Target-focused library design by pocket-applied computer vision and fragment deep generative linking

Merveille Eguida, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa, Didier Rognan

Target-focused library design by pocket-applied computer vision and fragment deep generative linking

Merveille Eguida, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa, Didier Rognan

Target-focused library design by pocket-applied computer vision and fragment deep generative linking.

Merveille Eguida,[†] Christel Schmitt-Valencia,[≠] Marcel Hibert,[†] Pascal Villa,[≠] and Didier Rognan.[†]*

[†]Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS-Université de Strasbourg, F-67400 Illkirch

[≠]Plateforme de Chimie Biologique Intégrative de Strasbourg, UAR3286 CNRS-Université de Strasbourg, Institut du Médicament de Strasbourg, F-67400 Illkirch

* To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

Abstract

We here describe a computational approach (POEM: Pocket Oriented Elaboration of Molecules) to drive the generation of target-focused libraries while taking advantage of all publicly available structural information on protein–ligand complexes. A collection of 31 384 PDB-derived images with key shapes and pharmacophoric properties, describing fragment-bound microenvironments, is first aligned to the query target cavity by a computer vision method. The fragments of the most similar PDB subpockets are then directly positioned in the query cavity using the corresponding image transformation matrices. Lastly, suitable connectable atoms of oriented fragment pairs are linked by a deep generative model to yield fully connected molecules. POEM was applied to generate a library of 1.5 million potential cyclin-dependent kinase 8 inhibitors. By synthesizing and testing as few as 43 compounds, a few nanomolar inhibitors were quickly obtained with limited resources in just two iterative cycles.

INTRODUCTION

Fragment-based drug design (FBDD)[1] has gain considerable popularity in the last 20 years for identifying new lead compounds and guiding the optimization towards drug candidates, even up to the market with four recently approved drugs.[2] Common FBDD programs starts by screening libraries of low molecular weight compound (fragments)[3] by multiple biophysical methods such as nuclear magnetic resonance spectroscopy (NMR), surface plasmon resonance (SPR), isothermal titration calorimetry (ITC) or mass spectroscopy (MS) to cite just a few.[4] Key advantages of FBDD with respect to biochemical high-throughput screening (HTS) are the sampling of a much larger chemical space as well as higher hit rates, even for difficult targets for which other approaches failed. Despite low affinities, fragment hits can be progressed to leads by linking, merging or growing approaches.[5] Although not necessary, it is usually advisable to start from high quality X-ray diffraction data to position fragment hits in their cognate target.[6] Even if FDBB is now widely used for hit identification, not all targets and fragments are suitable to X-ray diffraction. One the one hand, some targets still proved to be hard to isolate, purify in large scale and produce high-quality crystals for X-ray diffraction. On the other hand, some fragments cannot be detected by the latter technique because of poor physicochemical properties or too low affinities. In such cases, computational approaches are the only alternatives to predict the most viable positions of fragment hits identified experimentally[7] or to identify new hits by *in silico* screening.[8]

Three computational approaches can be used to predict the relative orientation of a fragment in a target cavity: molecular docking, functional group mapping and deconstruction-reconstruction. Molecular docking[9] is by far the most popular structure-based approach and aims at identifying both the bound conformation and the orientation of the ligand in a target cavity from their respective stereochemical and topological complementarities. Although it has mostly been applied to drug-like compounds, docking can be used to pose fragments with an accuracy comparable to that of lead-like compounds.[10-11] Docking is the computational method that is the closest to experimental fragment screening, and can be directly applied to any fragment library. In addition to potential hit identification,

the fragment position in the target cavity is also predicted. Unfortunately, scoring weak-binding fragments remains a challenge and requires an efficient post-processing, e.g. knowledge-based protein-ligand interaction rescoring.[12-14]

Functional group mapping[15] uses probe atoms or groups to map a protein cavity at their preferential location. Probes can be positioned according to protein-ligand interaction energies at regular points of a three-dimensional (3D) lattice[16-17] or by molecular dynamics (MD) sampling.[18] Interestingly, exhaustive all-atom MD better captures protein flexibility and solvation issues, and may also unmask transient cavities hidden to conventional docking protocol. Key drawback is the computational burden limiting a wide applicability for virtual screening. Moreover, reconstructing a fully connected ligand from several discontinuous propensity maps is not straightforward.

Last, deconstruction-reconstruction approaches[19] aim at computationally splitting protein-bound ligand X-ray structures into fragments according to well-known retrosynthetic organic chemistry rules.[20-21] Resulting fragments can then be recombined into new chemical entities while taking into account the protein environment. The method still suffers from the tricky recombination step (linking, merging, scaffold hopping)[22] that may disturb the original fragment binding modes or generate conformational strains. Interestingly, deep generative models[23-25] for linking disconnected fragments have shown some promises as they learn from millions of existing bioactive ligands. Deconstruction-reconstruction is mainly target-specific and applicable to targets for which numerous co-crystallized ligands are already available, although docking poses may be used in principle.

None of the above-reported method really takes profit of the increasing amount of structural data on protein-ligand complexes and their druggable pockets.[26] Since low molecular weight fragments have been shown to bind to preferential protein microenvironments regardless of their evolutionary relationship,[27] a FBDD approach considering the whole universe of druggable ligands and pockets is desired. Capitalizing on our recent numerical image processing tool to describe and align protein cavities,[28] we here propose to pose fragments according to the local similarity of their respective

subpockets to the target cavity. Applying the transformation matrix leading to the optimal subpocket-cavity alignment, the corresponding fragments are directly positioned into the target cavity and connected, under topological constraints, by a deep generative linker to yield fully connected molecules. Applying the method to the catalytic site of human cyclin dependent kinase 8 (CDK8), a focused library of 1.5 million chemical entities could be quickly generated. Interestingly, most newly generated compounds exhibited unprecedented structures. *In vitro* biological evaluation of 43 carefully selected compounds identified several nanomolar inhibitors within just two design iterations and limited experimental efforts.

RESULTS AND DISCUSSION

Setting the scene. We herein present a novel method to design target cavity-focused libraries based on predicted similarities between the target cavity and a library of PDB fragment-bound subpockets (**Figure 1**). The underlying idea is to locate the most complementary fragments in the target cavity based on the estimated similarity of their corresponding subpockets, and then to link the prepositioned fragments into drug-like compounds using a deep generative linker. Accordingly, this approach can be implemented even in the absence of known ligands for the target protein. To assess its applicability and limits in a real-life drug design project, the method is here applied to CDK8, a target of pharmaceutical interest[29] and known X-ray structure.[30] In the following sections, we will describe, step by step, each part of the workflow until the experimental validation of newly generated inhibitors.

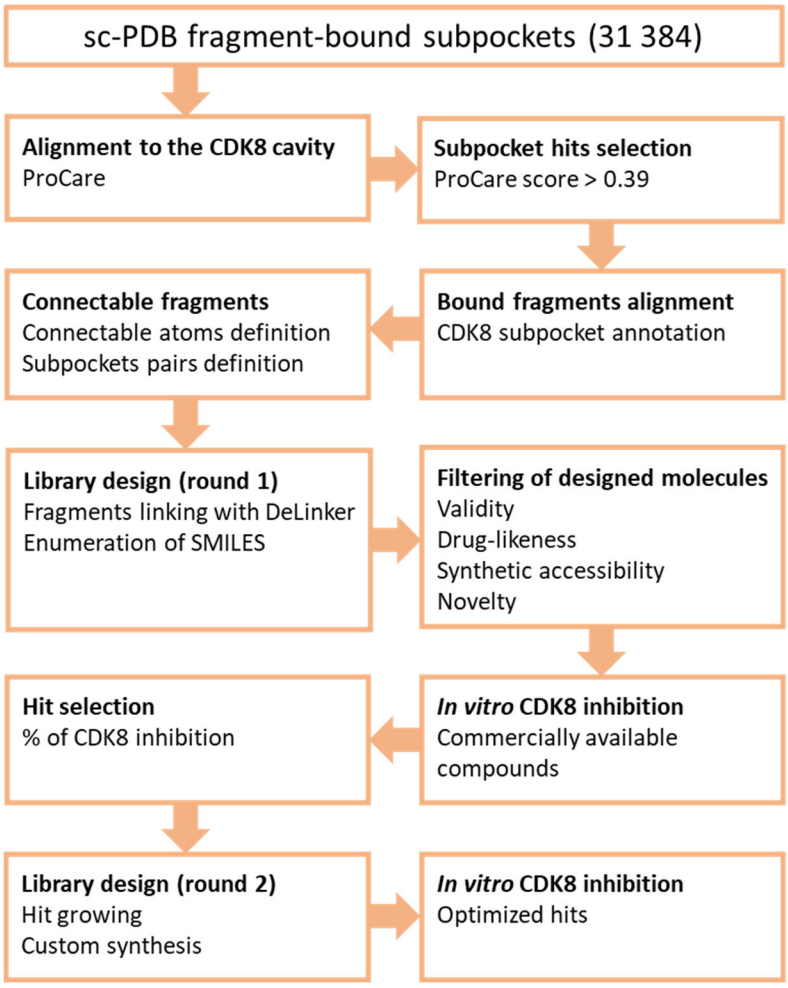


Figure 1. Overall workflow of the POEM computational method including in vitro experimental validation.

Alignment of fragments to the target cavity. Subpockets, defined as the immediate protein environment around bound fragments of druggable protein-ligand complexes (sc-PDB dataset),[31] were compared and aligned to the ATP pocket of CDK8 with the aim to use the hidden bound fragments for library design. The rationale of this implementation is that according to the similarity principle, fragments originating from similar subpockets are likely to reproduce favorable interactions with the target pocket. The term 'fragment' here refers to the molecular moieties obtained after interaction-aware 3D fragmentation of ligands bound to proteins so that each fragment exhibits at least one polar interaction and at least four interactions with its target.[32] The query CDK8 pocket and the sc-PDB subpockets are represented as a cloud of 1.5 Å-spaced points annotated by eight pharmacophoric properties (hydrophobic, aromatic, H-bond acceptor, H-bond donor, H-bond acceptor and donor, positive ionizable, negative ionizable, null).[33] The term 'pocket' describes the full druggable cavity available at the surface of the protein while a subpocket is defined from its bound fragment. Since we aimed at targeting the ATP binding site in its type-I 'DMG in' conformation, the druggable pockets were first detected from 19 available CDK8 structures (**Table S1**). The largest pocket (830.3 $Å^3$) selected as representative was retrieved from the 5HBH[30] PDB entry (**Figure 2**). This pocket incorporates regions around the hinge, the gatekeeper F97, extends to a solvent exposed area near the αD helix whereas on an opposite side. It covers the DMG motif and reaches the αC-helix (**Figure 2A**). It thus spans several already described kinase subpockets: the adenine pocket, the front pockets FP-I and FP-II, the back pockets BP-I-A and BP-I-B in the gate area.[34] The 31 384 sc-PDB subpockets were compared and aligned to the CDK8 cavity with the in-house ProCare method (**Figure S1**).[28] Briefly, ProCare finds the best possible local alignment of cavity-defining points using a point cloud registration algorithm[35-36] and scores the alignment according to the overlap of pharmacophoric properties of the aligned points. According to a preliminary study on the set of CDK8 structures, the original ProCare alignment c-FPFH fingerprint was modified to account only for the spatial distribution of pharmacophoric features (**Figure S2-S3**), a modification leading to a better alignment of CDK8 subpockets and fragments to the

corresponding full cavities. Noteworthy, the novel c-FH fingerprint does not change the distribution of ProCare scores with respect to alignments generated by the original c-FPFH fingerprint.

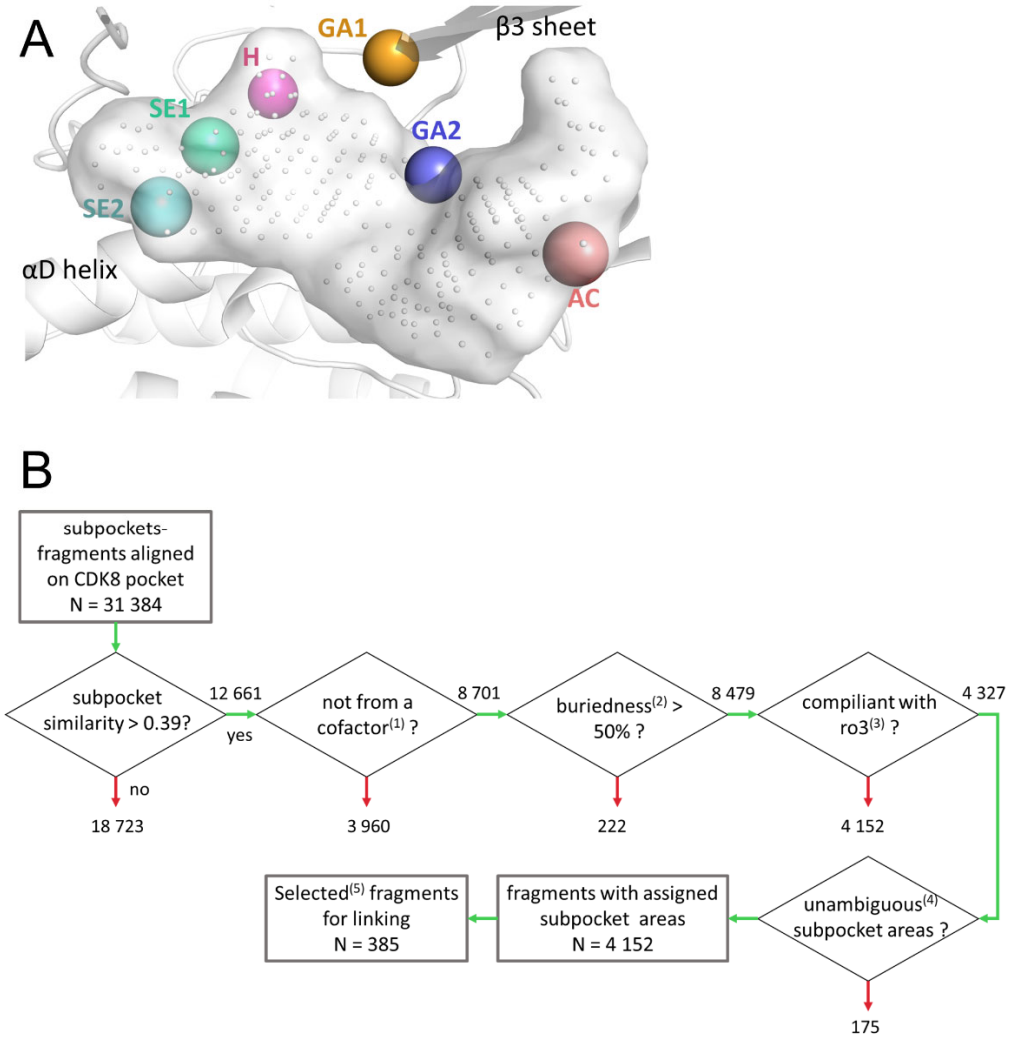


Figure 2. Seed fragments selection to fill the CDK8 query cavity. **A)** Description of the reference CDK8 pocket (PDB ID: 5HBH). Cavity points (grey dots, 246 points) delineate a ligand-accessible envelope (solid surface, 830.3 Å^3) and areas (hinge, H; gate area 1, GA1; gate area 2, GA2; solvent-exposed area 1, SE1; solvent-exposed area 2, SE2; αc area, AC) according to the distance to key CDK8 atoms (spheres). **B)** Fragments selection workflow. (1) A list of cofactors (PDB HET code) is provided in the sc-PDB database. (2) Fragments buriedness is approximated as the percentage of heavy atoms within 1.5 Å of one CDK8 cavity point. (3) fragment rule-of-three:[37] molecular weight ≤ 300 g.mol-1, logP ≤ 3, H-bond donor count ≤ 3 and H-bond acceptor count ≤ 3. (4) ambiguous annotation denotes assignment of two or more incompatible areas (Methods section) out of the six possible areas. (5) All annotated fragments from H, GA1, SE2 areas and a random sampling of 100 fragments from GA2 were selected.

Once transformation matrices of the alignment of sc-PDB subpockets to the target cavity were obtained, the same rotation/translation matrices were applied to the corresponding sc-PDB fragments to position them in the CDK8 cavity. Posed fragments were then filtered according to five criteria

(**Figure 2B**). Fragments originating from subpockets exhibiting a similarity score to the CDK8 pocket above a threshold value of 0.39 (previously shown to optimally discriminate known similar from known dissimilar binding sites)[28] were first selected, leading to a set of 12 661 fragments. Remaining fragments were further pruned according to three criteria: (i) belonging to a cofactor (therefore avoiding purine-base fragments), (ii) insufficient buriedness in the target cavity, (iii) no compliance to the fragment rule-of-three.[38] Remaining fragments were then annotated by one of the six CDK8 areas in which they were positioned: hinge (H), gate (GA1, GA2), solvent-accessible (SE1, SE2), αC helix (AC) (**Table 1**, **Figure 3**). 4 152 fragments could be unambiguously assigned to one CDK8 area: H (1.4%), GA1 (2.7%), GA2 (22.5%), SE1 (61.9%), SE2 (2.8%) and AC (8.7%) (**Figure 3A**).

Table 1. Annotation of the CDK8 target cavity by key pharmacophoric atoms.

Area	Label	Key CDK8 atoms	KLIFS subpockets[a]
Hinge area	H	Asp98.O, Ala100.N, Ala100.O	AP
Gate area 1	GA1	Phe97.CA (gatekeeper residue)	AP, BP-I-A, BP-I-B
Gate area 2	GA2	Lys52.NZ	AP, FP-I, FP-II
Solvent-accessible area 1	SE1	Arg366.CZ	-
Solvent-accessible area 2	SE2	His106.CE1	-
αC helix area	AC	Ser62.CA	-

[a] Full or partial overlap with KLIFS[34] subpockets: AP: adenine pocket, BP: back pocket, FP: front pocket

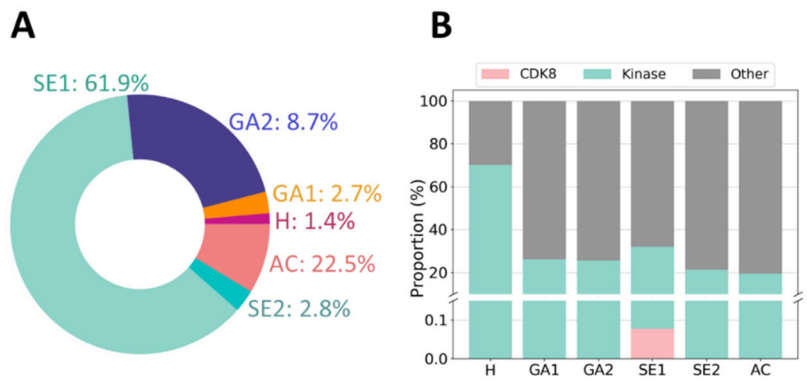


Figure 3. CDK8 subpocket occupancy of sc-PDB fragments. **A)** Assignment of CDK8 pocket areas to 4 152 sc-PDB fragments. **B)** Proportion of sc-PDB fragments per area.

We next analyzed the origin of the sc-PDB ligands these fragments were derived from. As to be expected, 70% of fragments assigned to the hinge area (H) come from protein kinase inhibitors, the remaining 30% originating from a ligand co-crystallized with a protein that belong to a non-kinase family (**Figure 3B**). However, it should be noted that fragments from known CDK8 inhibitors were not selected as occupying the hinge region. Two simple reasons explain this absence: (i) the seven CDK8 ligands in the sc-PDB dataset are type II inhibitors binding to a DMG-out conformation and occupy the back pocket, (ii) the only CDK8 ligand (3RGF) that binds to the hinge could not be fragmented by our protocol and therefore did not pass our filters. The other areas (GA1, GA2, SE1, SE2, AC) were assigned fragments from both kinase (~25%) and non-kinase ligands (~75%). While the initial sc-PDB subpocket database contains 16% of entries from protein kinases, the enrichment observed for hinge-selected fragments (4.4) is logically due to the specific stereoelectronic features of the hinge area, notably the hydrogen bonding capacity of Asp98 and Ala100 backbone heteroatoms imposing complementary features on the ligand side. To limit the size of the library, all fragments were not considered for full enumeration of complete molecules. Whereas all fragments bound to H (n=57), GA1 (n=111) and SE2 (n=117) subpockets were selected, only 100 GA2-bound fragments were randomly chosen. Duplicates, in other words 2D identical fragments were kept as they do not originate from the same 3D subpocket, therefore resulted in different alignments that may differently impact molecules design. Comprehensive statistics of the pairwise fragment similarity (**Figure S4**) and the observed distribution of their physicochemical properties (**Figure S5**) clearly evidence their chemical diversity. 385 fragments were selected at this stage for the next linking stage.

Round-1 library generation. The DeLinker deep generative model[23] was used to link the above-selected fragments. Briefly, DeLinker uses a graph-based deep generative model, trained on the ZINC[38]

or PDBbind[39] databases, to expand bond by bond the two fragments to be connected until final SMILES strings are generated by a variational autoencoder while keeping 3D constraints through a set of distances and angles between connectable atoms.[23] In the current work, all possible connectable atoms of hinge-annotated fragments (H) were used as seeds to find potential connectable atoms in fragments filling three remaining subpockets (GA1, GA2, SE2) (**Figure S6**).

An atom is considered connectable if it is a heavy atom covalently bonded to a hydrogen, that bond being used as exit vector for the linking. Pairs of atoms belonging to different fragments are then associated by restricting the angle between the exit vectors and distances between the corresponding heavy atoms (see Methods) in order to avoid pointless connections and lower the number of combinations (**Figure S7**). Starting from 385 fragments, 1 517 488 SMILES strings were generated by linking fragment pairs with DeLinker. 15% of the proposed solutions were discarded since they correspond to uncomplete molecules where the SMILES consisted of a linker moiety attached to only one of the two fragments (**Figure 4**).

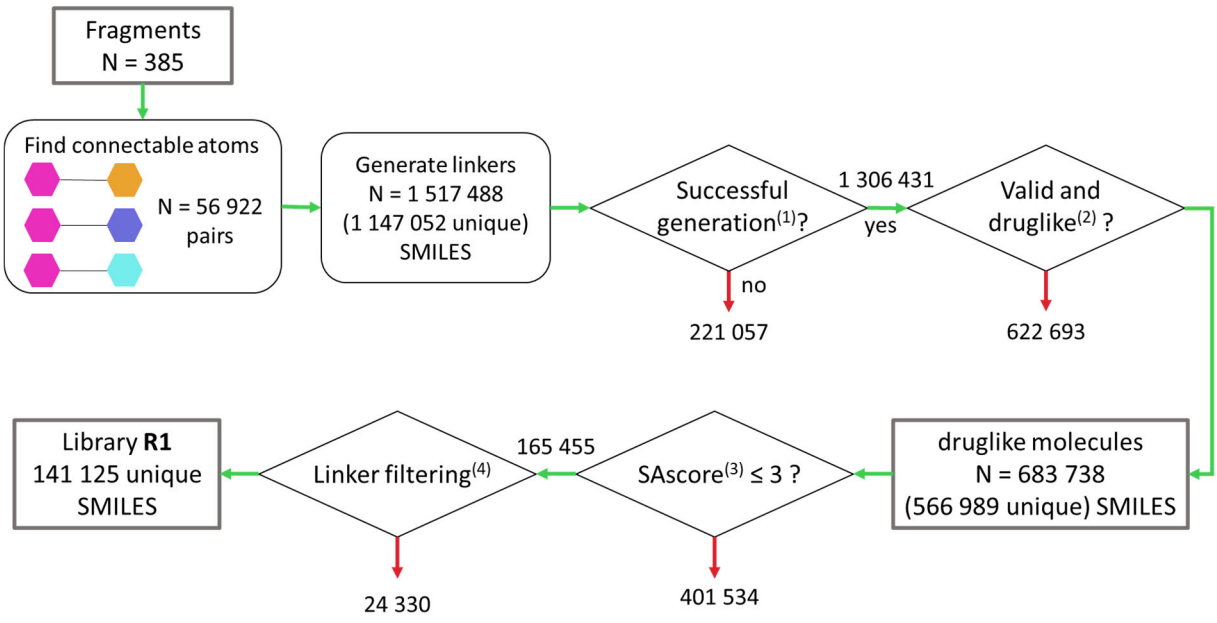


Figure 4. Focused library design via linking selected fragments. Fragments aligned in the H area were paired with fragments from GA1, GA2 and SE2 areas. SMILES were generated by linking fragment pairs with DeLinker[23] and filtered to compose the first-round library R1. (1) Successful linking signifies that both fragments have been attached to the linker whereas cases where only one of the fragments was linked were considered unsuccessful. (2) Druglikeness is defined by customized OpenEye Filter rules available in **Table S2**. (3) Synthetic accessibility score.[40] (4) Filter to remove unwanted aliphatic linkers.

The remaining molecules were filtered for drug-likeness (**Table S2**) resulting in 566 989 unique SMILES. Although the redundant SMILES per pair of connectable atoms were removed during the linking process, duplicated molecules still arose when connecting the same 3D fragments via equivalent exit atoms (symmetry cases) or connecting the same duplicated fragments originating from different subpockets. After keeping only molecules that are likely to be synthesized (SAscore[40] ≤ 3), only those having a linker compliant with defined rules (**Figure S8**) were finally kept. The remaining 141 125 molecules composed the first-round R1 library (**Figure 4**). A majority of the generated molecules arose from combining the hinge and the solvent-exposed SE2 fragments which account for more than 50% of the sets (**Figure 5**).

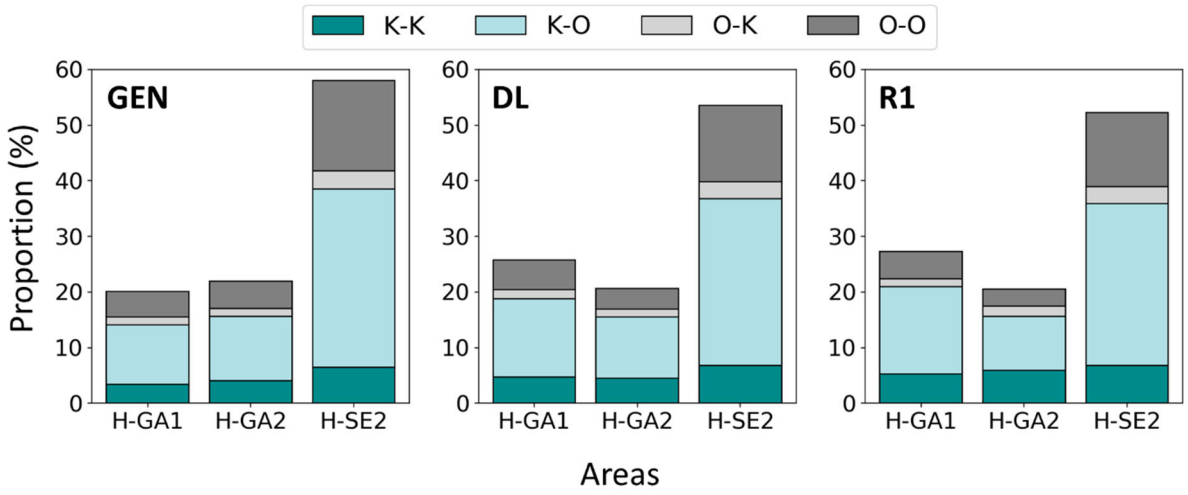


Figure 5. Protein origin of fragments pairs in newly generated molecules. From left to right, the full set after cleaning unsuccessful generation out (GEN), the drug-like subset (DL) and round-1 library (R1). The distribution is given for the combinations annotated by the targeted CDK8 area (H, hinge; GA1, gate area 1; GA2, gate area 2, SE2, solvent-exposed area 2) and color-coded according to the protein origin (co-crystallized target) of the two connected fragments (K, protein kinase; O, other; K-K, both fragments were derived from a protein kinase structure; K-O, H-fragment derived from a protein kinase and the other fragment from a non-kinase protein structure; O-K, H-fragment derived from a non-protein kinase and the other fragment from a kinase protein structure; O-O, both fragments were derived from a non-kinase protein structure).

Indeed, the average number of generated SMILES strings per pair of H-SE fragments is higher than for the two other areas, a consequence of having more pairs of connectable atoms and more generated linkers per connectable atoms for the H-SE subpockets. While it was expected that kinase-derived

fragments would contribute to most of the generated molecules, only 14% of SMILES strings were generated by linking two kinase-bound fragments. Interestingly, around 26% of the molecules were made of two fragments originating from a non-kinase protein. Interestingly, the observed proportions do not vary between the full set, the drug-like subset and the R1 set (**Figure 5**). Most of the generated molecules (> 90 %) were already compliant with the Lipinski's rule of five[41] (**Figure S9**). Albeit two fragments were assembled, many generated molecules still remained in the fragment space with around 10 % of SMILES strings being compliant with the fragment rule-of-three[37] (**Figure S9**). Filtering the designed molecules to R1 library members did not bias our selection towards molecules with particular properties as the distribution of the molecular properties, although reported individually, remained comparable among the sets (full, drug-like and R1; **Figure S9**). To give insights on the chemical space covered by R1 library members, we further assessed its overlap with either a broad purpose bioactive chemical space[41] (1.7 million ChEMBL compounds) or a recently described kinase-focused ligand space (6.7 million KinFragLib library members).[42] 259 unique R1 library molecules were exactly found in ChEMBL among which only a few have been assayed against protein kinases, while only five R1 library compounds were identical to KinFragLib molecules. Considering similarity, only 0.85% and 13% of R1 library members were found similar to KingFragLib and ChEMBL molecules, respectively, according to a Tanimoto coefficient, computed from Morgan2 fingerprints[43] higher than 0.60. The herein proposed computational workflow is therefore able to generate really new chemical entities, the chemical diversity of the generated molecules stemming from the diversity of the seed fragments pool, the connectivity and the possible linkers.

As a first validation of the structure-based workflow, we verified whether the drug-like subset contains molecules highly similar to 302 submicromolar human CDK8 inhibitors retrieved from the ChEMBL database. Using the similarity search protocol described in the methods section, we found 44 molecules that matched with 35 unique known CDK8 inhibitors (representing three series of congeneric molecules). While these molecules were built with fragments from all possible areas, most

of them were assembled from hinge-fragments originally co-crystallized with protein kinases, linked to fragments originally co-crystallized with non-kinase proteins.

The round-1 library contains novel and potent CDK8 inhibitors. To identify chemically novel hits, we filtered first-round R1 library members by dissimilarity (Tanimoto coefficient < 0.5, RDKit7 fingerprints) to all CDK8 compounds available in ChEMBL[41] and to all seed sc-PDB fragments. With respect to the previously used Morgan 2 fingerprint that is best suited to ligand-based virtual screening, the RDKit7 descriptor [44-45] was here chosen for its ability to account for substructure similarities between library members and known inhibitors. [44-45]

Hits were then searched for availability among 8.2 million commercially available drug-like compounds (**Table S3**) to select 37 compounds that are identical or very similar (Tanimoto coefficient > 0.90, RDKit7 fingerprints) to their queries (**Table S4**). These compounds were purchased and tested for CDK8 inhibition in a homogeneous time-resolved fluorescence (HTRF) assay aimed at measuring the FRET signal between a fluorescent-labelled ATP competitive inhibitor and the fluorescent-tagged CDK8 soluble kinase (see Methods). Six out of the 37 tested molecules (compounds **9**, **11**, **12**, **29**, **32**, **37**) inhibited the CDK8 kinase by more than 50% at the single concentration of 10 µM (**Figure 6**). Notably two related compounds **(12** and **37**), exhibiting more than 80% inhibition were assembled from the same pair of 3D fragments by just inverting the ester linkage (**Figure 6**). They differ from the original R1 library members by just a carbon atom (methoxy for ethoxy substitution, **Table S4**).

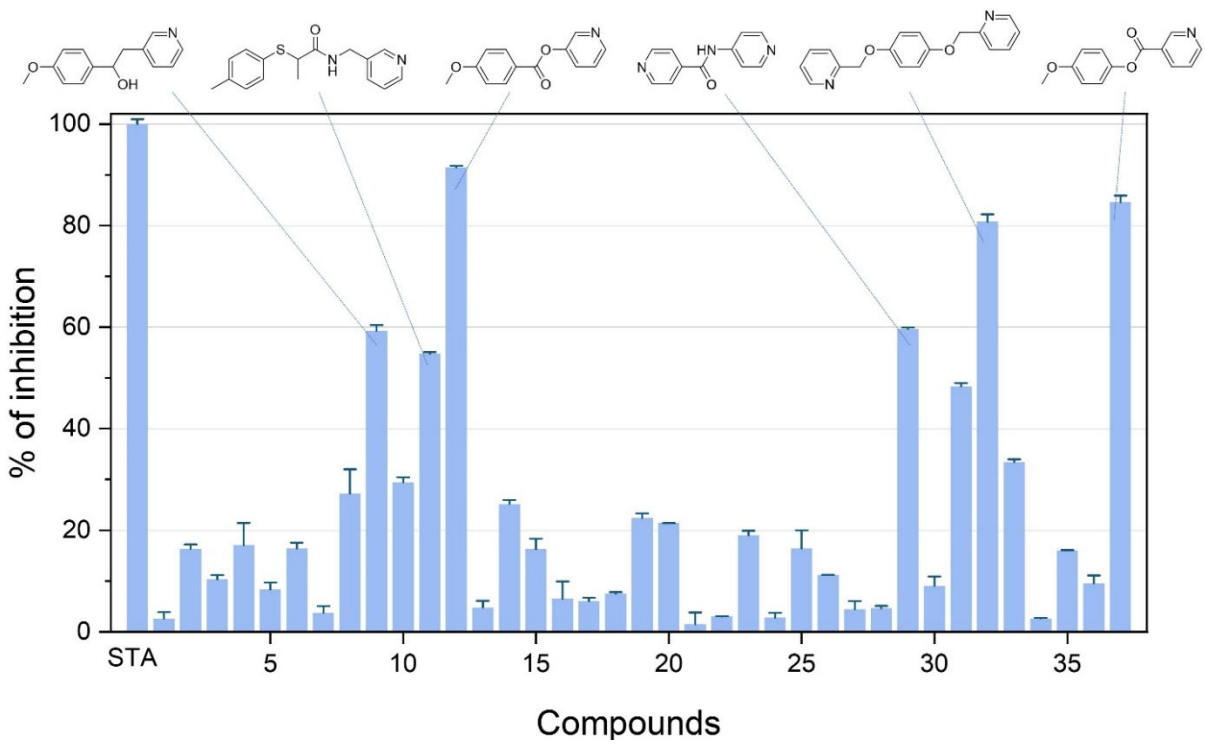
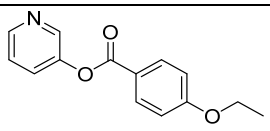
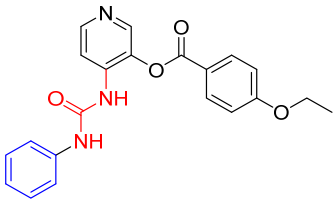
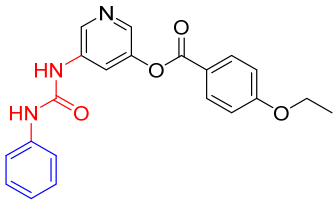
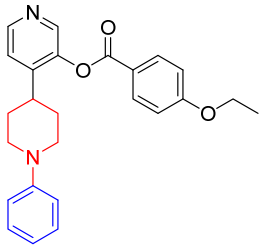
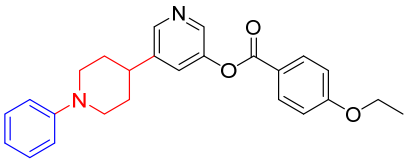
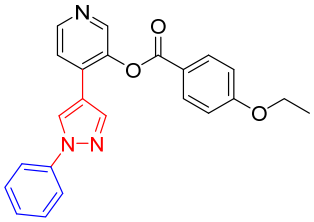
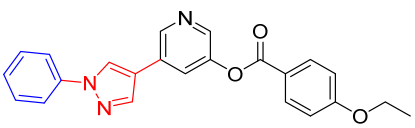


Figure 6. CDK8 inhibition (LanthaScreen Eu kinase competitive binding assay) by 37 commercially available compounds identical or very similar to R1 library members. Results are expressed as mean ± SEM of two independent experiments using a 10 µM concentration of competitor (STA, staurosporine control).

Round-2 library design by fragment hit growing. The most potent hit (**12**) from round-1 library , generated by linking a H-area pyridine fragment to a GA2-area p.methoxyphenyl fragment, is still a fragment-like compound (MW = 229 g.mol^{-1}) that can be optimized by growing towards the nearby and yet unexploited SE2 and GA1 subpockets. We thus explored the possible connections between the hinge-binding fragment of **12** and all remaining SE2 or GA1-anchored fragments, to generate a second-round library R2 of 5 700 compounds. R2 library members were filtered by physicochemical properties (number of rotatable bonds ≤ 6, no chiral centers) and synthetic accessibility (SAscore ≤ 3) to yield a final set of 151 candidates (**Table S5**). Six representative compounds (**Table 2**) were chosen for their ease of synthesis (i.e. availability of building blocks, costs of goods, number of synthetic steps) and predicted buriedness upon preliminary docking to CDK8. Three linkers (urea, piperidine, pyrazole) were chosen for their capacity to connect the H-anchoring pyridine ring to a SE2-anchored phenyl

fragment. Two positions of the pyridine ring (ortho and meta position to the benzoyl ester) were predicted compatible, therefore leading to six possible analogs (**Table 2**).

Table 2. Round-2 library of optimized hits and their CDK8 inhibitory potency.

Compound	Structure[a]	IC_{50}, nM[b]	CI 95%, nM[c]
12		376.9	245.2-579.5
39		354.6	203.4-618.0
41		>25 000	-
44		144.1	88.8-233.9
47		>25 000	-
49		6.4	4.57-8.95
51		> 25 000	-

[a] A phenyl moiety (blue) is attached via different linkers (red) to round-1 compound **12**. [b] Inhibition of CDK8 measured in a LanthaScreen Eu kinase competitive binding assay. Results are expressed as mean ± SEM of three independent experiments. [c] confidence interval at a 95% confidence level

The six compounds were synthesized (**Scheme S1**), checked for purity (**Figures S10-S15**) and tested for *in vitro* CDK8 inhibition using the same HTRF assay as described above, to build concentration-response curves (**Figure 7**).

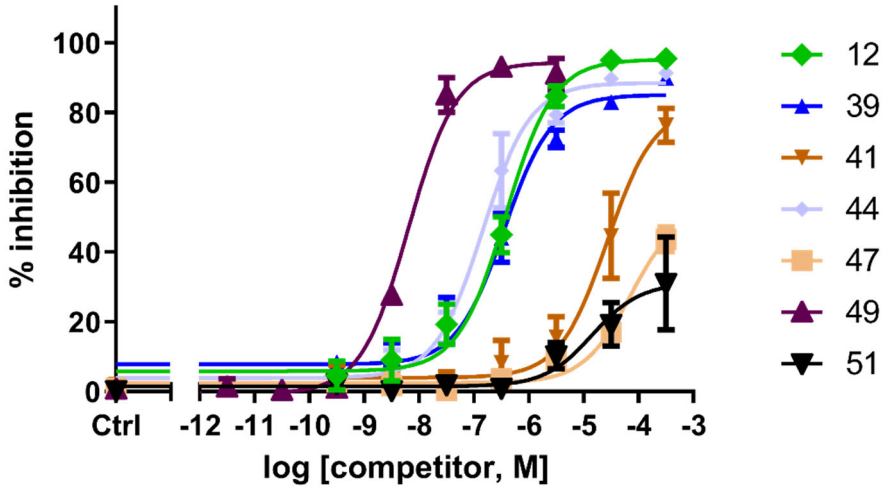


Figure 7. Inhibition of human CDK8 by six selected round-2 library compounds. Concentration-response curves are derived from three independent experiments with duplicates per experiment.

Out of the six round-2 library compounds, three molecules (**41**, **47**, **51**) are weak CDK8 inhibitors, one compound (**39**) is equipotent to the primary hit **12**, and two analogues (**44**, **49**) exhibit a higher potency than the parent compound **12** (**Table 2**, **Figure 6**). 3,4-disubstituted pyridines (**39**, **44**, **49**) were systematically more potent than their 3,5-disubstituted congeners (**41**, **47**, **51**). Noteworthy, the single-digit nanomolar inhibitor **49** could be obtained from scratch within just two design iterations and limited experimental efforts.

The ProCare pose of the three fragment-bound subpockets used to yield compound **49** (**Figure S16**) illustrate the alignment of key residues within each subpocket to the target cavity, and the capacity of ProCare to detect local similarities extending to 5-7 conserved amino acid pairs. As part as the fragment selection workflow, only fragments bound to subpockets whose pharmacophoric points share enough polar matches with that of the target query were considered for linking. In the present cases, several

polar matches (h-bond donor/acceptor, positive/negative ionizable, aromatic) could be detected between each of the three subpockets and the CDK8 cavity (**Figure S16**). The topological relationships (distance between connectable atoms, angles between exit vectors) between all selected fragments, notably that leading to the final hits **12** and **49** are given **in Figure S17**.

The~~ir~~ putative binding mode of compound **49**, deduced form molecular docking, suggests that the pyridine nitrogen atom h-bonds to the hinge backbone atoms (E98, A100) while the ethoxyphenyl and the newly introduced pyrazole moieties exhibit π- π interactions to H106 (SE2 subpocket) and the gatekeeper F97 (GA1 subpocket). Last, the terminal phenyl ring is oriented towards K52 (GA2 subpocket) for a putative π-cation interaction (**Figure 8**). While the parent hit **12** showed two possible docking poses (ethoxyphenyl towards GA2 or SE2), growing by a pyrazole prioritized the SE2 orientation, still with exhibited interactions compatible with the rationale of the initial fragment alignments.

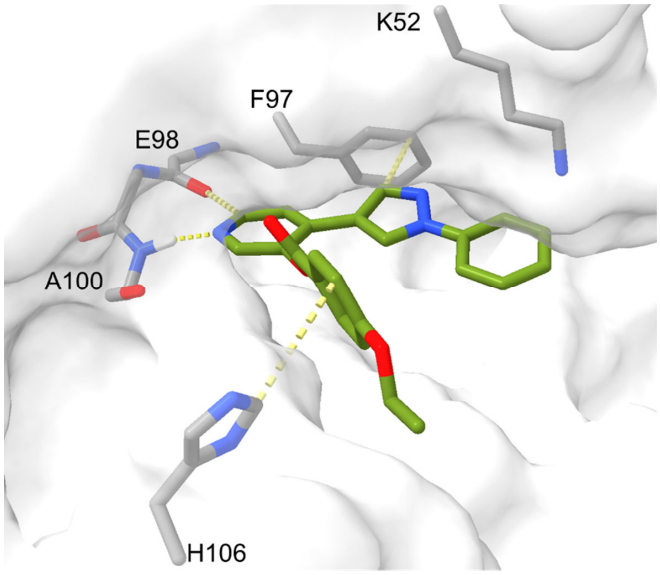


Figure 8. PLANTS docking pose of compound **49** (green sticks) to the catalytic site of CDK8 (PDB ID 5HBH, solid surface). H-bond to the hinge (E98, A100) and π- π interactions to F97, H106 are displayed by yellow broken bonds.

At this point, we should recall that neither early safety (e.g. kinase selectivity) nor pharmacokinetic properties (e.g. metabolic stability) have been considered in either generating or post-processing the

target-focused library members. Although technically feasible, target selectivity assessment requires applying the same workflow to different cavities and prioritizing compounds generated only for the target of interest. This approach is feasible for a comparing a few targets but is rapidly impracticable at a larger scale (e.g. whole kinome). It has not been applied in the current study aimed at demonstrating the proof-of-concept of the structure-based workflow.

CONCLUSION

We herewith propose a novel fragment-based library design method to generate target-focused compound libraries. The originality of the approach is that seed fragments are chosen from a large repertoire of protein-bound fragmented ligand X-ray structures, and positioned in the target according to the local similarity of their protein subpocket to the target cavity. This ligand-agnostic posing protocol does not require scoring protein-ligand interactions and is fuzzy enough to transfer ligand information across unrelated target spaces. Once fragments have been posed, they are linked by a deep generative model to enumerate full molecules which are later post-processed to account for drug-likeness and synthetic accessibility. The linking step still deserves improvement, notably to enumerate candidate molecules directly in the original target 3D coordinate frame. Hence, the variational autoencoder used here generates SMILES strings and just accounts for the target binding site topology in the form of topological relationships between fragment atoms to be connected. A true 3D deep generative model[46] considering complementarity to the binding site shape and the ligand conformational freedom would be highly desirable to link subpocket-selected seed fragments. It would

avoid a tedious post-processing of unrealistic solutions and the necessary docking of candidates to verify whether the starting binding hypothesis of the seed fragments is conserved.

When applied to the test case of the CDK8 kinase, the method was able to quickly suggest potential inhibitors. Within two iterations and 43 compounds, a single digit nanomolar inhibitor could be identified thereby demonstrating a first proof-of-concept of the underlying methodology. Interestingly, the POEM method is technically applicable to any target of known 3D structure and does not require prior ligand knowledge.

~~MATERIAL AND METHODS~~EXPERIMENTAL SECTION

CDK8 cavity detection. All publicly available X-ray structures of human CDK8 (UniProt accession number P49336; **Table S1**) were downloaded from the Protein Data Bank[47-48]. Type I structures (DMG-in, α-C helix-out) were put in the same coordinates frame by subsequent structural alignment to the 4F7S reference with Maestro v.2019-3 (Schrödinger, New York, NY 10036, U.S.A.) and refinement to ensure that the hinge residue Ala100 heavy atoms were fitted. Aligned structures (proteins, co-factors, ligands) were then protonated with Protoss v.4.0,[49] while optimizing the intra and inter-molecular hydrogen bond network. After discarding crystallization additives, each PDB entry was split to afford a protein (no water molecules) and a ligand in separate mol2 files using SYBYL-X 2.1.1 (Certara USA, Inc., Princeton, NJ 08540, U.S.A.). For each protein file, entire cavities ("*CAVITY_ALL*" output) were next computed with the VolSite[33] module of the IChem v.5.2.9 package,[50] using default parameters and saved as point clouds annotated by pharmacophoric features. Only cavities corresponding to the catalytic site were retained for the next steps. Upon visual inspection, the corresponding three clouds for PDB entry 5HBH were merged into a single cavity in mol2 file, yielding the reference pocket for CDK8.

sc-PDB subpocket-fragment database. 16 034 drug-like ligands in their protein-bound X-ray structure were retrieved from the sc-PDB database[31] of druggable protein-ligand complexes and fragmented in three dimensional (3D) space within their protein binding site using the IChem fragmentation tool.[32] Only fragments exhibiting at least 4 non-covalent interactions[12] (out of which one is polar, hydrogen-bond or electrostatic interaction) with the protein target were retained. The fragments exit bonds (dummy atoms 'Z') were converted into hydrogen atoms. The immediate protein environment of each selected fragment was considered to compute VolSite point clouds, keeping only those with at least 3 points, each being closer than 4.0 Å from any fragment heavy atom ("*CAVITY_4*" output), thereby defining a subpocket point cloud in mol2 file format for 31 384 fragments.

CDK8-focused library design. In the first stage, 31 384 sc-PDB subpocket point clouds (**Figure S1**) were aligned to the reference 5HBH CDK8 cavity point clouds with ProCare[28] v.0.1.1 using default parameters and the c-FH color-based descriptor (**Figure S2**) corresponding to the eight terminal bins of the c-FPFH descriptor.[28] For each subpocket-cavity pair, the optimal alignment matrix was used to position the corresponding sc-PDB fragment into the CDK8 cavity. The comparison protocol was validated by successful cross-comparison of CDK8 subpockets from type I PDB entries (**Figure S3**).

In the second stage, aligned sc-PDB fragments were filtered according to their subpocket similarity to the CDK8 cavity (ProCare score ≥ 0.39), their compliance to the fragment rule-of-three,[37] and their embedding into the CDK8 cavity such that at least half of the fragment atoms are less than 1.5 Å away to the closest CDK8 cavity point. Fragments originating from the sc-PDB list of cofactors were excluded. Resulting fragments were further annotated with the CDK8 cavity area to which they have been aligned based on their distance (closest heavy atom should be within 6 Å) to subpocket-specific preliminary defined atom centers (hinge H area, Asp98 O atom and Ala100 N and O atoms; gate area 1 GA1, Phe97 CA atom; gate area 2 GA2, Lys52 NZ atom; solvent-exposed area 1 SE1, Arg356 CZ atom; solvent-exposed area 2 SE2 subpocket, His106 CE1 atom; αC area AC, Ser62 CA atom). For selecting hinge-binding fragments, hydrogen bonds to Asp98 O or Ala100 N or O was mandatory. Since a few fragments were assigned to multiple subpockets, the following prioritization scheme was applied: H annotation takes precedence over all the other annotations, therefore a fragment interacting with the hinge centers is only annotated as such. SE1 and SE2 were defined compatible so that fragments annotated as from both areas were automatically assigned only SE2. Similarly, fragments annotated as from both AC and GA2 areas were automatically assigned only GA2. In any other case of combination (e.g. fragments annotated as from GA2 and SE1), the annotations were considered ambiguous and the fragments were discarded.

In the third stage, H fragments were defined connectable to either GA1, GA2 or SE2 fragments (in the current work, although other connections are possible). Selected fragments were converted into sdf

format with OpenEye v.2.5.1.4. toolkit.[51] For each pair of fragments with hydrogen atoms connected, pairs of connectable atoms were searched based on their respective orientation as follows. A right circular cone (half-angle=π/4) is projected along the bond axis between any heavy atom A_i and a bound hydrogen atom H_i. A connectable atom pair A_1A_2 is selected if heavy atoms A_1 and A_2 are located in the projection cone of their counterpart (**Figure S7**).

In the fourth stage, the recently-described DeLinker[23] deep learning method was employed to generate linkers between above-described connectable atom pairs using the default model distributed with the package and a batch size of 1. Input data were prepared as ZINC atom types features to be ready for DeLinker using the 'prepare_data' module and by setting the 'test' parameter of the 'preprocess' function to 'True' as molecules are to be found. The linker length was set to a minimum of on and a maximum of six heavy atoms. Other parameters were kept by default. Generated molecules were saved as SMILES strings and further processed to remove redundancy for each connectable atom pair. In the final stage, unsuccessful linking attempts where only a single fragment is attached to the linker were removed using the function 'get_linker' in the 'frag_utils' utility. The remaining SMILES were filtered to keep only drug-like compounds according to in-house rules (**Table S2**). Next, the synthetic accessibility scores were computed with the the SAscore[40] method distributed with RDKit[52] to remove molecules with SAscore higher that three. Finally, molecules made of long flexible linkers were discarded according to our in-house filtering workflow (**Figure S8**), resulting in the first-round library (R1).

Comparison with ChEMBL and KinFraglib ligands. Standardized ChEMBL (1.7 million compounds) and KinFragLib (6.7 million) data were retrieved from the KinFragLib website.[53] Pairwise 2D fingerprint similarity to R1 molecules were assessed with RDKit[52] Morgan (radius = 2) topological fingerprint (default parameters, maximum path = 7).

Comparison to known CDK8 inhibitors. A search in the ChEMBL database[54, 41] for human CDK8 target assays resulted in three target report cards (CHEMBL3038474, CHEMBL5719 and CHEMBL3885556) from which bioassay data were joined and processed to keep compounds with a half maximal inhibitory concentration IC_{50} inferior or equal to 1 µM. Duplicates were then removed according to and the SMILES were standardized with OpenEye Filter v.3.0.1.2 (OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.). The final list of 302 inhibitors was searched in the generated drug-like subset described above for substructure 2D similarity using both RDKit Morgan (radius = 2) and topological (maximum path = 7) fingerprints and a combination of Tanimoto (Tc) and Tversky (Tv) metrics. Pairs were reported when morgan2 Tc ≥ 0.6 or morgan2 Tv ≥ 0.8 or RDKit7 Tc ≥ 0.75 or RDKit7 Tv ≥ 0.9.

Search for new potential CDK8 inhibitors. R1 library members were considered as potentially new at the condition that their similarity to any of 946 unique human CDK8-tested compounds (both active and inactive) reported in ChEMBL (target card reports CHEMBL3038474, CHEMBL5719 and CHEMBL3885556) and any of the 31 384 sc-PDB fragment is inferior to 0.50 (Tanimoto coefficient from RDKit topological fingerprints). Last, the subsequent list was searched for substructure similarity (RDKit topological fingerprint Tanimoto ≥ 0.90) to an in-house library of 8 280 193 commercially available drug-like compounds (**Supporting Table S3**).

Molecular docking. Virtual hits were drawn as 2D sketches with ChemAxon MarvinSketch v.16.10.17, (ChemAxon Ltd., 1031 Budapest, Hungary) saved in sdf file format, ionized at physiological pH with OpenEye Filter v.2.5.1.4 and finally converted in 3D structures (mol2 file) with Corina v.3.40 (Molecular Networks GmbH, 90411 Nürnberg, Germany), generating all possible stereoisomers and ring conformers simultaneously. The prepared molecules were docked into the above-described CDK8 cavity using PLANTS[55] v.1.2 The search space was set at 13 Å from the binding site center with a search speed of 1 (highest accuracy). 10 poses were generated per ligand, scored by the ChemPLP scoring

function and clustered using a root-mean square deviations (RMSD) of 2 Å on ligand heavy atoms. The flipped/rotated side chains were reconstructed in the protein structure for each corresponding PLANTS pose when applicable.

Molecular data analysis. Molecular descriptors (molecular weight ($g.mol^{-1}$), the count of heavy atoms (non-hydrogen atoms), logP, polar surface area (Å), count of H-bond acceptor, count of H-bond donor, count of rotatable bonds, count of ring systems, count of heteroatoms, bonds) were computed with RDKit. Data were processed with Python v.3.7.

Data visualization. Molecules were drawn in 2D with RDKit and MarvinSketch v.16.10.17, (ChemAxon Ltd., 1031 Budapest, Hungary). Three-dimensional structures were analyzed with Maestro v.2019-3 (Schrödinger, New York, NY 10036, U.S.A.) and Pymol v.2.1 (Schrödinger, New York, NY 10036, U.S.A.). Plots were generated with Matplotlib v3.0.2[56] in Python v.3.7.

Chemistry. All reactions were carried out under usual atmosphere unless otherwise stated. Chemicals and solvents were purchased from Enamine (LV-1035 Riga, Latvia) and were used without further purification. Yields refer to isolated compounds, estimated to be >95% pure as determined by 1H NMR or HPLC. 1H NMR spectra were recorded at 298 K on Bruker Avance III Spectrometer operating at 400 MHz. All chemical shift values δ and coupling constants J are quoted in ppm and in Hz, respectively, multiplicity (s = singulet, d = doublet, t = triplet, q = quartet, quin = quintet, sex = sextet m = multiplet, br = broad).

Preparative HPLC was performed using two methods: Method A) 2-10 min 30-70% acetonitrile, 30 ml/min ((loading pump 4 ml acetonitrile); column: YMC-ACTUS TRIART (C18; 100 mm x 20 mm; 5 µm);

Method B) 2-10 min 0-50% acetonitrile, 30 ml/min ((loading pump 4 ml acetonitrile); column: SunFire C18; 100 mm x 19 mm; 5 µm)

Analytical RP-HPLC-MS was performed using Agilent Technologies 1260 Infinity LC/MSD system with DAD\ELSD Alltech 3300 and Agilent LC\MSD G6120B mass-spectrometer using the following acquisition parameters: column, Agilent Poroshell 120 SB-C18 4.6x30mm 2.7 µm with UHPLC Guard Infinity Lab Poroshell 120 SB-C18 4.6x 5mm 2.7 µm; Temperature 60 C; Mobile phase A – acetonitrile : water (99:1%), 0.1% formic acid, B – water (0.1% formic acid); Flow rate 3 ml/min; Gradient : 0.01 min –99% B, 1.5 min – 0% B, 1.73 min - 0% B, 1.74 min - 99% B; Injection volume 0.5µl; Ionization mode Electrospray ionization (ESI); Scan range m/z 83-600; DAD 215 nm, 254nm, 280 nm. Purities of all tested compounds used in the biological assays were determined by HPLC/MS using the area percentage method on the UV trace recorded at a wavelength of 254 nm. All compounds were found to have >95% purity.

*1-(3-hydroxypyridin-4-yl)3-phenylurea (**38**).* To a stirred solution of phenylisocyanate (0.4 g, 3.4 mmol) in DMF (5 ml) was added a solution of 4-aminopyridin-3-ol hydrochloride (0.5 g, 3.4 mmol) in DMF (5 ml) followed by the addition of triethylamine (1.4 ml, 10.2 mmol) at room temperature (r.t.). The resulting mixture was stirred at room temperature overnight. The reaction mixture was concentrated under reduced pressure and the crude residue was purified by HPLC to afford 50 mg (6%) of the 1-(3-hydroxypyridin-4-yl)-3-phenylurea **38** as a white solid which was used for the next step without further purification.

*4-(3-phenylureido)pyridin-3-yl 4-ethoxybenzoate (**39**).* To a stirred solution of 4-ethoxybenzoic acid (36 mg, 0.22 mmol) in DMF (2 ml), compound **38** (50 mg, 0.22 mmol), EDC (50 mg, 0.26 mmol) and DMAP (27 mg, 0.22 mmol) were added. The resulting mixture was stirred at r.t. for 16 h. After completion of the reaction, the mixture was diluted with water (7 ml) and extracted with chloroform (3x7 ml). The combined organic layers were washed with saturated aqueous $NaHCO_3$, dried over anhydrous Na_2SO_4,

and concentrated under reduced pressure. The residue was purified by HPLC (method A) to afford compound **39** (40 mg, 49%) as a white solid. ^{1}H NMR (400 MHz, DMSO-d6) δ 9.25 (s, 1H), 8.60 (s, 1H), 8.38 – 8.25 (m, 3H), 8.17 (d, J = 8.6 Hz, 2H), 7.43 (d, J = 8.1 Hz, 2H), 7.30 (t, J = 7.7 Hz, 2H), 7.17 (d, J = 8.7 Hz, 2H), 7.01 (t, J = 7.3 Hz, 1H), 4.18 (q, J = 7.0 Hz, 2H), 1.38 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 378.2 [$(M+H)^+$, calcd. $C_{21}H_{20}N_3O_4$, 378.1].

*1-(5-hydroxypyridin-3-yl)-3-phenylurea (**40**).* Compound **40** was prepared as described above for compound **38**, starting from 5-aminopyridin-3-ol hydrobromide (0.65 g, 3.4 mmol). The reaction mixture was concentrated under reduced pressure and the crude residue was purified by HPLC (method B) to afford 60 mg (8%) of 1-(3-hydroxypyridin-5-yl)-3-phenylurea **40** as a white solid which was used for the next step without further purification.

*5-(3-phenylureido)pyridin-3-yl 4-ethoxybenzoate (**41**).* Compound **41** was prepared as described above for compound **39**, starting from 1-(5-hydroxypyridin-3-yl)-3-phenylurea **40** (60 mg, 0.264 mmol). The residue was purified by HPLC (method B) to afford compound **41** (36 mg, 45%) as a white solid. ^{1}H NMR (400 MHz, DMSO-d6). δ 9.01 (s, 1H), 8.83 (s, 1H), 8.46 (q, J = 2.7 Hz, 1H), 8.16 (d, J = 2.7 Hz, 1H), 8.08 (td, J = 5.5, 2.2 Hz, 2H), 7.99 (t, J = 2.5 Hz, 1H), 7.45 (d, J = 7.8 Hz, 2H), 7.28 (t, J = 8.0 Hz, 2H), 7.11 (dd, J = 9.1, 2.3 Hz, 2H), 6.98 (t, J = 7.4 Hz, 1H), 4.16 (dt, J = 10.1, 6.6 Hz, 2H), 1.36 (td, J = 6.9, 2.4 Hz, 3H). LC-MS (ESI) m/z 378.2 [$(M+H)^+$, calcd. $C_{21}H_{20}N_3O_4$, 378.1].

4-(1-phenyl-3,6-dihydro-2H-pyridin-4-yl)pyridin-3-ol ***(42)***. To a stirred solution of 4-iodopyridin-3-ol (0.63 g, 2.86 mmol, 1.1 eq.) and 1-phenyl-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-3,6-dihydro-2H-pyridine (0.74 g, 2.6 mmol, 1 eq.) in a mixture of 1,4-dioxane and water (20 ml, v/v=4:1), K_2CO_3 (1.8 g, 13 mmol, 5 eq.) was added and purged with argon for 30 min followed by the addition of Pd(dppf)Cl_2 (0.1 g, 0.05 eq.) and stirred at 90°C overnight. After completion, the reaction mixture was cooled to room temperature, diluted with ethyl acetate and water. The organic layer was washed with water and brine, dried over anhydrous sodium sulfate and evaporated under reduced pressure. The crude

product was purified by column chromatography on silica gel (hexane/EtOAc) to afford **42** (251 mg, 38%).

*4-(1-phenyl-4-piperidyl)pyridin-3-ol (**43**).* Compound **42** (251 mg, 1 mmol) was dissolved in MeOH (20 ml), followed by addition of Pd (10 wt % on activated carbon, 50 mg), and then the resulting suspension was stirred at room temperature under 1 atm. hydrogen pressure overnight. The resulting reaction was filtered, concentrated under reduced pressure, and dried under vacuum, to afford **43** (201 mg, 79%) which was used for the next step without further purification.

*[4-(1-phenyl-4-piperidyl)-3-pyridyl] 4-ethoxybenzoate (**44**).* A solution of compound **43** (201 mg, 1 eq.), 4-ethoxybenzoic acid (131 mg, 1 eq.), Et_3N (0.27 ml, 2.5 eq.) and HATU (360 mg, 1.2 eq.) in dry DMSO (2 ml) was stirred at room temperature for 12h. The completion of the reaction was monitored by LCMS. The mixture was purified by HPLC (Method A) to give compound **44** (120 mg, 38% yield) as a white solid. 1H NMR (400 MHz, DMSO-d6). δ 8.46 (d, J = 5.4 Hz, 2H), 8.15 – 8.09 (m, 2H), 7.50 (d, J = 5.1 Hz, 1H), 7.22 – 7.10 (m, 4H), 6.93 (d, J = 8.2 Hz, 2H), 6.75 (t, J = 7.3 Hz, 1H), 4.16 (q, J = 6.9 Hz, 2H), 3.78 (d, J = 12.3 Hz, 2H), 2.87 – 2.79 (m, 1H), 2.63 (t, J = 10.0 Hz, 2H), 1.82 (t, J = 5.1 Hz, 4H), 1.37 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 403.2 [(M+H)+, calcd. $C_{25}H_{27}N_20_3$, 403.2].

*5-(1-phenyl-3,6-dihydro-2H-pyridin-4-yl)pyridin-3-ol (**45**)*. To a stirred solution of 5-iodopyridin-3-ol (0.63 g, 2.86 mmol, 1.1 eq.) and 1-phenyl-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-3,6-dihydro-2H-pyridine (0.74 g, 2.6 mmol, 1 eq.) in a mixture of 1,4-dioxane and water (20 ml, v/v=4:1), K_2CO_3 (1.8 g, 13 mmol, 5 eq.) was added and purged with argon for 30 min followed by the addition of $Pd(dppf)Cl_2$ (0.1 g, 0.05 eq.) and stirred at 90 °C overnight. After completion, the reaction mixture was cooled to room temperature, diluted with ethyl acetate and water. The organic layer was washed with water and brine, dried over anhydrous sodium sulfate and evaporated under reduced pressure. The crude product was purified by column chromatography on silica gel (hexane/EtOAc) to afford compound **45** (326 mg, 49%).

*4-(1-phenyl-4-piperidyl)pyridin-3-ol (**46**).* Compound **45** (251 mg, 1 mmol) was dissolved in MeOH (20 ml), followed by addition of Pd (10 wt% on activated carbon, 50 mg), and then the resulting suspension was stirred at room temperature under 1 atm. hydrogen pressure overnight. The resulting reaction was filtered, concentrated under reduced pressure, and dried under vacuum, to afford compound **46** (220 mg, 86%) which was used for the next step without further purification.

*[5-(1-phenyl-4-piperidyl)-3-pyridyl] 4-ethoxybenzoate (**47**)*. A solution of compound 46 (200 mg, 1 eq.), 4-ethoxybenzoic acid (131 mg, 1 eq.), Et_3N (0.27 mL, 2.5 eq.) and HATU (360 mg, 1.2 eq.) in dry DMSO (2 ml) was stirred at room temperature for 12h. The completion of the reaction was monitored by LCMS. The mixture was purified by HPLC (Method B) to give compound **47** (140 mg, 44% yield) as a white solid. 1H NMR (400 MHz, DMSO-d6) δ 8.48 (d, J = 1.8 Hz, 1H), 8.41 (d, J = 2.4 Hz, 1H), 8.12 – 8.05 (m, 2H), 7.71 (t, J = 2.2 Hz, 1H), 7.21 (dd, J = 8.6, 7.1 Hz, 2H), 7.15 – 7.09 (m, 2H), 6.98 (d, J = 7.8 Hz, 2H), 6.76 (t, J = 7.3 Hz, 1H), 4.16 (q, J = 7.0 Hz, 2H), 3.82 (d, J = 12.1 Hz, 2H), 2.88 – 2.71 (m, 2H), 2.54 (d, J = 1.0 Hz, 1H), 1.92 (d, J = 11.8 Hz, 2H), 1.81 (qd, J = 12.4, 3.9 Hz, 2H), 1.37 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 403.2 [(M+H)+, calcd. $C_{25}H_{27}N_20_3$, 403.2].

*4-bromopyridin-3-yl 4-ethoxybenzoate (**48**).* A solution of 4-bromopyridin-3-ol (300 mg, 1.7 mmol, 1 eq.), 4-ethoxybenzoic acid (310 mg, 1.87 mmol, 1.1 eq.), DIPEA (0.89 ml, 5.1 mmol, 3 eq.) and HATU (760 mg, 2 mmol, 1.2 eq.) in DMF (10 ml) was stirred at 25°C for 16 h. The reaction mixture was poured into 50 ml of water and extracted with ethyl acetate (3x15 ml). The combined organic layers were washed with saturated ammonium chloride solution (50 ml) and brine (50 ml), dried over anhydrous sodium sulfate, and concentrated under reduced pressure to afford compound **48** as a brown solid (320 mg, purity 85%), which was used in the next step without further purification.

*4-(1-phenyl-1H-pyrazol-4-yl)pyridin-3-yl 4-ethoxybenzoate (**49**).* A mixture of compound **48** (200 mg, 0.62 mmol, 1 eq.), 1-(phenylpyrazol-4-yl)boronic acid (130 mg, 0.68 mmol, 1.1 eq.), cesium carbonate (400 mg, 1.24 mmol, 2 eq.) and Pd(dppf)Cl_2 (25 mg, 0.03 mmol, 0.05 eq.) in dioxane/water (5 ml, 10:1 v/v) was degassed and stirred at 105°C for 16 h under inert atmosphere. After cooling, the reaction

mixture was poured into 30 ml of water and extracted with ethyl acetate (4x10 ml). The combined organic layers were washed with brine (20 ml), dried over anhydrous sodium sulfate, and concentrated under reduced pressure. The crude material was purified by HPLC (Method A) to afford compound **49** as a white solid (235 mg, 36% yield after 2 steps). ^{1}H NMR (400 MHz, DMSO-d6). δ 9.06 (s, 1H), 8.61 – 8.51 (m, 2H), 8.22 (d, J = 8.8 Hz, 2H), 8.14 (s, 1H), 7.88 (d, J = 5.1 Hz, 1H), 7.75 (d, J = 8.0 Hz, 2H), 7.50 (t, J = 7.8 Hz, 2H), 7.34 (t, J = 7.4 Hz, 1H), 7.16 (d, J = 8.8 Hz, 2H), 4.19 (q, J = 6.9 Hz, 2H), 1.38 (t, J = 6.9 Hz, 3H). LC-MS (ESI) m/z 386.0 [$(M+H)^+$, calcd. $C_{21}H_{20}N_3O_4$, 386.1].

*5-bromopyridin-3-yl 4-ethoxybenzoate (**50**).* Compound **50** was prepared as described above for compound **48**, starting from 5-bromopyridin-3-ol (300 mg, 1.7 mmol, 1 eq.) to afford a yellow solid (260 mg, purity 90%), which was used in the next step without further purification.

*5-(1-phenyl-1H-pyrazol-4-yl)pyridin-3-yl 4-ethoxybenzoate (**51**).* Compound **51** was prepared as described above for compound **49**, starting from 5-bromopyridin-3-yl 4-ethoxybenzoate **50** (200 mg, 0.62 mmol, 1eq.). The crude material was purified by HPLC (method B) to afford compound **51** as a white solid (50 mg, 8% yield after 2 steps). ^{1}H NMR (400 MHz, DMSO-d6). δ 9.21 (s, 1H), 8.95 (d, J = 1.9 Hz, 1H), 8.44 (d, J = 2.6 Hz, 1H), 8.39 (s, 1H), 8.17 – 8.10 (m, 3H), 7.92 – 7.85 (m, 2H), 7.55 (t, J = 7.8 Hz, 2H), 7.35 (t, J = 7.2 Hz, 1H), 7.18 – 7.11 (m, 2H), 4.17 (q, J = 6.8 Hz, 2H), 1.38 (t, J = 6.8 Hz, 3H). LC-MS (ESI) m/z 386.0 [$(M+H)^+$, calcd. $C_{21}H_{20}N_3O_4$, 386.1].

In vitro CDK8 inhibition. Inhibitory activity of compounds was tested by using the LanthaScreen® Eu kinase binding assay optimized for CDK8/CyclinC (Invitrogen). This assay is based on the binding and displacement of an Alexa Fluor® 647-labeled ATP-competitive kinase inhibitor scaffold (kinase tracer) to the kinase. Binding of the tracer to the kinase is detected using a europium-labeled anti-tag antibody, which binds to the tagged CDK8/CyclinC. Simultaneous binding of both the tracer and antibody to the kinase results in a close proximity suitable for a high degree of FRET (fluorescence resonance energy transfer) from the europium (Eu) donor fluorophore to the Alexa Fluor® 647

acceptor fluorophore on the kinase tracer. Binding of an inhibitor to CDK8/CyclinC competes for binding with the tracer, resulting in a loss of FRET. Binding assay was performed into 384-well small volume plates (CORNING 3824) using kinase buffer provided by supplier (HEPES 50mM pH7.5, MgCl2 10mM, EGTA 1mM, Brij-35 0.01%) in a final volume of 15 µL. Briefly, 5µL of 3X compound (increasing concentrations from 3.10^{-11} to 3.10^{-5} M) prepared in kinase buffer are added to 5µL of 3X kinase/Ab solution (15nM kinase, 6nM biotin anti-His-tag antibody, 6nM Eu-streptavidin) and 5µL of 30nM kinase tracer236 (Kd 8 nM). The plate was incubated 1h at room temperature before reading with a TRF-compatible multi-well plate reader (Envision, PerkinElmer) using a classic TRF reading protocol (excitation at 337 nm; donor emission measured at 620 nm; acceptor emission measured at 665 nm). The TR-FRET signal was collected both at 665 and 620 nm, and TR-FRET ratios were calculated (acceptor signal value divided by donor signal value). IC_{50} and K_i values of the tested compounds were determined from competitive binding curves using GraphPad Prism software (version 6.07) as follows:

$$S = S_{min} + \frac{(S_{max} - S_{min})}{(1 + 10^{(X - logIC_{50})})}$$

S is the TR-FRET ratio value

X is the compound concentration

$$logIC_{50} = log10^{(logKi * \left(1 + \frac{[tracer]}{K_d}\right))}$$

[tracer] is the tracer concentration used in the competition assay

K_d is the dissociation constant value of the tracer

~~**Safety Statement.** No unexpected or unusually high safety hazards were encountered. All experiments were conducted under ISO 9001 compliance.~~

ASSOCIATED CONTENT

SUPPORTING INFORMATION

Supplementary Methods section and additional figures and tables including the comparison and alignment of sc-PDB subpocket and fragments to CDK8 ATP binding site, the colored feature histogram (c-FH descriptor) used to align sc-PDB subpockets to the target cavity, the validation of the subpocket comparison protocol, the pairwise similarity of selected fragments, the properties of selected fragments, the definition of connectable fragments, the topological requirements to connect fragment atoms by a linker, the filters for DeLinker-generated linkers, the properties of generated molecules, the LC-MS analysis of compounds **39**, **41**, **44**, **47**, **49** and **51**, the ProCare alignment of selected fragment-bound subpockets to the CDK8 cavity, the topological relationships between the 385 fragments selected for the linking stage the synthesis of round-2 library compounds, the list of CDK8 X-ray structures, the filtering rules to select drug-like compounds, the in-house catalog of commercially available drug-like compounds, the list of 37 commercially available compounds structurally similar or identical to round-1 library members, the list of 151 round-2 library members (PDF).

Molecular formula strings–SMILES codes (CSV)

Docking pose of compound 49 (PDB file format)

This material is available free of charge on the ACS Publications website at http://pubs.cas.org

DATA AVAILABILITY

Data. All scripts and input data necessary to run the workflow (CDK8 structures, ProCare subpockets alignment and scoring, fragment selection, DeLinker generative linking, library processing) have been made available at https://github.com/kimeguida/POEM.

Software. ProCare (version 0.1.1) is available at https://github.com/kimeguida/ProCare. IChem (version 5.2.9) was downloaded from http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html.

IChem is freely available for non-profit academic research and subjected to moderate license fees for companies. DeLinker was retrieved from https://github.com/oxpig/DeLinker.

ACKNOWLEDGMENTS

This work was funded by fellowship of the French Ministry of Higher Education, Research and Innovation (MESRI) to M.E. and of the Drug discovery and Development Institute (IMS, Strasbourg, grant nr. IMI-HIB-2022.). The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) is acknowledged for allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg, Germany) for providing an executable version of Protoss, OpenEye Scientific Software for the generous allocation of an academic license, and contributors to open-source software used in this study. We acknowledge F. Imrie (University of California, Los Angeles, USA) for discussions on DeLinker. Last, we warmly thank M. Semenova and the Enamine team (Kyiv, Ukraine) in charge of the synthesis of round-2 library compounds.

ABBREVIATIONS

2D, two-dimensional; 3D, three-dimensional; AC, αC helix; CDK8, cyclin-dependent kinase 8; FBDD, fragment-based drug design; FRET, fluorescence resonance energy transfer; GA, gate area; HPLC, high performance liquid chromatography; HTRF, homogeneous time-resolved fluorescence; HTS, high-throughput screening; ITC, isothermal titration calorimetry; MD, molecular dynamics, MS, mass spectrometry; NMR, nuclear magnetic resonance spectroscopy; RMSD, root-mean square deviations; SE, solvent-exposed; SMILES, simplified molecular input line entry system; SPR, surface plasmon resonance; TR-FRET, time-resolved FRET.

REFERENCES

1. Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H., Twenty Years On: The Impact of Fragments on Drug Discovery. *Nat Rev Drug Discov,* **2016,** *15*, 605-619.
2. Li, Q., Application of Fragment-Based Drug Discovery to Versatile Targets. *Front Mol Biosci,* **2020,** *7*, 180.
3. Troelsen, N. S.; Clausen, M. H., Library Design Strategies to Accelerate Fragment-Based Drug Discovery. *Chem. Eur. J.,* **2020,** *26*, 11391-11403.
4. Coyle, J.; Walser, R., Applied Biophysical Methods in Fragment-Based Drug Discovery. *SLAS Discov,* **2020,** *25*, 471-490.
5. Erlanson, D. A.; McDowell, R. S.; O'Brien, T., Fragment-Based Drug Discovery. *J Med Chem,* **2004,** *47*, 3463-3482.
6. Erlanson, D. A.; Davis, B. J.; Jahnke, W., Fragment-Based Drug Discovery: Advancing Fragments in the Absence of Crystal Structures. *Cell Chem Biol,* **2019,** *26*, 9-15.
7. Bian, Y.; Xie, X. S., Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *AAPS J,* **2018,** *20*, 59.
8. de Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijer, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J.; Leurs, R., Crystal Structure-Based Virtual Screening for Fragment-Like Ligands of the Human Histamine H(1) Receptor. *J Med Chem,* **2011,** *54*, 8195-8206.
9. Brooijmans, N.; Kuntz, I. D., Molecular Recognition and Docking Algorithms. *Annu Rev Biophys Biomol Struct,* **2003,** *32*, 335-373.
10. Sandor, M.; Kiss, R.; Keseru, G. M., Virtual Fragment Docking by Glide: A Validation Study on 190 Protein-Fragment Complexes. *J Chem Inf Model,* **2010,** *50*, 1165-1172.
11. Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W., Docking Performance of Fragments and Druglike Compounds. *J Med Chem,* **2011,** *54*, 5422-5431.
12. Marcou, G.; Rognan, D., Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J Chem Inf Model,* **2007,** *47*, 195-207.
13. Jacquemard, C.; Drwal, M. N.; Desaphy, J.; Kellenberger, E., Binding Mode Information Improves Fragment Docking. *J Cheminform,* **2019,** *11*, 24.
14. Chachulski, L.; Windshugel, B., Leads-Frag: A Benchmark Data Set for Assessment of Fragment Docking Performance. *J Chem Inf Model,* **2020,** *60*, 6544-6554.
15. Guvench, O., Computational Functional Group Mapping for Drug Discovery. *Drug Discov Today,* **2016,** *21*, 1928-1931.
16. Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S., The Ftmap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat Protoc,* **2015,** *10*, 733-755.
17. Radoux, C. J.; Olsson, T. S.; Pitt, W. R.; Groom, C. R.; Blundell, T. L., Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J Med Chem,* **2016,** *59*, 4314-4325.
18. Guvench, O.; MacKerell, A. D., Jr., Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol,* **2009,** *5*, e1000435.
19. Chen, H.; Zhou, X.; Wang, A.; Zheng, Y.; Gao, Y.; Zhou, J., Evolutions in Fragment-Based Drug Design: The Deconstruction-Reconstruction Approach. *Drug Discov Today,* **2015,** *20*, 105-113.
20. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., Recap--Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J Chem Inf Comput Sci,* **1998,** *38*, 511-522.
21. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M., On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem,* **2008,** *3*, 1503-1507.
22. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M., Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J Chem Inf Model,* **2007,** *47*, 390-399.
23. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M., Deep Generative Models for 3d Linker Design. *J Chem Inf Model,* **2020,** *60*, 1983-1995.

24. Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H., Syntalinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chem Sci,* **2020,** *11*, 8312-8322.
25. Imrie, F.; Hadfield, T. E.; Bradley, A. R.; Deane, C. M., Deep Generative Design with 3d Pharmacophoric Constraints. *Chem Sci,* **2021,** *12*, 14577-14589.
26. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N., An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3d Structure. *Structure,* **2018,** *26*, 499-512 e492.
27. Gao, M.; Skolnick, J., A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput Biol,* **2013,** *9*, e1003302.
28. Eguida, M.; Rognan, D., A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J Med Chem,* **2020,** *63*, 7127-7142.
29. Dale, T.; Clarke, P. A.; Esdar, C.; Waalboer, D.; Adeniji-Popoola, O.; Ortiz-Ruiz, M. J.; Mallinger, A.; Samant, R. S.; Czodrowski, P.; Musil, D.; Schwarz, D.; Schneider, K.; Stubbs, M.; Ewan, K.; Fraser, E.; TePoele, R.; Court, W.; Box, G.; Valenti, M.; de Haven Brandon, A.; Gowan, S.; Rohdich, F.; Raynaud, F.; Schneider, R.; Poeschke, O.; Blaukat, A.; Workman, P.; Schiemann, K.; Eccles, S. A.; Wienke, D.; Blagg, J., A Selective Chemical Probe for Exploring the Role of Cdk8 and Cdk19 in Human Disease. *Nat Chem Biol,* **2015,** *11*, 973-980.
30. Mallinger, A.; Schiemann, K.; Rink, C.; Stieber, F.; Calderini, M.; Crumpler, S.; Stubbs, M.; Adeniji-Popoola, O.; Poeschke, O.; Busch, M.; Czodrowski, P.; Musil, D.; Schwarz, D.; Ortiz-Ruiz, M. J.; Schneider, R.; Thai, C.; Valenti, M.; Brandon, A. D.; Burke, R.; Workman, P.; Dale, T.; Wienke, D.; Clarke, P. A.; Esdar, C.; Raynaud, F. I.; Eccles, S. A.; Rohdich, F.; Blagg, J., Discovery of Potent, Selective, and Orally Bioavailable Small-Molecule Modulators of the Mediator Complex-Associated Kinases Cdk8 and Cdk19. *Journal of Medicinal Chemistry,* **2016,** *59*, 1078-1101.
31. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E., Sc-Pdb: A 3d-Database of Ligandable Binding Sites--10 Years On. *Nucleic Acids Res,* **2015,** *43*, D399-404.
32. Desaphy, J.; Rognan, D., Sc-Pdb-Frag: A Database of Protein-Ligand Interaction Patterns for Bioisosteric Replacements. *J Chem Inf Model,* **2014,** *54*, 1908-1918.
33. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D., Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J Chem Inf Model,* **2012,** *52*, 2287-2299.
34. van Linden, O. P.; Kooistra, A. J.; Leurs, R.; de Esch, I. J.; de Graaf, C., Klifs: A Knowledge-Based Structural Database to Navigate Kinase-Ligand Interaction Space. *J Med Chem,* **2014,** *57*, 249-277.
35. Rusu, R. B.; Cousins, S., 3d Is Here: Point Cloud Library (Pcl). *Ieee Int Conf Robot,* **2011**.
36. Zhou, Q.-Y.; Park, J.; Koltun, V., Open3d: A Modern Library for 3d Data Processing. *arXiv:1801.09847,* **2018**.
37. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discov Today,* **2003,** *8*, 876-877.
38. Sterling, T.; Irwin, J. J., Zinc 15--Ligand Discovery for Everyone. *J Chem Inf Model,* **2015,** *55*, 2324-2337.
39. Wang, R.; Fang, X.; Lu, Y.; Wang, S., The Pdbbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J Med Chem,* **2004,** *47*, 2977-2980.
40. Ertl, P.; Schuffenhauer, A., Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J Cheminform,* **2009,** *1*, 8.
41. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., Chembl: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res,* **2012,** *40*, D1100-1107.
42. Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A., Kinfraglib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination. *J Chem Inf Model,* **2020,** *60*, 6081-6094.

43. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J Chem Inf Model,* **2010,** *50*, 742-754.
44. Riniker, S.; Landrum, G. A., Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J Cheminform,* **2013,** *5*, 26.
45. O'Boyle, N. M.; Sayle, R. A., Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. *J Cheminform,* **2016,** *8*, 36.
46. Li, Y. B.; Pei, J. F.; Lai, L. H., Structure-Based De Novo Drug Design Using 3d Deep Generative Models. *Chemical Science,* **2021,** *12*, 13664-13675.
47. The Protein Data Bank, Https://Www.Rcsb.Org (Accessed 05-10-2022).
48. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res,* **2000,** *28*, 235-242.
49. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M., Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J Cheminform,* **2014,** *6*, 12.
50. Da Silva, F.; Desaphy, J.; Rognan, D., Ichem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem,* **2018,** *13*, 507-510.
51. Openeye Scientific Sofware, Santa Fe, Nm 87508, U.S.A.
52. Rdkit: Open-Source Cheminformatics Software, Http://Www.Rdkit.Org/ (Accessed 05-09-2022)
53. Kinfraglib, Https://Zenodo.Org/Record/3956580 (Accessed 01-23-2022).
54. Chembl, Https://Www.Ebi.Ac.Uk/Chembl (Accessed 10-14-2021).
55. Korb, O.; Stutzle, T.; Exner, T. E., Empirical Scoring Functions for Advanced Protein-Ligand Docking with Plants. *J Chem Inf Model,* **2009,** *49*, 84-96.
56. Matplotlib, Https://Matplotlib.Org/3.0.2/ (Accessed 04-10-2022).

Table of content graphic

~~**Target-focused library design by pocket-applied computer vision and fragment deep generative linking.**~~

~~Merveille Eguida, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa, and Didier Rognan.~~