Chemical Proteomics-based Target Prioritization through a Residue Agnostic Ligandability Assessment Platform

Fettah Erdogan*[a], Raiyan Chowdhury[c], Serap Beldar[a], Tom Bobby Chandy[d], Rebecca G. Allan[a], Elvin D. de Araujo[a], Patrick T. Gunning*[a,b]

[a] Department of Chemical and Physical Sciences, University of Toronto Mississauga, Ontario L5L 1C6, Canada

[b] Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada

[c] Cheriton School of Computer Science, University of Waterloo, Ontario N2L 3G1, Canada

[d] Faculty of Health Sciences, Western University, London, ON N6A 3K7, Canada

Corresponding Authors

*Patrick T. Gunning, Department of Chemical and Physical Sciences, University of Toronto Mississauga, 3359 Mississauga Rd N., Mississauga, Ontario L5L 1C6

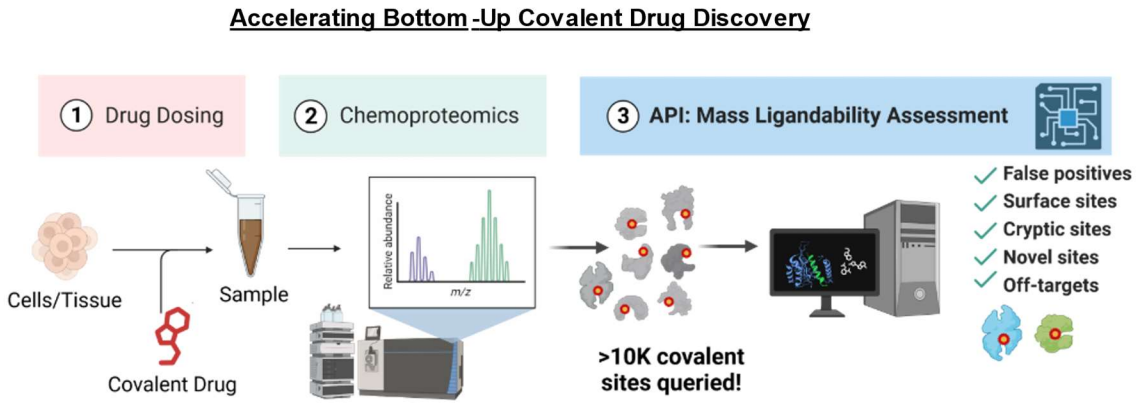
Email: patrick.gunning@utoronto.ca

*Fettah Erdogan, Department of Chemical and Physical Sciences, University of Toronto Mississauga, 3359 Mississauga Rd N., Mississauga, Ontario L5L 1C6

Email: fettah.erdogan@alum.utoronto.ca

Subject terms: covalent binding site, computational ligandability assessment, chemoproteomics, chemical proteomics, computational target prioritization, computer-aided drug discovery, machine learning

Graphical Abstract



Abstract

The landscape of drug discovery is undergoing a transformative phase with the influx of structural biology and omics data. Identifying optimal drug targets amid this data surge presents a multifaceted challenge. Covalent inhibitors, once undervalued, now hold substantial promise, especially targeted covalent inhibitors (TCIs), effectively engaging 'undruggable' proteins and overcoming resistance mechanisms. Existing ML software can proficiently model covalent ligands but lack comprehensive utility across large chemoproteomics sites. Challenges persist in predicting and assessing cryptic ligandable sites and sites beyond cysteine, demanding advanced computational tools. As cysteine-ligandable proteins represent only ~20% of the quantifiable proteome, there is a requirement for ligandability mapping of other nucleophilic amino acids. This study introduces a pioneering computational pipeline leveraging an AI-based ligandable predictor for meticulous evaluation of chemical proteomics-based reactive sites. The pipeline offers a scalable framework to assess covalent ligandability on a large scale, filter out improbable hits and systematically evaluate potential drug targets. Our work addresses covalent drug design challenges through a pipeline that fills crucial gaps in predicting cryptic ligandable and covalent sites in addition to cysteines to foster more efficient drug discovery methodologies.

Introduction

The magnitude of structural biology and omics information arriving at the service of drug discovery and medicinal chemistry research is expediting the rate of hit-to-lead development[1]. However, harnessing this deluge of information to identify optimal drug targets poses a multifaceted challenge. The efficacy and success of drug development largely hinges upon precise target identification, necessitating innovative technologies to address these challenges. Covalent inhibitors, long overlooked due to concerns about reactivity and off-target effects[2], are now gaining substantial attention and reverence in both academic and pharmaceutical drug design programs. Improved understanding of the factors influencing reactivity and emergence of new types of warheads in the form of targeted covalent inhibitors (TCIs), represent a promising avenue within drug discovery, particularly in targeting proteins previously deemed 'undruggable'.[3–8] Over 40 covalent drugs are currently under clinical development[2] and many previously so called "undruggable" targets and mechanisms of resistance have now been effectively tackled by covalent compounds.[3,9,10] Recently developed and FDA approved TCIs, AMG-510 (sotorasib) and MRTX-849 (adagrasib), effectively engage Cys12 of $KRas^{G12C}$, a famously difficult target implicated in 40% of lung cancers.[6,7,11] Additionally, ibrutinib, a first-in-class inhibitor known to successfully bind to Cyst of Bruton's tyrosine kinase (BTK), which is linked to overexpression of B cells in B cell cancers, has FDA approval for lymphoma.[12–14] Other FDA approved drugs such as Aspirin, Penicillin G, and Fosfomycin are all reported to engaged targets via covalent interaction[15].

TCIs enact specificity and irreversible binding through a combination of covalent and non-covalent interactions at their protein's target site. Although non-covalent interactions make up the majority of contacts between a TCI and the binding pocket residues of the protein, target engagement via covalent bond formation (kinact) increases drugging efficiency (KI) by prolonging duration of action and increasing the degree of protein-drug occupancy[16] in some cases de-coupling PK-PD[17], creating an opportunity to target shallow binding sites of challenging targets[18] and improving selectivity[19] towards target, target isoforms and disease-linked mutants. Given these appealing features, development of new tools to support the rational design of TCIs has become an indispensable task.

Moreover, drug discovery remains a prolonged and expensive process characterized by a notably low (13.8%) success rate[20]. The integration of machine learning (ML) technologies into drug discovery research emerges as a crucial avenue to address these challenges. Numerous

ML models focused on small-molecule design and quantitative structure-activity relationship (QSAR) have been specifically designed for virtual screening (VS) of drugs against targets. These innovations represent pioneering steps toward reducing overall costs and timelines in drug discovery research.[21,22] Various available computer-aided drug design (CADD) and screening programs such as Schrödinger CovDock[23], DUckCov[24], Cov_DOX[25], WIDOCK[26], Reactive Docking[27], and BIreactive[28], specialize in modeling close contacts and predicting the binding mode of covalent ligands or the reactivity of warheads against a model sidechain. However, these programs efficiently serve their purpose only with prior knowledge of the binding site. As a result, they often require independent and laborious use in combination with non-/covalent binding site predictor programs such as P2Rank[29], PocketFinder[30], SiteMap[10], and PocketMiner[31], to initially identify non-covalent binding cavities, followed by calculations that suggest whether the binding pocket is amenable to covalent modification. Nevertheless, while some of these programs can be impractically coupled with nascent cysteine reactivity predictors such as DeepCoSi[32], sbPCR[33], and HyperCys[34], many of these programs are intended for top-down proteomics discovery or have limitations. None of these platforms are designed with a utility of assessing ligandabilities across large numbers of chemoproteomics sites or are tailored for exploring sites other than cysteines. Recently, an MS-based quantitative proteolysis method (LiP-Quant) lightly integrating ML was used to prioritize true drug targets in chemoproteomics output dataset[35]. Although an expensive method anticipating establishment of acceptable detection rates on genuine ligandable sites and false-positives, LiP-Quant method can be promising for future chemical proteomics-based drug discovery research. In this work, we sought to develop an inexpensive method capable of analyzing existing chemoproteomics target sites in public or private repositories for proximal binding cavities and prioritize them using a ligandability score.

Covalent fragment-based drug discovery (FBDD) approaches using quantitative mass spectrometry (QMS) have conventionally been used to study inaccessible protein targets, reveal cryptic pockets and identify new potential targets in the proteome[36–38]. However, the success rate of chemical proteomics-based drug screening is comparable to crystallographic fragment-screening (5-10%).[39] Given the molecular dynamism of protein structures and the potential for occlusion of many transitory binding cavities, subpockets or PPI sites, an integrative *in-silico* pipeline that searches for potential druggable sites near nucleophilic residues specifically is a necessary part of the chemical proteomics-based covalent drug design pipeline.[27,35,40,41] Typical cell-based shotgun/bottom-up proteomics experiments can suggest >10,000 implicit protein

targets[42] however, visual inspection or *in-vitro* biophysical validation of each suggestive chemical proteomics-based hit is impractical and impossible. Given this challenge and in meeting the growing demand for TCI development[5], we developed a computational pipeline that uses an AI-based ligandable predictor (DeepPocket[43]-developed) to scrupulously assess and mass-validate chemical proteomics-based reactive sites for apparent and cryptic ligandable cavities adjacent to each site. This comprehensive platform allows researchers to upload a proteomics-based covalent hits and perform a mass-scale ligandability evaluation of each suggestive target site for filtering out improbable hits. DeepPocket has emerged as a state-of-the-art model which builds on an established computational method, namely FPocket[44], to map out the precise boundaries of ligandable sites and detect subcavities on a protein structure. Although the protein data bank (PDB) provides one of the most detailed descriptors on protein structures, it is imperative to emphasize that such structures are only simplified snap-shot models of the target macromolecules. Regeneration of these structures into conformer archetypes that contain slight perturbations in residue side chain coordinates (using tools such as CONCOORD[45]) can reveal transitory structural patterns for opportunistic pocket detection while preserving macromolecule's unique architectural style. CONCOORD produces protein conformers around an experimental structure using geometric restrictions. Studies of Molecular Dynamics (MD) simulations indicate that collective degrees of freedom are crucial to protein conformational changes, which are often vital to protein function. These internal constraints and configurational barriers can be used be used to simulate conformers without the need for more CPU intensive MD simulations[45]. Conformer generation using CONCOORD has been used in a similar fashion before[46,47] however we report its first use in covalent site ligandability assessment. Molecular structure descriptors such as PDB files are not to be thought of as a set of fixed coordinates but rather a framework for generating hypotheses based on molecular patterns to be explored[48]. The recent decade witnessed development of powerful generative AI models trained on omics data leading to algorithm that can recognize molecular features when faced with new data bearing similar characteristics. Our choice of DeepPocket as the pocket-predicting platform was inspired by a combination of its reliability on long-established methods, top-level performance and partitioned architecture which allowed us to further develop it towards detection of cryptic sites.

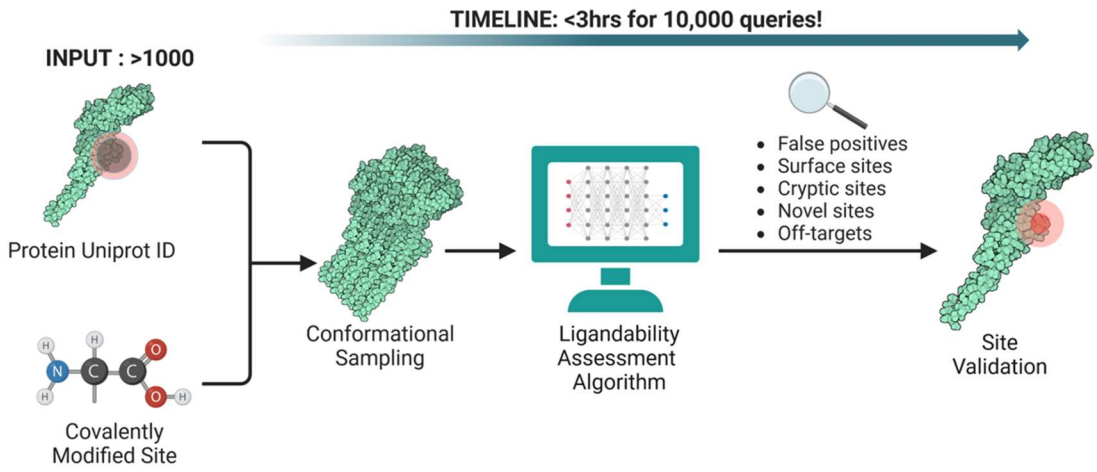


Figure 1: Ligandability assessment pipeline architecture. Thousands of chemoproteomics-based covalent site information is simultaneously fed into the pipeline using the Uniprot ID and residue ID of the modified targets and their sites, respectively. Protein structures are subject to random perturbations according to a predefined set of rules and constraints to produce conformers. Each protein structure is passed through a geometry-based candidate pocket detection algorithm and re-ranked using an ML algorithm[43,44]. Surface voxels of the pockets are then used to compute distances from the respective target site to the pocket surfaces and analyzed by a ligandability assessment algorithm. The pipeline outputs ligandability score for each query originally input at an extraordinary rate.

With diverse nucleophilic amino acids in ligandable proteins and the emerging need for comprehensive mapping of the human proteome, covalent drug design approaches are poised for significant advancements. The identification and targeting of cryptic ligandable sites within proteins present immense potential for novel therapeutic interventions. However, the challenges in predicting and assessing these sites, particularly beyond cysteine residues, underscore the critical need for advanced computational tools. The developed pipeline (**Figure 1**), leveraging AI-driven predictive models robustly assesses and validates chemical proteomics-based reactive sites with exceptional speed. These advancements not only enable the detection of cryptic pockets but also pave the way for a more efficient and systematic evaluation of potential drug targets. By offering a scalable platform for ligandability assessment and rational hit filtration, this work takes a significant step towards addressing the challenges in covalent drug discovery, ultimately aiding for more efficient, precise, and successful drug development strategies in the future.

Methods

Data Curation and Preprocessing

Unless otherwise mentioned, all data preparation and processing pipelines were built in-house using Python 3.10.0 and the Biopython[49] library. Since most PDB files contain multiple chains of proteins or macromolecules, each PDB file assess in this study was dissected for chain specific assessment, cleaned off from all ligands and/or double checked for errors with the associated covalent residue in the PDB. Any PDB chain alone with a covalent ligand sitting on a non-cavity site (absence of non-covalent interaction), missing the covalent residue in questions, or imperceptibly cryptic was omitted from analysis. Where applicable and unless otherwise mentioned, the positive covalent residue for each PDB was extracted and its respective negative was assigned by taking the most distant matching residue (matching false positive) within the chain.

Holo Sites. The initial list of 2,294 PDB coordinate files for high-resolution co-crystal structures of experimental covalent ligand-bound proteins (herein referred to as holo-protein) was scraped from the covPDB database[50]. A summary data file containing residue- and chain-specific annotation information was obtained directly from the covPDB website. Each annotated covalent site was subject to visual inspection (manual) for containment of the covalent ligand in a pocket-like cavity of the residue and saved as refined list (covPDBs, SupplementaryFile1).

Apo sites. Using the list of Uniprot accession codes referenced to the covPDB proteins, a list 3,527 apo counterparts of the covPDBs (apo-covPDBs) was initially scraped from the RCSB database[51] while filtering out entities with non-polymer molecules (except waters, metal ions, and small-molecules typically used in buffers). The refined list of apo-CovPBDs used for analysis was arduously compiled by manual alignment to the covPDB counterparts and cross-checking for presence of all pocket-forming residues (apo-covPDBs, SupplementaryFile1).

Cryptic sites. The initial list of 124 apo-cryptic PDB files and their corresponding experimental holo-proteins was curated from Meller, et al[31] and Cimermancic, et al[52] and/or scraped from the RCSB repository[53] and refined further to a finalized list (cryptic-PDBs, SupplementaryFile1). By aligning the holo-protein counterparts of cryptic PDBs to apo-cryptic structure coordinates, a mock-covalent residue closest in distance to the ligand in the cryptic pocket of the apo-cryptic PDB was assigned to each cryptic PDB.

Pipeline Architecture and Data Processing

Except for Alphafold PDB coordinate files, each PDB chain primed for processing in this work is initially examined for potential conformational states that may reveal transitory or occluded binding pockets on protein structure. These conformational states alongside the experimental state and Alphafold predicted structures are scanned for identification of druggable binding sites. Methodologies describing protein conformer generation and pocket prediction are detailed in CONCOORD[45] and DeepPocket[43], respectively. Briefly, each PDB chain and/or conformer structure was initially run on Fpocket[44] to calculate the barycenter of candidate pockets. Fpocket is a protein pocket (cavity) detection algorithm based on Voronoi tessellation and detects pocket curvatures in most protein structures with high accuracy. Subsequently, constant-sized grids are then placed at the barycenter of each candidate pocket and are scored using convolutional neural networks (CNN). A final 3D-shaped pocket structure of the top-ranked centers are then generated using a CNN segmentation model. The 3D-structure of pockets is constructed via voxelization, and the indices of the constructing voxel indices are converted to cartesian coordinates. Distance calculations are then taken between the surface of the binding pockets and the centers of the atoms of covalent residue(s) (Cys, Lys, Ser, Tyr, Thr, His, Asp or Glu). this study.

Ligandability assessment. Ligandability assessment is done to gauge the likelihood of ligandable cavity presence within rational distance of the residue within limits of standard small molecule dimensions. Whereas for the simplistic ligandability assessment, a distance threshold of 10Å was set for determining ligandability of a covalent site, to minimize false-positive rates, this threshold was set to 9Å for the advanced ligandability platform. The distance threshold is set based on the maximum of the shortest distance computed from pocket surface to covalent residue across three datasets; covalent sites of 1,647 covPDBs, 230 apo-covPDBs and mock sites of 90 cryptic-PDBs(**Figure 2**).

Results

Simplistic Ligandability Assessment

Using a list of empirically validated ligand binding sites, covalent site ligandability assessment platform was initially subject to performance validation using a simplistic design method of taking distance computations from a query residue to the nearby predicted pockets. Initially, four pools of PDB datasets were prepared; i) a list of empirically determined holo protein co-crystal structures with a covalent ligand bound at a ligandable sites (“Holo sites”), ii) a list of experimentally solved apo counterpart structures of the holo protein list (“Apo sites”), iii) a list of cryptic proteins with a mock covalent site designated in their respective occluded binding cavity (“Cryptic sites”) and, iv) a list of the most extreme surface residues lacking cavities or visible ligandable sites in their vicinities designates as “False-positive sites”. The dataset used in ligandability detection power against holo sites comprised of 2,251 holo-protein structures from the covPDB database[50], apo sites comprising of 248 apo protein structures, cryptic sites comprising of 118 cryptic protein structures from the from RCSB database[51], and false-positive sites comprising of 112 false positive sites in the Cryptic-PDBs list (SupplementaryFile1).

Overall, covalent site ligandability assessment using a simple distance (10Å) threshold approach on rigid protein structural models demonstrated a competent detection performance of 85% when tested on holo site and satisfactorily (72%) when tested on apo site data (**Figure 2**). These scores correlate with formerly reported F-pocket and DeepPocket binding site predictions on holo and apo protein structures, respectively[43,44]. Generally, detection performance of many algorithms including F-pocket, Kalasanty, DeeplyTough, DeepSite, and DeepCoSI albeit considerate of signature structural features representing dynamic protein regions, exhibit suboptimal detection performance on holo sites[43] as demonstrated in this study. The inability of the simplistic approach to capture ligandability of apo sites could associated with multiple factors including mitigation of high false positive rates and inability to recognize unstructured features used in training the pocket detection algorithms. Apo sites are expected to possess lower solvent accessible surface area (SASA), unstructured architecture of the ligand cavity and imprecise positions of pocket-forming atoms from the ideal expected coordinates of the ligandable site. Despite their consideration of protein dynamics, these distinct features could pose challenges for algorithms trained on holo pockets, especially when identifying valid sites within apo structures.

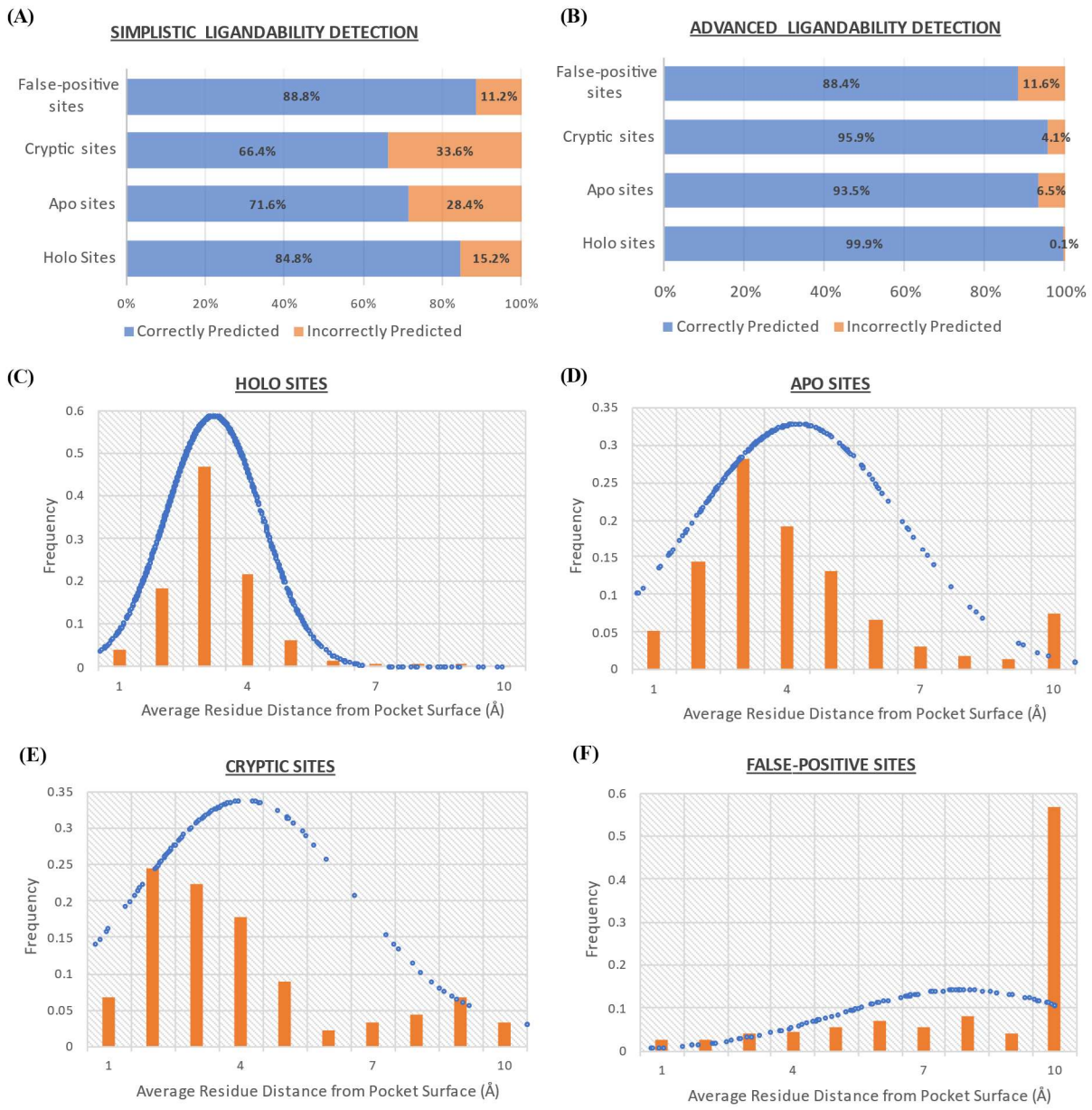


Figure 2: Performance metrics of the covalent ligandability assessment platforms. Panels (**A**) and (**B**) depict the various types of binding sites tested for correct prediction (filtering out false-positives or detecting experimental covalent binding sites) and incorrect prediction (predicting false-positives as covalent binding sites or failing to detect experimental covalent binding sites. Predictions made by (**A**) the simplistic platform designed on utilizing rigid protein structural models are compared to (**B**) predictions made by the advanced platform. Panels (**C-F**) depict the distribution scatter of distances from the covalent residue centers to nearest pocket surfaces in the various categories of dataset analyzed by the advanced platform. The bars in the distribution plots depict the relative number (frequency) of sites that fall in residue-pocket distance interval bins of 1Å (i.e. 1 for 0-1Å bin, 2 for 1-2Å bin, 3 for 2-3Å bin, etc.) in each of the dataset types. For simplicity, any distance higher than the set threshold of 9Å in the various datasets was placed in the 9-10Å bin.

Advanced Approach: Chemoproteomics Site Ligandability Assessment using an Amalgamated Platform

The vital role of protein conformational sampling in detecting hidden or unformed covalent ligandable sites was assessed using an advanced platform designed to consider protein conformational fluctuations (**Figure 2**). When tested on the same data set of holo sites, the advanced covalent ligandability assessment platform performed remarkably by correctly detecting >99% of these sites. This distinct performance is similarly reflected on apo sites data where the advanced ligandability assessment platform was able to capture 94% of the covalent ligandable sites on apo protein structures. The binding sites on apo proteins correspond to the liganded site on their holo structure counterparts. Since applications of conformational sampling of proteins is expected to widen of existing pockets or reveal subpockets thereby decreasing average distance between pocket surface and covalently modified residue in the PDB structure (**Suppl. Figure 2**), a distance threshold of 9Å was used by the advanced ligandability assessment platform. Overall, in going from the simplistic distance computation approach (**Suppl. Figure 3**) to the advanced platform (**Figure 2**), a distance shift (0.5Å to 2.0Å) in the average distribution of residue-to-pocket distances and a shift in the relative number of sites towards the lower distance interval bins occur across the datasets. Decrease in the average residue-to-pocket distance incurred by opening or widening of pockets is more obvious when comparing apo and cryptic dataset across the two platforms.

The proportion of empirical holo and apo sites missed by the advanced platform was respectively 0.1% and 6.5% compared to 11.2% and 28.4% using the simplistic ligandability assessment approach. Each “incorrectly predicted” holo site missed by either of the simplistic and advanced approaches was subject to visual inspection for presence of a ligandable cavity occupied by a ligand in the holo protein. For the incorrectly predicted apo sites; using the 3D atomic coordinate file of the subjective apo protein structure, the apo site in apo-covPDBs was aligned to its holo site counterpart from the covPDB pool and was visually inspected for experimental truncations and presence of residues fully forming the cavity in the ligand-complexed covPDB counterpart. Details on these corresponding covPDB IDs and their alignment analysis at the global scale as well as covalent ligand binding region can be found in SupplementaryFile1.

Platform Validation on Ligandable Cryptic Sites

In high probabilistic cases of ligandable covalent sites, binding of a covalent small molecule on a protein site is expected to induce structural change and potential opening or enlargement of a ligandable pocket or ostensibly a surface site. These structural transformations can give rise to discovery of hidden ligandable sites on proteins. Certainty around whether a previously unreported cavity detected in a conformational state as a false positive or, yet an undiscovered cryptic site is questionable. However, it is often possible to distinguish between superficial PPI-like sites and a buried cryptic binding site. In the past decade, many PPIs which once considered "undruggable" have now been successfully targeted by drug-like molecules[54,55]. Protein surface cavities and novel ligandable pockets are revealed from loop motions, secondary structure element motions and changes, and interdomain motions that increase SASA on the target. It is imaginable that such perturbations cause exposure of covalently reactive residues that may otherwise be buried as reported for most covalent cysteine sites.[56–58] This may give rise to indefinite estimates for setting a proximity distance threshold between covalent residue and pocket (**Figure 2, Suppl. Figure 1**). Notably, the power of the advanced covalent ligandability assessment platform is well demonstrated by its high detection scores on covalent sites in the apo-covPDB dataset including those which are buried (**Suppl. Table 1**) as well as in the cryptic-PDB dataset. Compared with a simplistic distance-based simplistic approach, the advanced ligandability assessment platform showcased an improved detection rate on missed cryptic-PDBs by 8-fold (33.6% vs 4.1%, **Figure 2**). This indicates that the advanced platform is able to capture a diversity of conformational changes that lead to formation of cryptic pockets while maintaining a comparable detection rate of false-positives sites with the simplistic ligandability assessment platform (**Suppl. Figure 3**).

Eliminating Non-Ligandable Sites

To assess the capacity of the advanced platform to successfully disqualify false-positive covalent sites and minimize the chance of a false-negatives in spite of increased positive detection rates, we selected a new mock list of false-positive sites using the apo-covPDB dataset. This list comprises of the most the superficial (solvent accessible) residue on the apo-covPDB structure whose identity matches the covalent ligand-bound residue in the CovPDB counterpart (Matching-Negative Residue column in the SupplementaryFile1). For instance, the covPDB structure 6CGE is covalent ligand-bound at HIS221 and the apo-covPDB counterpart of this protein is 1BHS. With reference to HIS221, HIS179 represents the most distant matching residue ID in 1BHS. Since

HIS179 remained unliganded in the experimental structure of the covalent ligand-bound structure of the protein, it represents the most reliable matching false-positive site. Although a similar list of matching false positives was created for the covPDBs and made available in SupplementaryFile 1, due to the large number of covPDB matching false positive sites and unfeasibility for visual inspection to confirm correct/incorrect prediction for each site, we sufficed our discussion with apo-covPDB matching false-positives (and Cryptic-PDB false-positives, **Figure 2**) evaluation trials of the two platforms. Any matching false positive not found in the apo-covPDB structure or not be superficially exposed although being most distant from the covalently modified residue was omitted from analysis. The cryptic-PDB false positives list comprises residue sites on the most extreme or superficial areas of the protein structure tested to ensure that false-positive detection rates remain loyal when assessing non-ligandable surface residues (SupplementaryFile1). Overall, although the simplistic platform demonstrates a false positive detection rate (11.2%, **Suppl. Figure 3**) comparative to the advanced approach (11.6%), it does so at the expense of missing real ligandable sites.

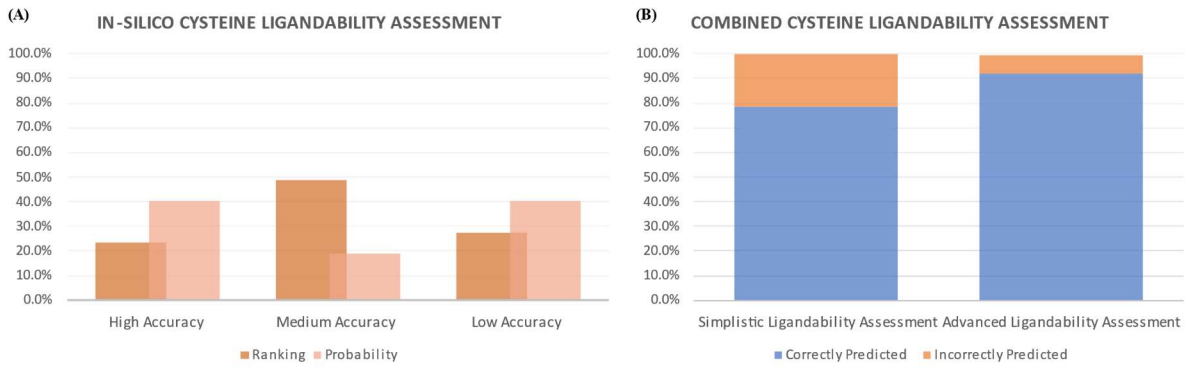


Figure 3: Cysteine ligandability detection. Covalent ligandability assessment against a list of buried cysteine residues from the apo-covPDB dataset which are empirically confirmed to be ligandable was compared across two different methods. The first approach (**A**) entirely dependent on predictions made by an ML algorithm[32] was able to rank 28% of the buried cysteines as first or second most ligandable in the queried targets. The success rate of cysteine ligandability detection of the algorithm on the queried dataset is slightly higher when taking the probability score it assigns to each cysteine (by setting 3 bins for clarity: high for probability score 1.0-1.2, Medium for probability score of 0.8-1.0 and Low for less than 0.8). Details and exact score can be found in the SupplementaryFile1 (**B**) The second approach uses a combination of chemoproteomics-based (empirical) covalent sites and in-silico ligandability assessment techniques presented in this work. In the combined approach, the simplistic ligandability approach was able to capture 78% of the ligandable cysteines compared to the advanced method which correctly judged 92% of the empirical sites.

Cysteine Ligandability Detection: Computational versus Combined Methodologies

Although various standalone residue reactivity prediction algorithms[32,59] take key ligandability features such as SASA and flexibility into account in detecting reactive residues, our experience suggests that protein conformation sampling and/or separate assessment for the presence a binding pocket in the vicinity of the predicted reactive residue site may be necessary for accurate results. Although the differentiating advantage of such an approach can be sensed from the data presented above, we further pursued a case in which we compared prediction performance of a purely computational method on a list of cysteine sites buried in an occluded binding pockets. A list of 47 apo-covPDB structures empirically confirmed to contain a buried ligandable cysteine covalent site were selected and tested for ligandability prediction by DeepCoSI[32] (apo-covPDBs_buriedCys, SupplementaryFile1). DeepCoSI performs global analysis of all cysteine residues on each query PDB file and ranks the cysteines according to their "ligandability" score. Albeit considerate of SASA around the target residue, the DeepCoSI platform was able to capture at least 28% (high accuracy ranked) and at most 76% of the ligandable cysteines (the total of high and medium accuracy predictions using ranking-based scores). In contrast to ranking-based score, the proportion of top predicted ligandable cystines based on probability score of the algorithm matched that of the empirical data at least 19% and at most 40% of the time. (Figure 3). Treating the same apo-covPDB samples as empirical cysteine chemoproteomics data points and assessing them for ligandability using the simplistic and advanced approaches correctly assessed 78% and 92% of the ligandable cysteines, respectively. As exemplified, covalent ligandability evaluation techniques purely dependent on computational methods may misrepresent a real ligandable site whereas that which uses the simplistic rigid protein models for site ligandability evaluation may fail to detect hidden or novel sites. By maximizing the number of correct predictions, the advanced ligandability approach which considers protein conformational fluctuations in assessing ligandable cavities around experimental sites is an ideal method to be adopted in site-directed and covalent drug discovery pipelines.

Discussion

The identification of covalent site ligandability through machine learning presents a tremendous opportunity to expand the druggable protein space, especially those lacking well-defined binding sites. Our study aimed to introduce and validate an advanced computational platform, leveraging chemoproteomics studies, to accelerate covalent-based drug discovery processes. This platform integrates protein conformational fluctuations and structural dynamics, providing a robust scaffold for assessing and validating experimental chemoproteomics sites and identifying potential ligands for drug development.

In the evaluation of ligandability using a simplistic approach based on a distance threshold from pocket surfaces, our initial findings revealed a competent detection performance, particularly in detecting empirical holo sites. However, the simplistic approach exhibited limitations in identifying ligandable sites within apo structures, where structural irregularities and lower solvent accessible surface areas challenged the detection algorithm's efficacy. These results highlight the inherent challenges faced by algorithms trained on holo pockets when applied to detect valid sites within apo structures, emphasizing the need for methodologies that account for protein dynamics and conformational changes.

The majority of the human genome remains undrugged encoding thousands of proteins that have experimental evidence linking them to human disease. With only about 5% of the human proteome drugged and multiple targeted drugs available for a small number of driver gene targets, the vast majority of driver gene targets (>3500 for cancer driver) remain untargeted and/or inaccessible.[60,61] These proteins which need to be immediately explored for *de novo drug* discovery, represent a potential of multi-fold increase beyond the total number of available FDA-approved drugs existing today. Whereas membrane proteins make-up about 30% of all known proteins, over 60% of the current drug targets are membrane proteins.[39] The difficulty associated with targeting many of these proteins is the absence of well defined binding sites or lack of target structural information, albeit the latter is rapidly changing with recent advancement in cryo-EM techniques. The advanced covalent ligandability assessment platform, designed to consider protein conformational fluctuations, exhibited remarkable performance gains in the detection of covalent ligandable sites, surpassing the simplistic approach both in detecting holo sites and significantly outperforming in identifying apo sites. Notably, the advanced platform showcased a reduced rate of missed sites for both holo and apo structures compared to the simplistic approach, underscoring the pivotal role of considering protein structural fluctuations in

enhancing the accuracy of ligandability predictions. The platform also enables mass-scale and holistic analysis of proteomics data with unprecedented speed (<1 second/site) using a multidisciplinary approached aimed at expediting drug discovery research and development. Furthermore, the computational platform designed in this study can be tailored for detection of non-covalent sites by providing platform with mock chemoproteomics sites proximal to expected non-covalent binding sites.

Chemical proteomics-based drug discovery research typically suggests a large number (1000s) of targets of which 1 to 5 potential targets are selected based on target quality for tractability analysis using wet-lab biochemical and biophysical validation assays. Target selection and prioritization is typically done using computational biology approaches that qualify targets using curated omics data available to us. Target prioritization is a long-standing issue in drug discovery research especially due to lack of techniques that assess binding site quality and downstream functional impact of targeting each site[62]. Target validation is highly time consuming (est. >1year/target) and is a monetary sink (est. >$1M/target validation) in the covalent drug discovery pipelines. Quite often, such tests eventually indicate that one or none of the elected targets are true therapeutic targets and are indeed identified as off-targets and/or false positives. Aside from being highly time and resource wasteful, this traditional approach of target validation misses on a vast opportunity that the remaining unvalidated targets pose. Ligandability evaluation techniques showcased in this work provide an opportunity to explore more than 75% of known proteome (767,580 protein structures)[63,64] using chemoproteomics-enabled covalent drug discovery experiments on a time and financial scale that is at least 100-fold less than the traditional approaches (Figure 4). To ensure the platform's efficacy in eliminating false-positive covalent sites, computational tools were employed to identify and disqualify superficial matching-negative residues in apo-covPDB structures. The comprehensive assessment revealed the platform's robustness in minimizing false-positive rates, crucial for ensuring the accuracy of ligandability predictions.

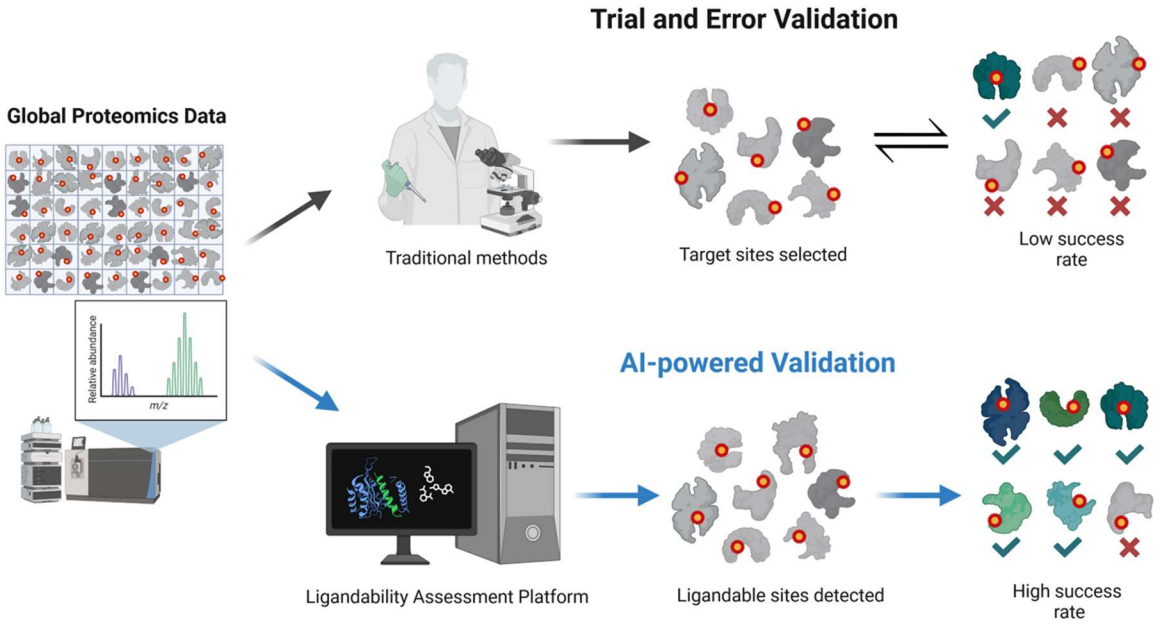


Figure 4: The traditional approach of target prioritization involves an exhaustive feedback loop of target selection and in-vitro validation, quite often leading to pursuance of many false-positive hits. In the advanced approach powered by AI-based computational platforms, target sites are assessed for ligandability and downstream functionality in-silico before attempting any in-vitro validation. Although targets picked by an algorithm are still bound for in-vitro validation, the number of necessary validation and their success rate is highly improved.

Application of such ligandability assessment platforms can be very broad, ranging from development of novel QMS-based techniques and site-directed fragment-based screening to identifying variants adjacent to novel ligandable pockets for allele-specific drug binding[65]. Our results indicated that as of today, a conglomeration of empirical and computational methods may provide the highest research throughput. We report for the first time a list of apo counterparts of covalent-ligand bound PDB structures, which are of highly value for further development of platforms that can accurately recognize both apo and holo covalent sites. Although the work here is focused on a platform more suitable for a bottom-up drug discovery approach, the data reported in this paper can be utilized in training and developing singly-sufficient computational platforms that, through a top-down approach, accurately predict valid covalent ligandable sites on target proteins (Suppl. Figures 3 and 4).

Moreover, the platform demonstrated its proficiency in identifying cryptic binding sites, which typically remain hidden due to structural transformations induced by ligand binding. The advanced platform's ability to detect a diverse range of conformational changes leading to the formation of cryptic pockets was significantly higher compared to the simplistic approach, highlighting its capability to unveil hidden ligandable sites while maintaining a low false-positive rate.

In conclusion, the developed computational platform, by harnessing protein conformational sampling, represents a paradigm shift in covalent site ligandability evaluation. Its ability to consider dynamic structural changes, detect cryptic pockets, minimize false-positive rates, and enhance the accuracy of ligandability predictions stands as a significant advancement in computational biology. This platform holds immense promise in expediting drug discovery research and development by providing a holistic and multidisciplinary approach for identifying potential therapeutic targets and facilitating the design of novel covalent drugs.

Acknowledgements

We thank the IT department of the University of Toronto Mississauga for their support in providing infrastructure for data storage. We thank the Department of Mathematical and Computational Sciences at the University of Toronto Mississauga for providing access to computational talent and resources. We especially thank our data analysis volunteers Hrithik Kumar Advani, Sibthain Kazmi, Ahmed Nasir, and advisors Tudor B. Radu, Lisa Zhang and Sirano Dhe-Paganon for their valuable support and commitment to the development of this project. We thank Digital Research Alliance of Canada for providing access to supercomputing resources and supporting the development of this project.

P.T.G. is supported by research grants from NSERC (RGPIN-2014-05767), CIHR (MOP-130424, MOP-137036), Canada Research Chair (950-232042), Canadian Cancer Society (703963), Canadian Breast Cancer Foundation (705456) and infrastructure grants from CFI (33536) and the Ontario Research Fund (34876).

Notes

Visualizations are generated using Biorender.com platform.

The authors declare the following competing financial interest(s): F.E., is co-founder and chief scientific officer of Gene2Lead, A.D.E, and P.T.G are co-founders of Gene2Lead.

References

1. Westbrook, J. D. & Burley, S. K. How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals. *Structure* **27**, 211–217 (2019).
2. De Vita, E. 10 years into the resurgence of covalent drugs. *Future Med Chem* **13**, 193–210 (2021).
3. Singh, J., Petter, R. C., Baillie, T. A. & Whitty, A. The resurgence of covalent drugs. *Nat Rev Drug Discov* **10**, 307–317 (2011).
4. Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angewandte Chemie International Edition* **55**, 13408–13421 (2016).
5. Abdeldayem, A., Raouf, Y. S., Constantinescu, S. N., Moriggl, R. & Gunning, P. T. Advances in covalent kinase inhibitors. *Chem. Soc. Rev.* **49**, 2617–2687 (2020).
6. Hallin, J. *et al.* The KRASG12C Inhibitor MRTX849 Provides Insight toward Therapeutic Susceptibility of KRAS-Mutant Cancers in Mouse Models and Patients. *Cancer Discov* **10**, 54–71 (2020).
7. Canon, J. *et al.* The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).
8. Gehringer, M. & Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J Med Chem* **62**, 5673–5724 (2019).
9. Bradshaw, J. M. *et al.* Prolonged and tunable residence time using reversible covalent kinase inhibitors. *Nat Chem Biol* **11**, 525–531 (2015).
10. Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **49**, 377–389 (2009).

11. Marx, M. A. *et al.* Abstract B30: Structure-based drug discovery of MRTX1257, a selective, covalent KRAS G12C inhibitor with oral activity in animal models of cancer. *Molecular Cancer Research* **18**, B30 (2020).
12. Feng, Y., Duan, W., Cu, X., Liang, C. & Xin, M. Bruton's tyrosine kinase (BTK) inhibitors in treating cancer: a patent review (2010-2018). *Expert Opinion on Therapeutic Patents* **29**, 217–241 (2019).
13. da Cunha-Bang, C. & Niemann, C. U. Targeting Bruton's Tyrosine Kinase Across B-Cell Malignancies. *Drugs* **78**, 1653–1663 (2018).
14. Byrd, J. C. *et al.* Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *N Engl J Med* **369**, 32–42 (2013).
15. Hoffer, L. *et al.* CovaDOTS: In Silico Chemistry-Driven Tool to Design Covalent Inhibitors Using a Linking Strategy. *J Chem Inf Model* **59**, 1472–1485 (2019).
16. Barf, T. & Kaptein, A. Irreversible protein kinase inhibitors: balancing the benefits and risks. *J Med Chem* **55**, 6243–6262 (2012).
17. Raouf, Y. S. *et al.* Discovery of YSR734: A Covalent HDAC Inhibitor with Cellular Activity in Acute Myeloid Leukemia and Duchenne Muscular Dystrophy. *J. Med. Chem.* (2023) doi:10.1021/acs.jmedchem.3c01236.
18. Petri, L. *et al.* A covalent strategy to target intrinsically disordered proteins: Discovery of novel tau aggregation inhibitors. *Eur J Med Chem* **231**, 114163 (2022).
19. Lonsdale, R. & Ward, R. A. Structure-based design of targeted covalent inhibitors. *Chem Soc Rev* **47**, 3816–3830 (2018).
20. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).

21. Hu, S., Chen, P., Gu, P. & Wang, B. A Deep Learning-Based Chemical System for QSAR Prediction. *IEEE J Biomed Health Inform* **24**, 3020–3028 (2020).
22. Muegge, I. & Oloff, S. Advances in virtual screening. *Drug Discov Today Technol* **3**, 405–411 (2006).
23. Zhu, K. *et al.* Docking Covalent Inhibitors: A Parameter Free Approach To Pose Prediction and Scoring. *J. Chem. Inf. Model.* **54**, 1932–1940 (2014).
24. Rachman, M. *et al.* DUckCov: a Dynamic Undocking-Based Virtual Screening Protocol for Covalent Binders. *ChemMedChem* **14**, 1011–1021 (2019).
25. Wei, L. *et al.* Cov_DOX: A Method for Structure Prediction of Covalent Protein–Ligand Bindings. *J. Med. Chem.* **65**, 5528–5538 (2022).
26. Scarpino, A. *et al.* WIDOCK: a reactive docking protocol for virtual screening of covalent inhibitors. *J Comput Aided Mol Des* **35**, 223–244 (2021).
27. Bianco, G. *et al.* Reactive Docking: A Computational Method for High-Throughput Virtual Screenings of Reactive Species. *J. Chem. Inf. Model.* **63**, 5631–5640 (2023).
28. Palazzesi, F. *et al.* BIreactive: A Machine-Learning Model to Estimate Covalent Warhead Reactivity. *J Chem Inf Model* **60**, 2915–2923 (2020).
29. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics* **10**, 39 (2018).
30. An, J., Totrov, M. & Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* **4**, 752–761 (2005).
31. Meller, A. *et al.* Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* **14**, 1177 (2023).

582 32. Du, H. *et al.* Proteome-Wide Profiling of the Covalent-Druggable Cysteines with a
583 Structure-Based Deep Graph Learning Network. *Research (Wash D C)* **2022**, 9873564
584 (2022).
585 33. Wang, H. *et al.* Sequence-Based Prediction of Cysteine Reactivity Using Machine Learning.
586 *Biochemistry* **57**, 451–460 (2018).
587 34. Gao, M. & Günther, S. HyperCys: A Structure- and Sequence-Based Predictor of Hyper-
588 Reactive Druggable Cysteines. *International Journal of Molecular Sciences* **24**, 5960 (2023).
589 35. Piazza, I. *et al.* A machine learning-based chemoproteomic approach to identify drug targets
590 and binding sites in complex proteomes. *Nat Commun* **11**, 4200 (2020).
591 36. Lanning, B. R. *et al.* A road map to evaluate the proteome-wide selectivity of covalent
592 kinase inhibitors. *Nat Chem Biol* **10**, 760–767 (2014).
593 37. Browne, T., Concheiro-Guisan, M. & Prinz, M. Semi quantitative detection of signature
594 peptides in body fluids by liquid chromatography tandem mass spectrometry (LC–MS/MS).
595 *Forensic Science International: Genetics Supplement Series* **7**, 208–210 (2019).
596 38. Lanman, B. A. *et al.* Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the
597 Treatment of Solid Tumors. *J Med Chem* **63**, 52–65 (2020).
598 39. Maveyraud, L. & Mourey, L. Protein X-ray Crystallography and Drug Discovery. *Molecules*
599 **25**, 1030 (2020).
600 40. Boatner, L. M., Palafox, M. F., Schweppe, D. K. & Backus, K. M. CysDB: a human cysteine
601 database based on experimental quantitative chemoproteomics. *Cell Chem Biol* S2451-
602 9456(23)00090–9 (2023) doi:10.1016/j.chembiol.2023.04.004.

41. Mortenson, D. E. *et al.* “Inverse Drug Discovery” Strategy To Identify Proteins That Are Targeted by Latent Electrophiles As Exemplified by Aryl Fluorosulfates. *J. Am. Chem. Soc.* **140**, 200–210 (2018).
42. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem Rev* **113**, 2343–2394 (2013).
43. Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V. & Priyakumar, U. D. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c00799.
44. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
45. de Groot, B. L. *et al.* Prediction of protein conformational freedom from distance constraints. *Proteins* **29**, 240–251 (1997).
46. Yang, Y., Kucukkal, T. G., Li, J., Alexov, E. & Cao, W. Binding Analysis of Methyl-CpG Binding Domain of MeCP2 and Rett Syndrome Mutations. *ACS Chem Biol* **11**, 2706–2715 (2016).
47. Patschull, A. O. M., Gooptu, B., Ashford, P., Daviter, T. & Nobeli, I. In Silico Assessment of Potential Druggable Pockets on the Surface of α1-Antitrypsin Conformers. *PLoS One* **7**, e36612 (2012).
48. Zheng, H., Hou, J., Zimmerman, M. D., Wlodawer, A. & Minor, W. The future of crystallography in drug discovery. *Expert Opin Drug Discov* **9**, 125–137 (2014).
49. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

50. Gao, M., Moumbock, A. F. A., Qaseem, A., Xu, Q. & Günther, S. CovPDB: a high-resolution coverage of the covalent protein-ligand interactome. *Nucleic Acids Res* **50**, D445–D450 (2022).
51. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
52. Cimermancic, P. *et al.* CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology* **428**, 709–719 (2016).
53. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
54. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nat Rev Drug Discov* **15**, 533–550 (2016).
55. Sijbesma, E. *et al.* Site-Directed Fragment-Based Screening for the Discovery of Protein-Protein Interaction Stabilizers. *J Am Chem Soc* **141**, 3524–3531 (2019).
56. Romany, A., Liu, R., Zhan, S., Clayton, J. & Shen, J. Analysis of the ERK Pathway Cysteinome for Targeted Covalent Inhibition of RAF and MEK Kinases. *J Chem Inf Model* **63**, 2483–2494 (2023).
57. Harris, R. C., Liu, R. & Shen, J. Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant pH Molecular Dynamics in Amber. *J Chem Theory Comput* **16**, 3689–3698 (2020).
58. Marino, S. M. & Gladyshev, V. N. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *J Mol Biol* **404**, 902–916 (2010).
59. Ustach, V. D. *et al.* Optimization and Evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a Tool for Target-Based Ligand Optimization. *J. Chem. Inf. Model.* **59**, 3018–3035 (2019).

60. Warner, K. D., Hajdin, C. E. & Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat Rev Drug Discov* **17**, 547–558 (2018).
61. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).
62. K. Brown, K. *et al.* Approaches to target tractability assessment – a practical perspective. *MedChemComm* **9**, 606–613 (2018).
63. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
64. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
65. Nichols, C. A. *et al.* Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nat Commun* **11**, 2517 (2020).