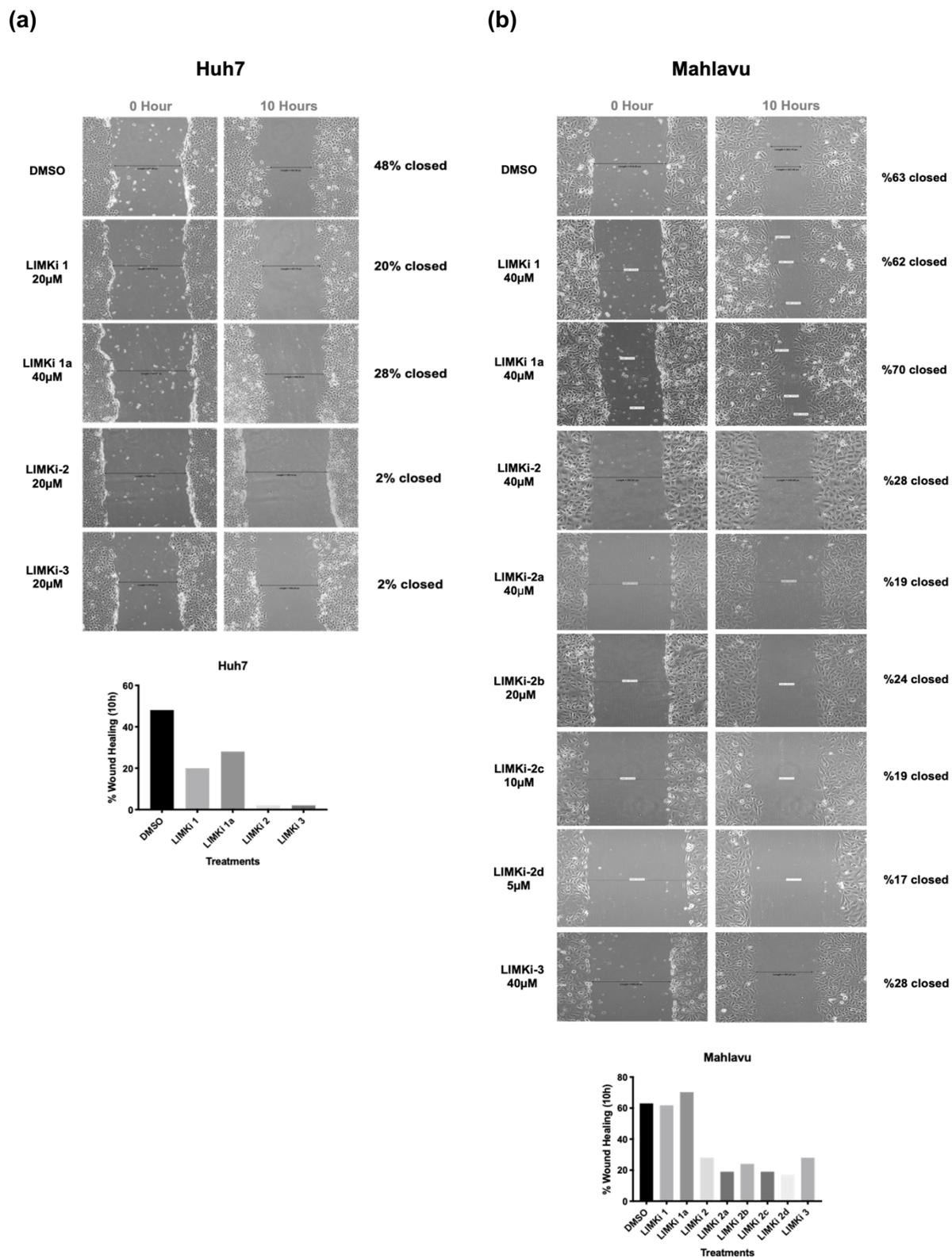
**(a)**                                                    **(b)**



**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.

**(a)**



**(b)**
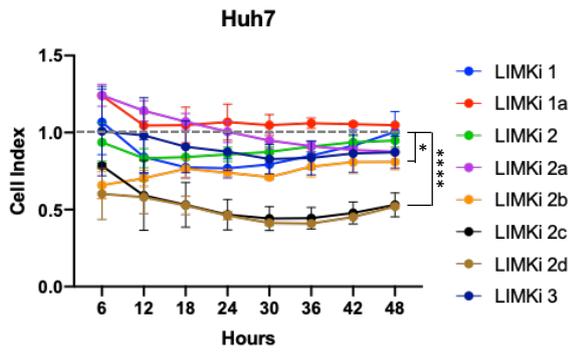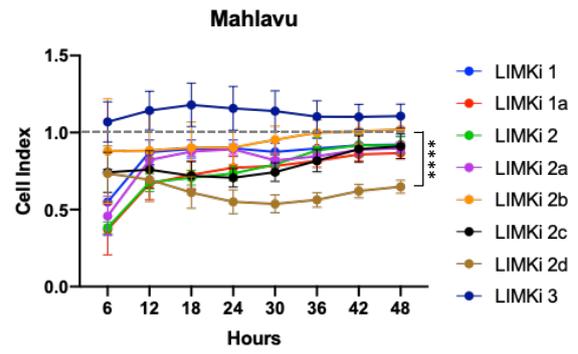
1391

**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 $\mu$M 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).

1399
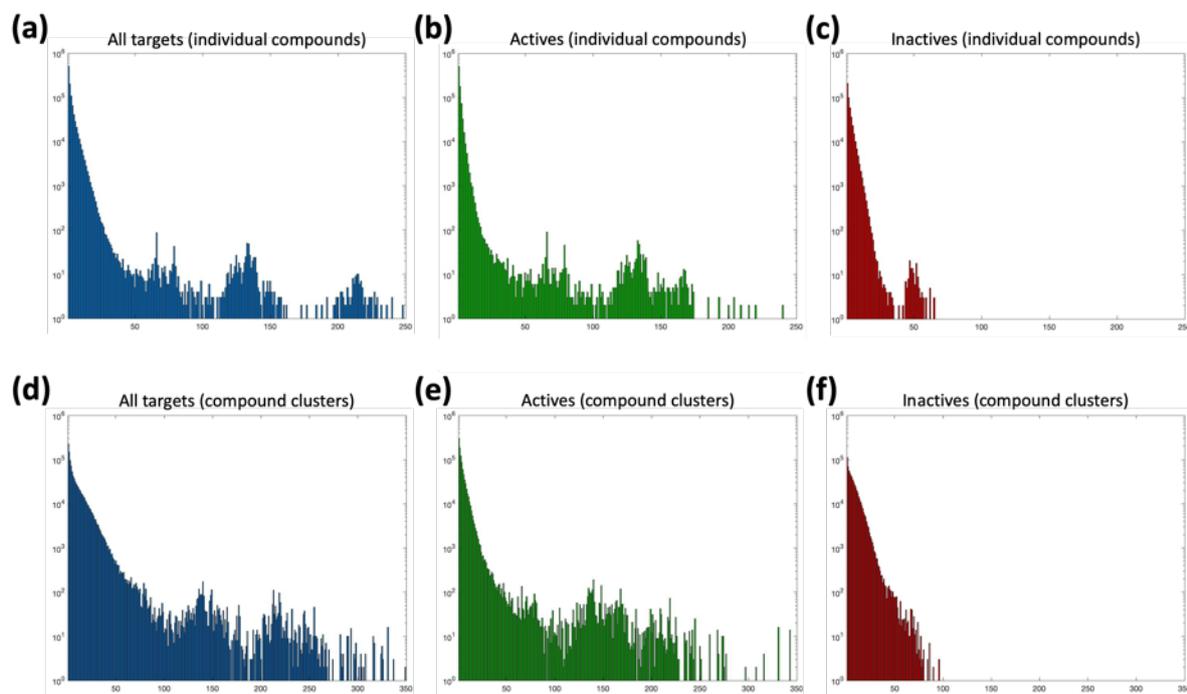


1400

1401

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.