An Integrated Covalent Drug Design Workflow using Site-Identification by Ligand Competitive Saturation

Wenbo Yu[1,2,3], David J. Weber[2,3] and Alexander D. MacKerell, Jr.[1,2,3*]*

[1]Computer-Aided Drug Design Center, Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland [2]Institute for Bioscience and Biotechnology Research (IBBR), Rockville, Maryland [3] Center for Biomolecular Therapeutics (CBT), School of Medicine, University of Maryland, Baltimore, Maryland

*Correspondence: wyu@rx.umaryland.edu and alex@outerbanks.umaryland.edu

Abstract

Covalent drug design is an important component in drug discovery. Traditional drugs interact with their target in a reversible equilibrium while irreversible covalent drugs increase the drug-target interaction duration by forming a covalent bond with targeted residues, and thus may offer a more effective therapeutic approach. To facilitate the design of this class of ligands, computational methods can be used to help identify reactive nucleophilic residues, frequently cysteines, on a target protein for covalent binding, to test various warhead groups for their potential reactivity, and to predict non-covalent contributions to binding that can facilitate drug-target interactions that are important for binding specificity. To further aid covalent drug design, we extended a functional group mapping approach based on explicit solvent all-atom molecular simulations (SILCS: Site Identification by Ligand Competitive Saturation) that intrinsically considers protein flexibility, functional group and protein desolvation along with functional group-protein interactions. Through docking of a library of representative warhead

fragments using SILCS-Monte Carlo (SILCS-MC), reactive cysteines can be correctly identified for proteins being tested. Furthermore, a machine learning model was trained to quantify the effectiveness of various warhead groups for proteins using metrics from SILCS-MC as well as experimental model compound warhead reactivity data. The ability to rank covalent molecular binders with similar warheads using SILCS ligand grid free energy (LGFE) ranking was also tested for several proteins. Based on these tools, an integrated SILCS based workflow was developed, named SILCS-Covalent, that can both qualitatively and quantitatively inform covalent drug discovery.

1. Introduction

Traditionally, drug molecules are designed to bind a specific biological target in equilibrium through noncovalent interactions such as hydrogen bonding, ionic bonding and hydrophobic interactions ***(1, 2)***. Covalent drugs take the drug-target interaction to another level by introducing chemical reactivity that enables covalent bond formation between drug and target ***(3, 4)***. Covalent drugs have been on the market since the late 19th century when the anti-inflammatory agent Aspirin was developed by Bayer, even though the covalent binding mechanism remained unrevealed at that time ***(5)***. Given the success and widespread use of aspirin, as well as more recent successes presented below, covalent drug design is now a widely used strategy in the development of new therapeutic agents.

By significantly shifting the binding equilibrium toward the inhibitor-protein complex product, covalent bond formation can help to improve drug effectiveness ***(3, 4)***. However, this approach was discouraged historically due to safety concerns related to potential off-target reactivities ***(6)***. The resurgence of covalent drugs after entering the 21th century was mainly accelerated by the concept of 'targeted covalent inhibitors' (TCIs) that combine both covalent and noncovalent components ***(6, 7)***. Instead of directly searching for molecules based on binding affinity alone, the design of TCIs involves multiple steps composed of identification of reactive residues on the target, selection of electrophilic functional groups or so-called warheads, as well as optimization of the non-reactive scaffold for binding site affinity and specificity ***(8)***. Such a finely tuned combination of a weakly electrophilic warhead with

site specific noncovalent interactions can yield drugs that have enhanced duration of action and potency as well as increased selectivity over noncovalent drugs ***(6-8)***. Successful examples include the Bruton's tyrosine kinase (BTK) inhibitor Ibrutinib targeting B-cell cancers ***(9)*** as well as the recently approved Nirmatrelvir targeting SARS-Cov-2 main protease, which is contributing to therapeutically expedite the ending of Covid-19 pandemic ***(10)***.

Like noncovalent drug design, computer-aided drug design (CADD) approaches ***(11)*** have become an indispensable component of covalent drug design. Covalent ligand design methods can predict the binding mode and affinity of covalent ligands mostly by integrating conventional noncovalent docking and scoring schemes with additional sampling and scoring treatment in order to model the newly formed covalent bond at the reaction site ***(12)***. Unlike noncovalent docking where the binding site is usually treated as rigid or with limited flexibility, covalent docking needs to handle local conformational changes of the target protein caused by covalent bond formation introducing additional computational cost compared to conventional docking. And most covalent docking schemes rely on conventional docking scores that are usually trained against noncovalent binding data where covalent bond formation is ignored. Examples include CovDock from Schrodinger ***(13)*** and DOCKovalent developed by London and colleagues, ***(14)*** among others ***(12)***. The apparent binding score used by CovDock combines conventional Glide scores ***(15)*** for the binding poses sampled before and after covalent bond formation ***(13)***, while DOCKovalent ***(14)*** depends on the conventional DOCK3.6 scoring function ***(16)***. In order to account for covalent bond, quantum mechanical (QM) calculations may be applied; one recent example is COV_DOX ***(17)*** that was shown to outperform all tested molecular mechanics (MM) based covalent docking methods. However, the method comes with the price of reduced computational efficiency, which limits its routine use. An alternative is direct noncovalent docking methods ***(12)*** and a recent comparative evaluation study showed such methods to have performance similar to covalent docking methods but with a significantly lower computing expense ***(18)***.

Beyond ligand design, an important aspect of covalent drug design is to predict reactive nucleophilic residues that can accommodate covalent binders. This is usually done by predicting residue pK_a as it

correlates with nucleophilicity, where a downshift in the pK_a in the protein environment compared to its free state value generally implies a higher reactivity ***(19)***. Multiple tools for pK_a prediction are in existence such as PROPKA, ***(20)*** among others. However, these methods typically employ implicit solvent models contributing to large inaccuracies ***(21)***. Constant pH molecular dynamics (MD) simulation methods using explicit solvent can predict pK_a to a higher level of accuracy but with a high computational cost ***(21, 22)***.

The site identification by ligand competitive saturation (SILCS) method ***(23)*** is a functional group mapping approach using explicit solvent all-atom oscillating excess chemical potential, μ_{ex}, Grand Canonical Monte Carlo (GCMC)/MD simulations ***(24)*** to model the interactions of selected probe solute molecules as well as water with a target protein. Probability distributions of the solute functional groups from SILCS simulations are normalized and Boltzmann weighted yielding a 3D free energy affinity pattern, termed grid free energy (GFE) FragMaps ***(25, 26)***. Such GFE FragMaps include contributions from protein-functional group interactions, protein flexibility, and desolvation of both the functional groups and the protein and have been used in various CADD applications ***(27-33)***. SILCS was shown to have better performance in terms of pharmacophore based virtual screening than other empirical docking methods ***(27, 28)*** and in the context of binding affinity predictions, it was verified to yield comparable outcomes with computationally expensive free energy perturbation (FEP) methods ***(32, 33)***.

In this work, we extend the SILCS method in the context of covalent drug design. The pre-computed SILCS GFE FragMaps hold the promise of addressing both accuracy and efficiency. Using the information content in the FragMaps and probability distributions of the sulfur atoms in cysteine residues along with the implementation of a machine learning (ML) model, we address i) the identification of reactive cysteine residues, ii) the selection of warheads targeting those residues, and iii) optimization of the non-covalent portion of the ligands to improve affinity. The study involves analysis of results on a training set of proteins for initial model development followed by validation on selected proteins for which experimental data is available. From these efforts, an integrated computational workflow for covalent drug design targeting cysteine residues is presented, termed SILCS-Covalent.

2. Methods

Cysteine, serine, threonine, tyrosine and lysine are all nucleophilic amino acid residues in a protein that can be pursued for covalent binding purposes. Due to the high nucleophilic property of thiolate and its noncatalytic role in many proteins, cysteine (Cys) is the most targeted amino acid in covalent drug development and most Food and Drug Administration (FDA) approved covalent drugs target Cys. Thus, the present study focuses on Cys related covalent drug design. Instead of explicitly considering bond formation and the product state, the developed noncovalent docking protocol involves Monte Carlo sampling in the fields of the SILCS GFE FragMaps, termed (SILCS-MC) ***(25, 26)***, is used to predict and optimize the affinity and specificity of irreversible inhibitors as well as to identify targetable Cys residues. In addition, the physics based SILCS method is supplemented with ML models for determination of optimal warhead functional groups targeting tractable residues.

2.1 Protein systems for test and structure preparation

Initially, six kinase protein systems from various branches of the kinase phylogenetic tree ***(34)*** were selected to explore the ability of SILCS for targetable Cys residue identification. A previous experimental work conducted a large-scale electrophilic warhead screen against all six proteins ***(34)***, which can serve as a useful benchmark to test the ability of SILCS for the determination of optimal warhead groups. The proteins include BTK, extracellular signal-regulated kinase 2 (ERK2), p90 ribosomal S6 kinase 2 (RSK2), mitogen-activated protein kinase kinase 6 (MAP2K6), Janus kinase 3 (JAK3) and Maternal embryonic leucine zipper kinase (MELK). After the initial test, we selected three non-kinase proteins for validation of the method against alternate protein classes, as indexed in the CovPDB database ***(35)***. They include cathepsin K (CATK), a hydrolase protein, interleukin-2 (IL-2), a signaling protein, and glutathione S-transferase omega-1 (GSTO1-1), a transferases protein.

The initial crystal structures were taken from the Protein Data Bank (PDB) ***(36)*** as follows: BTK (5p9j), ERK2 (4zzm), RSK2 (4d9u), MAP2K6 (3vn9), JAK3 (5lwm), MELK (5ih9), CATK (7nxm), IL-2 (1m4b) and GSTO1-1 (5yvo). For *holo* structures, the ligands were removed. Except for missing regions in the N- and C-termini, missing residues in short loops in the crystal structures were modeled using SWISS-

MODEL ***(37)***. This includes residues 553 to 555 for BTK and residues 74 to 82 and 99 to 102 for IL-2. All processed protein structures were submitted to the CHARMM-GUI ***(38)*** to generate coordinates for missing atoms based on internal coordinates. For example, coordinates of atoms CG, CD, CE and NZ in both residues K147 and K191 for CATK were missing in the crystal structure.

For the MAP2K6 crystal structure, since the missing N-terminus contains a Cys residue at the 38 position, the AlphaFold protein structure database ***(39)*** was utilized to generate a structure for the missing N-terminus residues 33 to 43. MAP2K6 structure (UniProt entry P52564) predicted by AlphaFold ***(40)*** was downloaded from the AlphaFold protein structure database hosted by the European Bioinformatics Institute at the European Molecular Biology Laboratory (EMBL-EBI) ***(39)***. The AlphaFold pLDDT score ***(40)*** for residues 33 to 43 ranges from ~ 53 to ~71, which implies a low model confidence, thus extra MD simulations were conducted for MAP2K6 to equilibrate the structure. MAP2K6 protein was solvated in a water box, the size of which was determined to have the protein extrema separated from the box edge by a minimum of 10 Å on all sides and ions were added to neutralize the full system. CHARMM36m protein force field ***(41, 42)*** and CHARMM modified TIP3P water model ***(43)*** were used to describe protein and water during the simulations, respectively. MD was conducted using the GROMACS program ***(44)***. The system was first minimized for 50,000 minimization steps with the steepest descent (SD) algorithm ***(45)*** in the presence of periodic boundary conditions (PBC) and harmonic restrains on protein backbone Cα carbon atoms with a force constant (k in 1/2 $k\delta x^2$) of 0.12 kcal/mol/Å to mainly relax the solvent. The second minimization was conducted with the same setup but with removal of restraints on protein residues 33 to 43 to relax the AlphaFold structure and then followed by a third minimization with the same setup but with removal of restraints on all protein residues to further relax both the solute and solvent. The minimized system was then subject to a MD equilibration under isochoric-isothermal (NVT) canonical ensemble for 500 ps with harmonic restrains on protein Cα atoms and followed by isothermal-isobaric (NPT) ensemble MD for another 500 ps to adjust the PBC box size at 300K and 1atm. Another 500ps NPT MD simulation was conducted without any restrains to further equilibrate the whole system. The

final production MD run was conducted for 20 ns for further equilibration and the structure from the last frame of the trajectory was extracted for the SILCS simulation.

2.2 SILCS simulations and map generation

SILCS simulations were conducted using the previously described protocol ***(32)*** for the 9 proteins. The simulation systems involved protein, water and eight solute molecules including benzene, propane, methanol, formamide, dimethylether, imidazole, methylammonium, and acetate. Ten individual simulation systems for each protein with explicit water and randomly positioned solutes at approximately 0.25 M each were generated and simulated independently for better convergence. Initial equilibration of each SILCS system involved 5,000 steps of energy minimization using the SD method ***(45)*** followed by 100 ps MD equilibration at 300 K using the velocity rescaling thermostat with randomized initial velocities and the Berendsen barostat ***(46)*** to allow for system volume relaxation. The equilibrated systems were then subjected to 25 cycles of GCMC comprising 200,000 MC steps to redistribute water and solute molecules. The GCMC simulations are based on a previously described protocol where an oscillating excess chemical potential, μ_{ex}, is applied to increase the insertion acceptance efficiency ***(24)***. The final coordinates from this procedure were fed to the production run of 100 cycles of iterative GCMC/MD protocol where GCMC drives the sampling of the solutes and water with all solutes, water and protein atoms subsequently propagated in the MD simulations. Each GCMC/MD cycle consists of 200,000 GCMC steps, followed by 1 ns of production MD run which yield a cumulative 200 million steps of GCMC and 1,000 ns MD production time over all ten systems. Weak harmonic restraints with a force constant (k in 1/2 $k\delta x^2$) of 0.12 kcal/mol/Å were applied to all protein Cα atoms during MD simulations. A time step of 2 fs was used, and the protein conformations and distributions of water and solutes were saved every 10 ps for analysis. Temperature and pressure were maintained using the Nosé–Hoover thermostat ***(47, 48)*** and the Parrinello–Rahman barostat ***(49)***, respectively. The GCMC simulations were conducted using in-house developed code in the SILCS software suite, version 2022.1 (SilcsBio LLC) and MD simulations were performed using GROMACS, with the protein, solutes and water being described using the CHARMM36m protein force field ***(41, 42)***, the CHARMM General force field

(CGenFF) ***(50, 51)*** and the TIP3P water model modified for the CHARMM force field ***(43)***, respectively.

FragMaps were generated by binning selected solute atoms into voxels on a 1 Å spaced grid spanning the simulation system and were combined to obtain both specific and generic FragMap types as previously described ***(32)***. 3D normalized probability distributions were obtained by normalizing the voxel occupancies computed in the presence of the protein by the respective values of the solutes alone in aqueous solution based on their average number relative to that of water, which was assumed to have a 55 M concentration. The normalized distributions were Boltzmann transformed to free energies for each functional group type to yield GFE FragMaps. Four generic FragMaps were generated including APOLAR (benzene and propane carbons), generic heterocycle carbon GEHC (imidazole carbons), HBDON (formamide and imidazole donor nitrogen) and HBACC (formamide and dimethylether oxygens and imidazole acceptor nitrogen) maps. In addition, four specific FragMaps were used including positive MAMN (methylammonium nitrogen) maps, negative ACEC (acetate carboxylate carbon) maps, alcohol MEOO (methanol oxygen) maps, and FORC (formamide carbon) maps. Alcohol maps are used as these functional groups can act as both hydrogen bond donors and acceptors complicating their inclusion in the generic HBDON or HBACC maps. And GEHC and FORC carbon maps are used explicitly as such carbons that are adjacent to polar atoms have different physiochemical properties than apolar carbons. Exclusion maps that represent the solute/water forbidden region during SILCS simulation were also generated. In addition to SILCS FragMaps, probability maps for Cys thiol sulfur atoms ($Prob_S$) were generated on the same 1 $Å^3$ grid to account for the flexibility of Cys residues during the MD simulations. Thiol sulfur atom occupancy values were normalized by the total number of MD frames yielding a 0 to 1 scale probability value for each Cys residue, and the sum of probability values for all voxels is equal to the number of Cys residues in the target protein.

2.3 SILCS-MC for covalent ligand reaction-competent stage docking

SILCS-MC sampling under exhaustive mode was used to dock 1) an electrophilic warhead library for development of an optimal warhead selection approach or 2) known covalent ligands for binding affinity

prediction. The docking is guided by intramolecular energies and the ligand grid free energy (LGFE) score. To calculate the LGFE, a GFE score is assigned to each classified atom in a molecule based on an atom classification scheme (ACS) with the summation of the GFE scores over all the classified atoms yielding the LGFE. The generic ACS as described previously ***(32)*** was used for the current study. The six SILCS-MC simulations were initialized from molecule conformations randomized in six spheres centered at six positions located 3.5 Å from the Cys sulfur atom in the six directions along the three axes (±X, ±Y and ±Z) as shown in Figure 1. A simulation radius of 6 Å was adopted for the warhead library considering their relatively small sizes while a 10 Å radius was used for the covalent ligands. For warhead library docking, the SILCS-MC simulations were conducted for all Cys residues in a protein while only designated Cys residue were targeted for the covalent ligands. Docking used the exhaustive mode where each molecule was subjected to an initial intramolecular energy minimization for 10,000 steps using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method ***(52)*** with a gradient tolerance of 3×10^{-8} kcal/mol/Å based on the intramolecular energy defined by the CGenFF energy function that includes a 4r dielectric constant for the electrostatic term. The minimized molecular structure was sampled through 10,000 steps of MC in the field of the GFE FragMaps with molecular translations, molecular rotations, and dihedral rotations sampled within a range of 0-1Å, 0-180°, and 0-180°, respectively. This was followed by simulated annealing (SA) from 300 to 0 K over 40,000 MC steps, where the molecule is allowed to sample within a range of 0-0.2 Å, 0-9°, and 0-9° for molecular translations, molecular rotations, and dihedral rotations, respectively. To assure the convergence of the MC sampling, six independent simulations were run for each molecule of up to 250 cycles where, after 50 cycles the sampling could exit based on a convergence criterion of 0.2 kcal/mol difference among the top three most favorable LGFE scores. Note that the standard SILCS-MC approach typically uses a value of 0.5 kcal/mol ***(32)***.

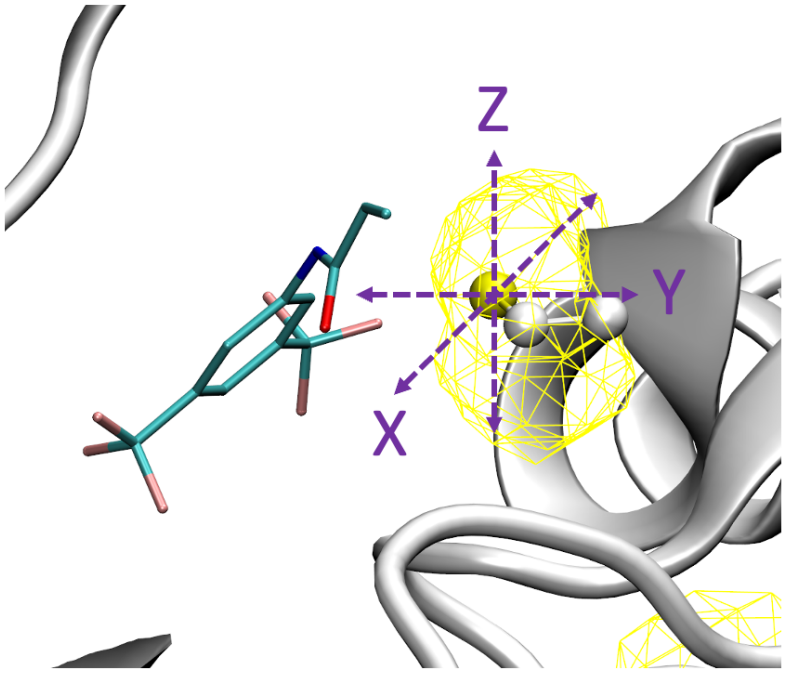

Figure 1. Illustration of SILCS-MC sampling around a Cys residue. The initial molecule configuration is randomized over six directions along the three axes near the Cys thiol sulfur atom (ball and stick representation for Cys sidechain and cartoon representation for target protein). The Cys thiol sulfur probability map is shown as a yellow mesh. An example is shown for a starting configuration of a warhead probe molecule for SILCS-MC generated in the -Y direction.

2.4 Metrics from SILCS-MC

Multiple SILCS metrics for model development were calculated. $LGFE_{Eq}$ is defined as the most favorable LGFE from SILCS-MC which represents the binding affinity associated with reversible equilibrium step. The more favorable the binding is during the first step, the longer the residency time thereby facilitating subsequent reorientation and covalent-bond formation. The contact probability, $Prob_{Cys}$ is defined as the overlap of a molecule's warhead reactive atom in a given SILCS-MC docking pose with the sulfur atom of the target Cys residue. $Prob_{Cys}$ values are only considered for ligand orientations with favorable interactions with the surrounding protein residues as determined by LGFE < 0. This quantifies the potential for a molecule to reorient itself for the reaction considering both geometric (close contacts between electrophilic and nucleophilic groups) and energetic (the molecule can interact

favorably with the protein) criteria. $Prob_{Cys}$ is evaluated for all unique ligand orientations from a SILCS-MC run using the Cys sulfur probability map as defined in equation 1 and illustrated in Figure 2.

$$Prob_{Cys} = \sum_{\substack{if\ warhead\ reactive\ atom\ is\ within\ d_{cutoff} \\ of\ a\ Cys\ S\ probability\ voxel \\ Cys\ S\ probability\ voxels}} Prob_S\ , \{if\ LGFE < 0\} \qquad (1)$$

Where, for each Cys sulfur probability voxel with a finite value, its distance to the warhead group reactive atom (the atom going to be covalently bonded to the Cys sulfur) in a docking pose with favorable LGFE is measured. If that distance is less or equal than a user defined cutoff value, d_{cutoff}, that voxel is recorded. Summation is performed over all the recorded probability voxels to get a contact probability for each docking pose. A $Prob_{Cys}$ value of 1 for a docked pose means all non-zero Cys sulfur probability voxels are within the user defined cutoff distance to the warhead group reactive atom in the docking pose. The variable d_{cutoff} is used to define the maximum distance at which an electrophilic atom to the Cys sulfur will have the potential to react. A previous QM/MM study showed that the distance between a thiolate sulfur and C_β in methyl vinyl ketone for the complex transition state is centered around 2.5 Å ***(53)***. However, since the Cys sulfur probability map, which represents a diffuse distribution instead of a fixed coordinate, is being used to define the reactive distance, a tight d_{cutoff} of 2.0 Å is adopted here. We also tested values from 2.0 to 3.0 Å with a step size of 0.2 Å, with the results being insensitive to the distance used. $LGFE_{Rx}$, the LGFE value associated with ligand orientations with reactive warhead atom to Cys sulfur probability distribution < d_{cutoff}, describes how favorably the ligand in reaction-competent orientation is interacting with the protein. For example, if a molecule samples conformations with both a high $Prob_{Cys}$ and a favorable $LGFE_{Rx}$, then it has a high potential for covalent bond formation.

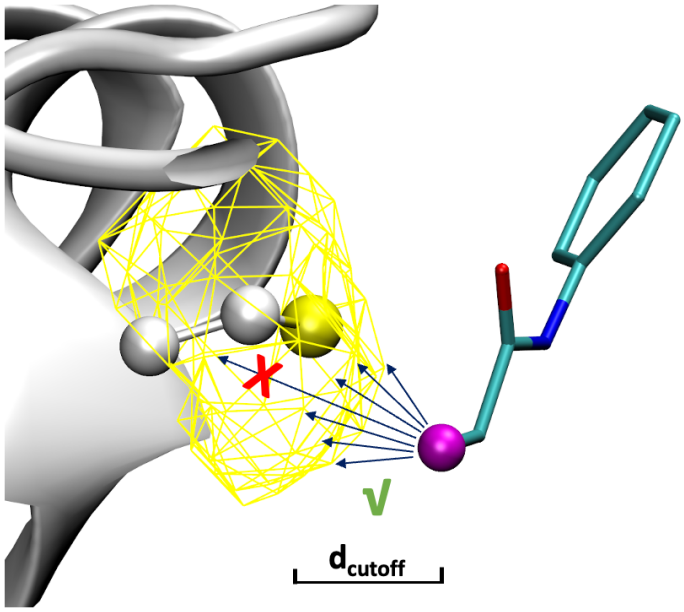

Figure 2. Calculation of the $Prob_{Cys}$ metric. For a ligand docking pose (licorice representation), the distances between the ligand reactive atom (the atom that will form the covalent bond with Cys as shown by a purple sphere representation) and Cys sulfur probability map grid (show as yellow mesh) are measured. Grid points within d_{cutoff} of the reactive atom will be counted toward the final $Prob_{Cys}$ value as shown by arrows with green mark. Grid points beyond d_{cutoff} will be discarded as illustrated by a red cross. Ball-and-stick representation is used for Cys sidechain and cartoon representation for target protein.

2.5 Warhead library and ligand data sets

To identify optimal warhead functional groups, a warhead library of a collection of electrophilic chemical groups is required. In the current work, the warhead library curated by Petri et. al. ***(34)*** that has 24 covalent probe molecules based on a noncovalent 3,5-bis(trifluoromethyl)phenyl scaffold, was adopted. This library covers a variety of representative electrophiles including acryl, nitrile, isothiocyanate, maleimide, halo-acetamide, thiol, and acetylene, among others. In addition, the library includes activity data against six kinases, that will be used for method development. Initial structures of the 24 probe molecules were generated using the Molecular Operating Environment (MOE) software ***(54)*** and minimized. For probe molecules 18, 19 and 21 which all have one chiral center, both isomers for each molecule were generated. During model development, metrics for both isomers were evaluated and the most favorable value was recorded to represent that molecule. Atom ID number of the reactive atom in each covalent warhead probe was also marked for evaluation of $Prob_{Cys}$. It should be noted that while

Petri's library ***(34)*** was used in the current proof-of-concept work, different libraries, for example, the Guo et al curated library of 113 warheads ***(55)*** could be used following the same protocol.

To validate the ability of SILCS-MC metrics to guide the optimization of the noncovalent scaffold portion of covalent ligands, covalent inhibitors which share the same or similar warhead group targeting a specific protein with experimental activity data available were used. Three such sets were selected from experimental studies targeting the BTK ***(56)***, RSK2 ***(14)*** and CATK ***(57)*** proteins. The number of ligands in the set ranges from 9 to 17. Initial structures of all covalent ligands were created with protonation states corresponding to pH 7.0 and minimized using MOE ***(54)***. Details about the test sets can be found in Table S1 in the supporting information.

2.6 Machine learning for classification of warhead functionalities

For ML model development, the warhead library includes experimental activity data for 24 warhead probes targeting 6 kinases, yielding 144 data points ***(34)***. As the experimental activity data is in the form of inhibition percentages that have high noise including, for example, negative values and activities of over 100%, pursing a regression model which tries to quantitatively reproduce the data was deemed to be inappropriate. Instead, a classification ML model was developed as it would be able to identify potential warhead groups for subsequent experimental testing. Accordingly, labels of high (H) and low (L) were assigned to all 144 data points (Table S3 in Ref ***(34)***) using an inhibition percentage cutoff of 50%, which turns the ML problem into a binary classification task. ML library Scikit-Learn for Python ***(58)*** was used for ML modeling. Six mainstream classification ML models were tested involving logistic regression (LR), support vector machine (SVM), K-nearest neighbors (KNN), decision tree (DT), random forest (RF) and naive Bayes (NB) using Scikit-Learn default hyperparameters. Four input features were used including the three SILCS-MC metrics $LGFE_{Eq}$, $LGFE_{Rx}$, and $Prob_{Cys}$ as well as the experimental metric $k_{GSH,}$ the GSH reactivity rate constant. The reported k_{GSH} values ***(34)*** were converted into ranking numbers to facilitate the ML training. This was due to the k_{GSH} values covering a wide range (0.001~>13.863 1/h) and including ambiguous values that were beyond the assay resolution. The assay minimal running time

window was 3 minutes, and some reaction rates were reported as being faster than 3 minutes. Ranking numbers were assigned to the following ranges as: 5 for >10 1/h rate, 4 for 1~10 1/h, 3 for 0.1~1 1/h, 2 for 0.01~0.1 1/h, and 1 for 0.001~0.01 1/h, such that higher ranking number indicates a faster reaction rate associated with higher intrinsic thiol reactivity. All input features were standardized in order to bring down all features to a common scale for efficient ML training. Outputs from the ML model will be 1 (for activity label "H") and 0 (for activity label "L"). The performance of all ML prediction models was evaluated using 5-fold cross validation. The training dataset was randomly divided into 5 subsets, wherein 4 subsets were used to train the model, and one remaining subset was used to validate the model. This process was repeated 5 times so that each fold is used once as the validation set and the average performance was reported as the performance of the model. Five performance metrics including accuracy, precision, recall, area-under-the-curve (AUC) and F1 score as described in the supporting information were reported to compare the different ML models. The final ML model was trained using all data points for future application purposes.

To further test the developed ML model, the protein glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was studied. The GAPDH crystal structure from PDB entry 1u8f with removal of nicotinamide-adenine-dinucleotide (NAD) was processed and used to initialize SILCS simulations. SILCS FragMaps and Cys sulfur probability maps were generated, and SILCS-MC simulations were performed for the 24 warhead fragments from the Petri's library targeting the catalytic Cys residue C152 in GAPDH. SILCS-MC metrics were then extracted from SILCS-MC simulations for all 24 warhead fragments and served as inputs for the ML model to predict warhead activity classes.

3. Results and Discussion

Multiple metrics extracted from SILCS-MC were tested for their potential to predict covalent ligand binding profiles. The steps involved in the covalent binding process are illustrated in Figure 3. During the binding process ***(3, 4)***, a ligand will first noncovalently interact with the protein in an equilibrium, $K_{eq} = k_{on}/k_{off}$, which in the context of binding free energies is modeled as $LGFE_{Eq}$. Subsequently, covalent bond

formation may occur, indicated by k_{inact}, which may be modeled as two consecutive processes. First the ligand can adjust its orientation within the binding site to allow its electrophilic warhead group to locate adjacent to the nucleophilic protein residue in a reactive mode, allowing for covalent bond formation to occur. These orientations, referred to as reaction-competent binding conformations, are identified based on the overlap of the warhead reactive functional group with the Cys S probability distribution. Ligand-protein affinities in these orientations are quantified through $LGFE_{Rx.}$ Final formation of the covalent bond then occurs based on the intrinsic reactivity of the different warheads determined experimentally, k_{GSH}. The k_{GSH} values have been measured experimentally through a glutathione (GSH) assay for the warheads included in the present study ***(34)***.

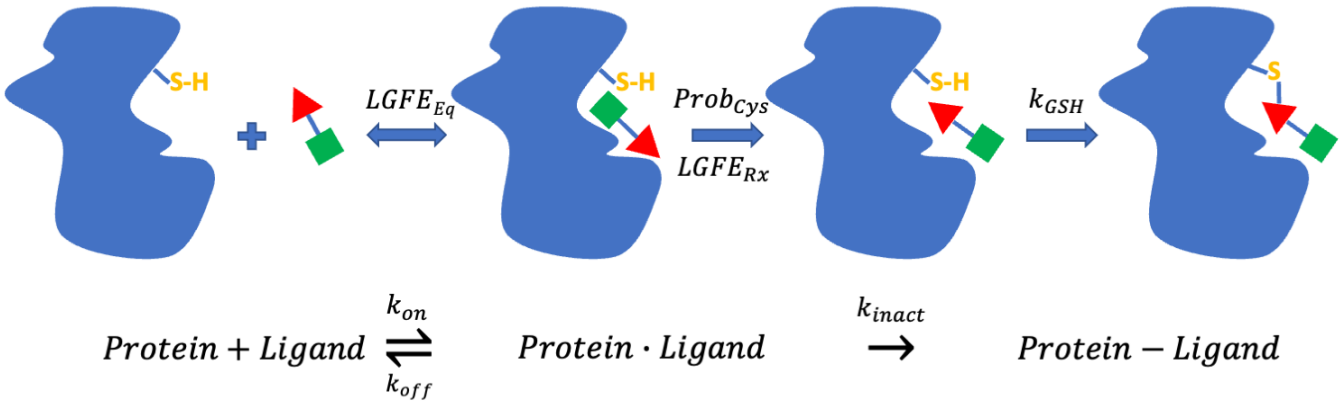

Figure 3. Illustration of the binding process for covalent ligands and relationship with SILCS-MC derived metrics. Target protein is shown in blue with Cys thiol group shown explicitly. Warhead in the covalent ligand is shown as a red triangle and noncovalent scaffold is shown as a green square. Covalent bonds between groups are shown as blue sticks. A covalent ligand will first bind with the target protein noncovalently which can be defined by k_{on} and k_{off} from which a reversible equilibrium constant and free energy of binding may be obtained. The $LGFE_{Eq}$ from SILCS-MC is relevant to this binding event. The second step towards covalent bond formation may be defined by the overall rate constant, k_{inact}. For reaction to occur, the ligand needs to adjust its orientation to position the warhead near the nucleophilic residue which can be described by the $Prob_{Cys}$ and $LGFE_{Rx}$ metrics from SILCS-MC. $LGFE_{Rx}$ indicates the binding free energy of the ligand in the potentially reactive orientation, which may be compared to $LGFE_{Eq}$ in the initial binding event and $Prob_{Cys}$ is related to the overlap of the warhead with the

nucleophilic Cys sulfur as required for covalent bond formation, which is represented by an experimental intrinsic reactivity metric k_{GSH}.

SILCS-MC was used to sample ligand binding poses around nucleophilic residues in a target protein yielding the equilibrium (reversible) and reaction-competent associated ligand binding affinities, $LGFE_{Eq}$ and $LGFE_{Rx}$, respectively. Variation of these and of the $Prob_{Cys}$ metric extracted from SILCS were tested for their relevance in guiding the covalent drug design. Exhaustive tests as described in the Supporting Information yielded optimal descriptors for covalent drug design model development. The variations of tested descriptors include the most favorable and mean values of the three SILCS-MC metrics described in section 2.4. Mean values over all docking poses under certain criterion or over the top 5 and 10 docking poses ranked by LGFE or $Prob_{Cys}$ metrics were tested. In addition, we tested a consensuses metric defined as $LGFE_{Rx} \times Prob_{Cys}$ that combines the two reaction-competent metrics. The final model uses the most favorable value for $LGFE_{Eq}$ and $Prob_{Cys}$ over all docking poses. This observation is consistent with the nature of covalent binding as the best $LGFE_{Eq}$ models how well a ligand can reside in the binding pocket during the initial noncovalent binding step, while the best $Prob_{Cys}$ and associated $LGFE_{Rx}$ indicates how close the warhead group in the ligand can approach he nucleophilic residue in an energetically favorable way, such that the ligand can assume a reaction-competent orientation to facilitate the reaction in the final covalent reaction step.

3.1 SILCS-Covalent can identify reactive Cys residues

Using Petri's library, 24 warhead probe molecules were docked by SILCS-MC targeting all Cys residues in the 6 target kinase proteins. The averaged $Prob_{Cys}$ value ($\langle Prob_{Cys} \rangle$) over all tested warheads is used for identification of tractable Cys residues since a larger $\langle Prob_{Cys} \rangle$ value for a Cys residue indicates that the residue samples conformations accessible to a broad range of warhead types, thereby potentially serving as a reliable nucleophilic target residue. From SILCS-MC, the best $Prob_{Cys}$ per warhead probe was

selected and the average value over all warhead probes, <$Prob_{Cys}$>, was evaluated for each Cys residue. First two rows in figure 4 show the calculated <$Prob_{Cys}$> for all Cys residues in the 6 tested kinases.

For BTK, only C481 has a substantial <$Prob_{Cys}$> value, consistent with this Cys being the front pocket N-terminal cap (FP Ncap) reactive Cys that FDA approved BTK drugs are targeting ***(9)***. The best two Cys residues with decent <$Prob_{Cys}$> for ERK2 are C161 and C166. Both residues have been previously confirmed to serve as covalent binder targets ***(59-62)***. Dalby et. al. previously found that a JNK inhibitor BI-78D3 binds to the D-recruitment site (DRS) of ERK and forms a covalent adduct with a conserved Cys residue, C159 in Rattus norvegicus ERK2, which correspond to C161 in Homo sapien ERK2 ***(59, 60)***. In another study, Gray et. al. used the multi-targeting ligand SM1-71 to scan the proteome for covalently modified kinases and found SM1-71 can covalently modify the DFG-1 C166 in ERK2 ***(61)***. In a crystallization effort, Reményi et. al. designed a ERK2 mutant, ERK2_AAG, to prevent crystal packing. During the study, they observed that β-mercaptoethanol, which was added to avoid oxidation during macroseeding, can form a (2-hydroxyethyl)thiocysteine adduct with C161 ***(62)***. Both these findings suggest that C161 and C166 in ERK2 can be covalently targeted.

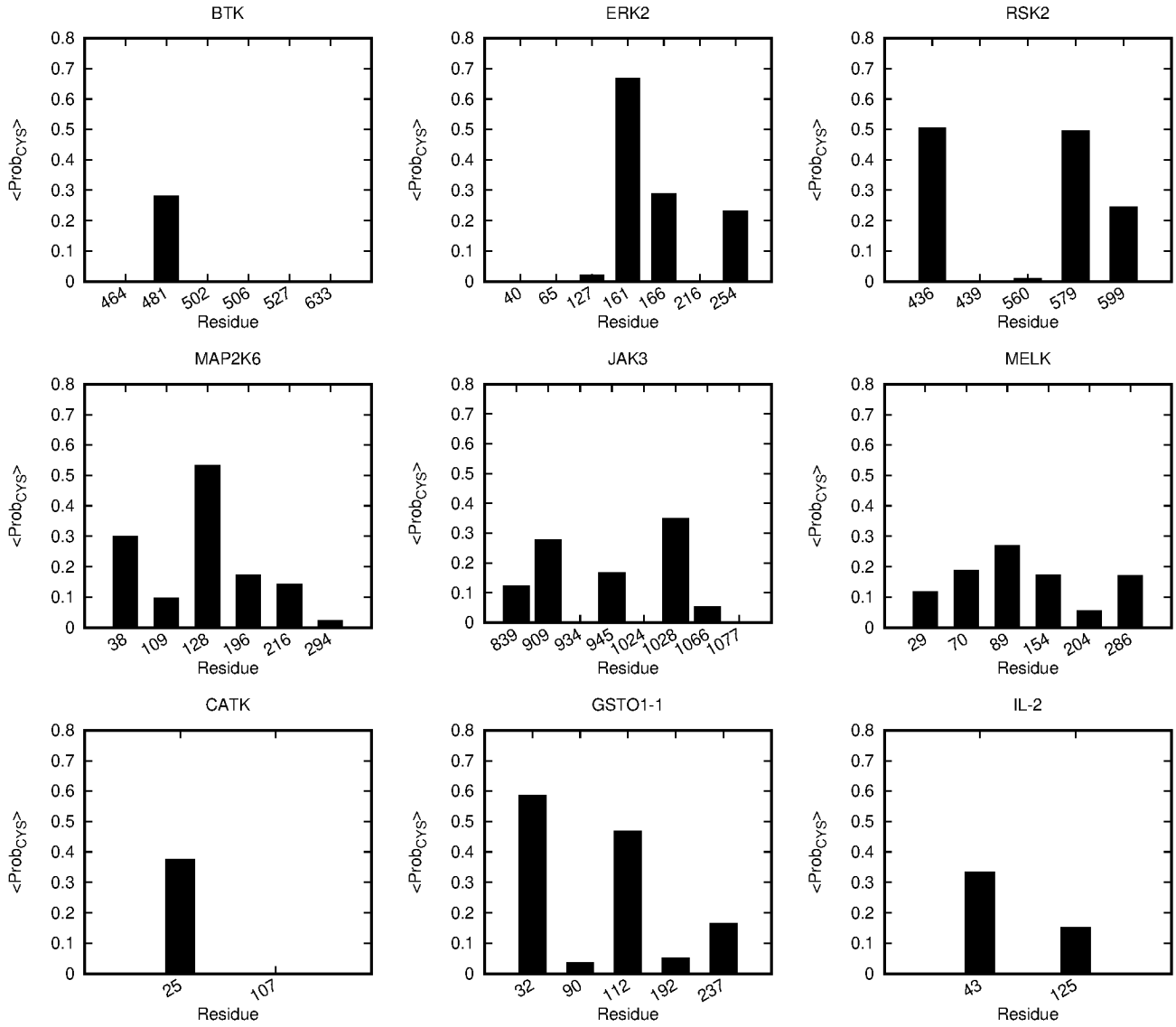

Figure 4. Average largest contact probabilities (<$Prob_{Cys}$>) over all 24 probes in the warhead library for each Cys residue in the tested protein systems. Cys residues that form disulfide bridge are excluded from all analysis.

Cys residue C436 on the regulatory C-terminal kinase domain (CTD) of RSK2 is a known targetable Cys with covalent inhibitors such as 1-[4-Amino-7-(3-hydroxypropyl)-5-(4-methylphenyl)-7H-pyrrolo[2,3-d]pyrimidin-6-yl]-2-fluoroethanone (FMK) bonding to this residue ***(63, 64)***. This is consistent with our prediction that C436 has the largest <$Prob_{Cys}$> value. In addition to C436, finite <$Prob_{Cys}$> values are also seen for C579, C599 and C560. A recent study used dimethyl fumarate (DMF) to target RSK2 CTD and revealed that all four Cys residue in RSK2 had DMF modifications using a peptide mapping

technique with C436 and C599 being totally modified ***(65)***. Crystallography confirmed covalent binding of DMF to C436 and C599. These experimental observations verify our predictions using the <$Prob_{Cys}$> metric for RSK2.

For MAP2K6, <$Prob_{Cys}$> predicts C128 followed by C38 and C196 to be accessible to warheads. To our knowledge, there are only few studies about covalent binding to this target and no crystal structure is available for a MAP2K6-covalent binder complex. However, previous cheminformatic and proteome-wide works implicated the Gatekeeper region C128 or DFG motif adjacent C196 as more accessible or reactive towards small-molecule electrophiles ***(66)***. And Chan et. al. discovered that ethacrynic acid (EA) inhibits MAP2K6 with nonconserved C38 being the site of covalent modification as confirmed by tandem mass spectrometry ***(66)***. These observations confirm our predictions about MAP2K6.

In JAK3, the FP Ncap C909 is known to be reactive to covalent molecules ***(67, 68)***. Recently, Liu et. al. studied several human kinases using the constant pH MD simulations (CpHMD) and discovered that C1028 on the αG helix of JAK3 is also reactive along with C909 ***(22)***. Consistent with these studies, C1028 and C909 are ranked as the top two Cys residues using the <$Prob_{Cys}$> metric. A limited number of studies on covalent binding to MELK are available to our knowledge. A previous study suggested that C70 and C89 may be exploited to develop selective and potent irreversible MELK inhibitors ***(69)***. These two Cys residues are ranked among the top two based on their <$Prob_{Cys}$> values which is consistent with the previous discussion ***(69)***.

We next extended our analysis beyond human kinases to cover other protein classes, including hydrolase, signaling and transferases proteins. The results are shown on the bottom row of Figure 4. For CATK, <$Prob_{Cys}$> successfully identified the catalytic C25 that has been targeted by covalent inhibitors ***(57, 70)***. Large <$Prob_{Cys}$> was obtained for C32 in GSTO1-1 consistent with its nucleophilic role as confirmed by many covalent inhibitor studies ***(71-73)***. For the IL-2 K43C mutant, the mutated residue C43 was calculated to have larger <$Prob_{Cys}$> value than the native C125 which is consistent with the previous study where K43 residue was engineered to Cys residue to enable covalent binding of guanidine

fragments ***(74)***. Thus, <$Prob_{Cys}$> was able to identify reactive Cys residues in the three tested protein systems beyond kinases.

3.2 ML model to determine optimal warhead

After a nucleophilic Cys residue is predicted, warhead groups with the highest potential to react with the residue need to be identified. To facilitate this step, metrics from SILCS-MC as well as k_{GSH}, the GSH reactivity rate constant, were used as inputs to train ML models based on the activity data from Petri's work for the 24 warhead fragments targeting the six kinase proteins ***(34)***. As discussed above, classification ML models were pursued due to the ambiguity in the inhibition percentage data.

For each kinase, the Cys residue which was either experimentally confirmed by crystallography or mass spectrometry to form a covalent bond with inhibitors or the most commonly targeted one when multiple Cys residues were confirmed, is assumed to bind the current set of warhead fragments. The selected Cys residues for model development are summarized in Table S2 in the supporting information. For MAP2K6, even though both C38, C128 and C196 were previously indicated to be reactive Cys residues, only C38 was experimentally confirmed by mass spectrometry to form covalent bond with small molecules ***(66)***. Both C436 and C599 in RSK2 were experimentally confirmed by crystallography to form covalent bond with molecules ***(63-65)***; however, C436 is the most targeted Cys residue ***(63, 64)***, so it was assumed to be the reactive residue. For MELK, there is no crystal structure confirmed Cys residue, although according to the docking and NMR analysis in Petri's work ***(34)***, C70 was suggested as the site of labeling.

SILCS-MC docking was undertaken targeting the residues listed in Table S2. From the docking, $LGFE_{Eq}$, $Prob_{Cys}$ and $LGFE_{Rx}$ for each warhead fragment targeting each kinase were calculated. Six typical classification ML algorithms, including LR, SVM, KNN, DT, RF and NB, were tried and their performances from 5-fold cross validation are shown in Table 2. Performances evaluated using the holdout method as shown by confusion matrixes can be found in Figure S1 in the supporting information.

As can be seen in Table 2, most classification ML algorithms performed satisfactorily on differentiating warhead fragments with high activities from low activity compounds. The metrics to define the

performance are described in the SI. Among the tested ML methods, the RF model has the overall best performance. Further analysis on the RF model reveals the input feature importance as shown in Figure 5. All the three SILCS-MC metrics as well as the k_{GSH} data have similar feature importance indicating that all properties associated with those metrics are important for warhead reactivity. This observation is expected since $LGFE_{Eq}$ captures the binding strength of warhead fragments in the first noncovalent binding step, $Prob_{Cys}$ represents the potential for close contact between the warhead and nucleophilic residue in the second binding step, k_{GSH} describes the warhead intrinsic reactivity with thiol group at ideal condition while $LGFE_{Rx}$ complements k_{GSH} with information about the ability of the ligand to interact with the local environment around the nucleophile. The final RF ML model was trained using all 144 data points for use as an ML tool to facilitate warhead selection, as applied below.

To test the utility of the final ML model, it was used to predict optimal warheads for glyceraldehyde-3- phosphate dehydrogenase (GAPDH) ***(75)***. For most protein systems, established covalent ligands were usually developed focusing on one or a few warhead groups that were pioneered by a single research group with further efforts focusing on optimization of the noncovalent scaffold instead of profiling different warheads. Accordingly, such systems are insufficient to validate our ML model. However, GAPDH has been widely studied leading to identification of multiple endogenous metabolites and xenobiotics as well as exogenous covalent inhibitors that contain multiple warhead types ***(75)*** making it a good testing ground for the ML model. Using the SILCS-MC metrics on the 24 warhead fragments for GAPDH as input features for the ML model, activity classification for each warhead was predicted. and the results can be found in Table S3 in the supporting information. Out of this ML classification effort, warheads 1 (acrylamide), 2 (acrylate), 5 (maleimide), 6 (maleimide), 11 (isothiocyanate), 12 (isothiocyanate), 20 (haloacetophenone), 21 (epoxide) and 22 (fluoride) out of the 24 warhead fragments were labeled to the high activity class by the ML model. Identification of warheads 1, 5, 6, 20 and 21 as high activity is consistent with the known GAPDH covalent warheads acrylamide (ACR), N-ethyl maleimide (NEM), α-halomethylcarbonyl and epoxide ***(75)***. This result is quite promising since the RF

ML model developed from kinase data was successful against a dehydrogenase protein implying its potential generic utility though additional tests on a wider range of systems are required.

Table 1. Performances of the six tested classification ML models from 5-fold cross validations on the 144 data points from the reference ***(34)***.

Model	Accuracy[a]	Precision[a]	Recall[a]	AUC[a]	F1[a]
LR	0.70 (0.09)	0.69 (0.08)	0.66 (0.07)	0.74 (0.04)	0.77 (0.08)
SVM	0.69 (0.07)	0.70 (0.06)	0.64 (0.06)	0.72 (0.08)	0.78 (0.06)
KNN	0.69 (0.05)	0.67 (0.04)	0.66 (0.03)	0.72 (0.04)	0.75 (0.06)
DT	0.62 (0.09)	0.60 (0.09)	0.61 (0.09)	0.61 (0.09)	0.68 (0.08)
RF	0.73 (0.05)	0.72 (0.05)	0.70 (0.04)	0.75 (0.05)	0.78 (0.06)
NB	0.69 (0.09)	0.69 (0.09)	0.66 (0.07)	0.74 (0.06)	0.76 (0.09)

[a] Average value across the 5 runs from the cross validation with standard deviation in parenthesis.

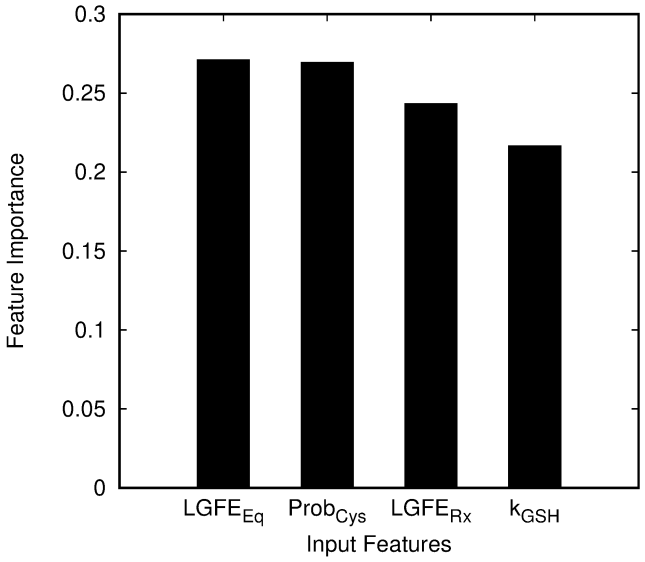

Figure 5. Input feature importance analysis for the RF ML model.

3.3 SILCS-Covalent can help with optimization of noncovalent scaffold

Once an optimal warhead targeting a reactive Cys residue is identified, development of the noncovalent scaffold that contributes to the compound specificity is required. This can be done in an iterative fashion through several rounds of design and evaluation combining computational and experimental methods. The noncovalent scaffold needs to have favorable, specific interactions with the binding pocket for the first binding step. For the second step, the ligands need to favorably interact with the protein in the vicinity of the target Cys to sample reaction-competent conformations. In the ideal scenario, the equilibrium binding site is the same as the reaction-competent site although in most cases the sites are not identical with the equilibrium site contributing to or adjacent to the reaction-competent site. In any of these situations, it is necessary to quantify the energetics of these ligand-protein interactions. This can be performed using the LGFE scores from SILCS-MC, which have been shown to quantify binding strengths of a wide selection of inhibitors targeting various protein systems for non-covalent inhibitors ***(32, 33)***. Accordingly, we explore the potential for use of LGFE scores to optimize the noncovalent scaffold in covalent inhibitors in the context of the same or similar warheads. This allows for the LGFE scores for the full ligand to be used as both equilibrium and reaction-competent binding involves the full ligand prior to covalent bond formation. Three covalent ligand data sets with three different warhead types targeting three proteins, respectively, were selected for analysis. These include 17 ligands with the acrylamide warheads targeting BTK, 9 ligands with a cyanoacrylamide warhead targeting RSK2 and 17 ligands with cyanamide warheads targeting CATK. Table S1 and Figure S2 summarize the data and structures of all the compounds.

$LGFE_{Eq}$ and $LGFE_{Rx}$ were compared with experimental binding free energies based on the experimental IC_{50} values for the three systems. We note that since covalent binding is a multistep process including the formation of an irreversible bond, full interpretation of the physical meaning of the experimental IC_{50} value is not possible. However, when the same or similar warhead is on all the ligands, the reaction chemistry may be assumed to be similar such that the overall binding differences are dominated by the interactions between the noncovalent scaffold and binding site residues at either the equilibrium or reaction-competent sites. Initial analysis involved comparing the average experimental binding free

energies along with the $LGFE_{Eq}$ and $LGFE_{Rx}$ averages and the difference between those values to understand overall trends in these terms. Initial analysis results are shown in Table 2 for the three systems. For all systems, the average ΔG_{exp} values are more favorable than the LGFE averages. While LGFE does not represent a formal free energy of binding, the values are representative of the binding affinities as shown in previous studies ***(32, 33)***. The ΔG_{exp} values being more favorable than the LGFE scores is expected as the IC_{50} values from which they are obtained include contributions from covalent bond formation. Comparing $LGFE_{Eq}$ and $LGFE_{Rx}$ shows the former to be systematically more favorable than the latter, consistent with full relaxation of the of the ligands yielding the $LGFE_{Eq}$ while with $LGFE_{Rx}$ the binding orientation is restricted to being in close contact with the target Cys sulfur atom. However, the difference between the two LGFE values differ for the three systems. This is indicative of the relationship of the equilibrium versus reaction-competent orientations, where smaller differences would indicate less reorganization of the ligand-protein interactions required to assume the reaction-competent conformation as required for covalent bond formation.

Table 2. Average experimental binding free energies, $LGFE_{Eq}$ and $LGFE_{Rx}$ scores and their differences (in kcal/mol).

System	$\langle\Delta G_{exp}\rangle$	$\langle LGFE_{Eq}\rangle$	$\langle LGFE_{Rx}\rangle$	$\langle LGFE_{Rx}\rangle$-$\langle LGFE_{Eq}\rangle$
BTK	-11.76	-9.80	-5.89	3.91
RSK2	-7.32	-6.32	-4.46	1.86
CATK	-9.05	-7.38	-4.12	3.26

Subsequent analysis involved the correlation between the experimental and LGFE free energies (Table 3 and Figure S3). As Table 3 shows, moderate correlations are observed between experimental binding data and both energy scores from SILCS-MC. Such correlations are promising since the data sets span large binding affinity ranges, for example, 5 nM to 100 μM for the CATK compounds. In addition, the compounds are from non-congeneric series for CATK and RSK2 with varied noncovalent scaffolds (Figure S2). For example, molecular weight varies from 96 to 291 g/mol for the CATK compounds.

Moreover, we reiterate that the three test sets have different warheads thus such moderate correlations for all three test cases implies the potential of using SILCS-MC energy scores to guide the non-covalent scaffold design.

Table 3. Pearson correlation coefficients (PR) for correlations between $LGFE_{Eq}$ or $LGFE_{Rx}$ with experimental binding free energies before and after BML reweighting for the three test systems.

	Original		BML	
System	$LGFE_{Eq}$	$LGFE_{Rx}$	$LGFE_{Eq}$	$LGFE_{Rx}$
BTK	0.57	0.40	0.75	0.44
RSK2	0.46	0.69	0.76	0.71
CATK	0.38	0.69	0.44	0.69

One notable observation is that the PR values for $LGFE_{Eq}$ and $LGFE_{Rx}$ for the individual targets are different. Table S10 lists the correlations between the two SILCS-MC binding metrics showing the two to not be fully correlated which is consistent with the results in Table 3 and Figure S3. This is reasonable since they describe the two different binding events. While the datasets are small, the differences may offer insights into the importance of two binding events with respect to ligand design. With BTK, $LGFE_{Eq}$ has the better correlation with experimental binding data and the difference between the average $LGFE_{Eq}$ and $LGFE_{Rx}$ is the largest of the three systems. This indicates that the equilibrium binding step dominates overall inhibition. For RSK2 and CATK, $LGFE_{Rx}$ has the better correlation with experimental data and the differences between the average $LGFE_{Eq}$ and $LGFE_{Rx}$ values are smaller, which may indicate that reaction-competent binding dominates overall binding. While the applied data sets are limited, such observations might be helpful to facilitate scaffold design offering guidance on which step to consider during ligand design.

While the correlations between the SILCS-MC metrics and experimental binding data suggest the potential quantitative use of the two metrics on noncovalent scaffold optimization, the level of correlation is moderate. To improve the predictability of the LGFE metrics, we applied the Bayesian machine learning (BML) technique previously developed for SILCS ligand binding predictions ***(32, 33)***. BML optimization

alters the contributions of the different types of SILCS GFE FragMaps to the calculated LGFE values targeting known experimental data. Given that the SILCS FragMaps represent high-quality priors, the optimization may be performed with a limited amount of experimental data. BML training was performed on the three systems for both $LGFE_{Eq}$ and $LGFE_{Rx}$ targeting the Pearson correlation with the experimental data. A force constant of 500 kcal/mol was used with a flat bottom potential to prevent overfitting of the weighting of the FragMaps with lower and upper limit of weights set to be 0.1 and 2.0, respectively, as previously described ***(32, 33)***. After training, SILCS-MC docking was rerun for all the ligands using the reweighted FragMaps from which new LGFE values were obtained. Figure S4 shows the correlation plots and Table 3 includes the PR for correlations between the BML-model $LGFE_{Eq}$ or $LGFE_{Rx}$ binding scores with experimental binding free energies. For all cases, the BML trained $LGFE_{Eq}$ or $LGFE_{Rx}$ values have better or equal level of correlations with experimental data with substantial improvements occurring for BTK and RSK2. For BTK, higher correlation still occurs with $LGFE_{Eq}$, while the correlation is now similar for the two LGFE metrics with RSK2 and $LGFE_{Rx}$ is still the most predictive with CATK. These observations indicated that the BML optimization, which was previously used for reversible drug design, can also be used for covalent ligand optimization purpose.

Further analysis involved the docking poses from SILCS-MC with an example from RSK2 shown in Figure 6. Figure 6A shows the reaction-competent pose of the cyanoacrylamide warhead fragment while the equilibrium binding pose associated with $LGFE_{Eq}$ for RSK2 ligand compound 24 ***(14)*** is shown in Figure 6B. The reaction-competent pose of compound 24 is shown in Figure 6C, along with the experimental crystal pose of that compound ***(14)***. Comparison of all three panels show the compounds to bind in the same region despite A and C being reaction-competent poses and B being an equilibrium pose. This is consistent with the similar correlations of the BML optimized $LGFE_{Eq}$ and $LGFE_{Rx}$ with experimental data. Going from the equilibrium to reaction-competent orientation for compound 24 primarily involves a rotation of the compound in the same binding region. The reaction-competent binding pose has the reactive atom in proximity to C436 and in a very similar location as the crystal binding orientation after covalent bond formation. Figure 6B includes the FragMaps used for the docking. The

overlap of the FragMaps with the compound indicates those moieties that are contributing to binding and the FragMaps in the vicinity of compound suggest new types of functionalities that may improve the binding affinity. This may also be performed for the ligand in the reaction-competent pose, which may be a preferable option with RSK2 given the similar correlations (Table 3). For example, the apolar FragMaps around the binding orientation of compound 24 suggest additional apolar groups could increase binding, consistent with the higher experimental binding affinity observed for compound 27 which has a larger pyrrolopyrimidine group at the terminus when compared with the smaller pyrazole group of 24 (Figure S2). Beyond the visual information in the FragMaps, the GFE contributions of the atoms and functional groups to the LGFE score, which were previously studied for reversible drug design ***(32, 33)***, can be quantified, as shown for the reaction-competent pose of compound 24 in Figure 6D. The alkene carbon in compound 24 which serves as the reactive atom in covalent bond formation has a small unfavorable atom GFE contribution (red label) as expected since the reaction-competent orientation has this atom in close contact with target Cys, while other atoms all have favorable GFE contributions to the total $LGFE_{Rx}$. This information indicates the role of different regions of the molecule that drive binding. Based on that information and the FragMaps around the ligand, synthetically accessible modifications of the compound may be designed, which may then be evaluated through SILCS-MC of the modified ligands.

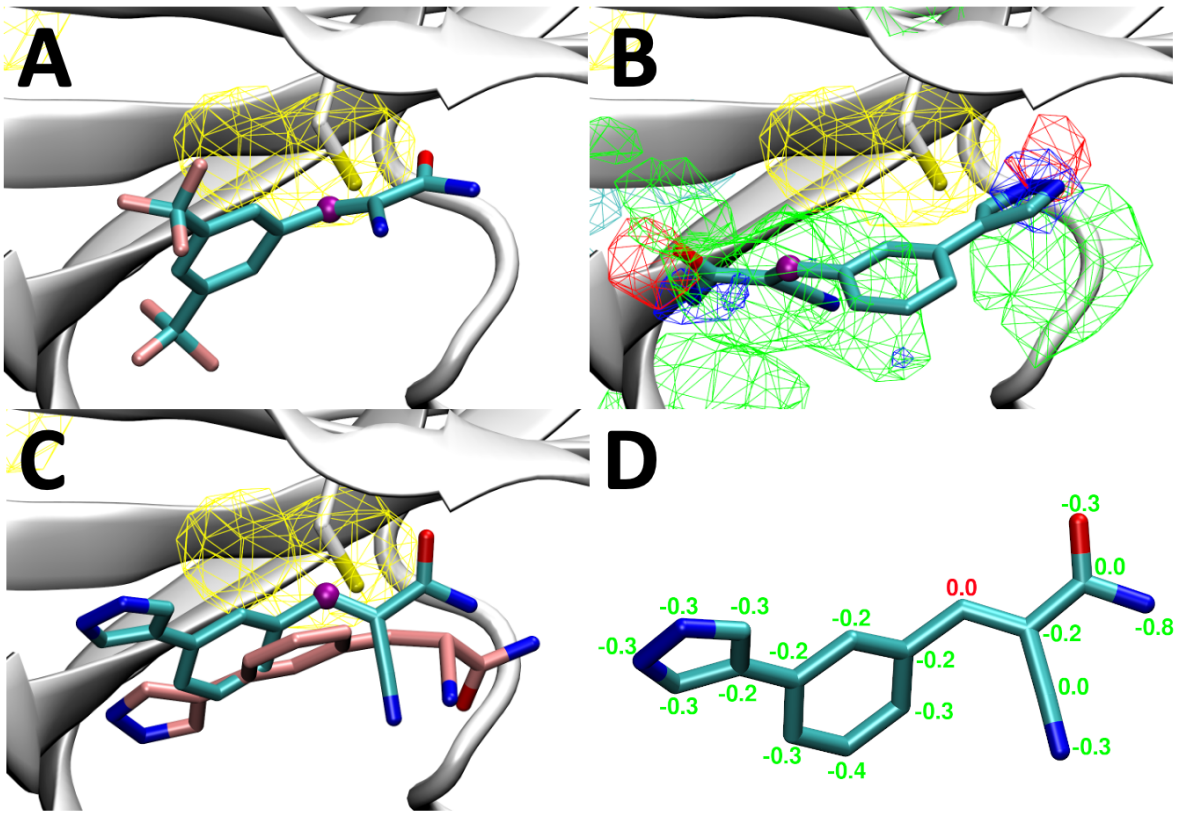

Figure 6. A) Docking pose of the cyanoacrylamide warhead compound from the Petri's library with the best $Prob_{Cys}$ to RSK2, B) docking pose of RSK2 inhibitor 24 with the best $LGFE_{Eq}$, C) docking pose of RSK2 inhibitor 24 with the best $Prob_{Cys}$ and D) atom GFE contributions (favorable energy in green and unfavorable energy in red) to $LGFE_{Rx}$ for docking pose in panel C. Protein is shown in cartoon representation with residue C436 in licorice representation with carbon in white color and the compounds are shown in licorice representation with carbon in cyan color and the reactive atom in the warhead group is marked by a magenta sphere. Crystal binding orientation of compound 24 from PDB entry 4m8t is shown with carbon in pink color in panel C. Apolar, hydrogen bond donor, acceptor FragMaps as well as thiol sulfur probability map are shown in green, blue, red and yellow color, respectively. FragMaps are at an isocontour level of -1.2 kcal/mol.

3.4 SILCS-Covalent workflow for covalent drug discovery

Based on the above analyses and observations, a full SILCS-Covalent workflow was implemented and is illustrated in Figure 7. The workflow encompasses the three modelled stages of covalent drug design from identification of potentially reactive Cys residues, selection of the optimal warhead for a specific Cys residue, and optimization of the noncovalent scaffold to improve binding and specificity. The workflow is initiated by performing SILCS simulations on a known protein structure followed by

FragMap and Cys thiol sulfur probability map generation (yellow meshes) as well as generation of the SILCS exclusion map. SILCS-MC docking is then conducted using the warhead library for all Cys residues and the best $Prob_{Cys}$ per warhead probe is recorded and the average value over all warhead probes, $<Prob_{Cys}>$, is used to identify reactive Cys residues. Larger $<Prob_{Cys}>$ values indicate Cys residues that may react with a broader range of warheads. Once a tractable Cys residue is selected, the $LGFE_{Eq}$, $LGFE_{Rx}$ and $Prob_{Cys}$ values, calculated for the previous step, targeting the selected Cys residue are used as input features along with k_{GSH} for the developed ML model. This outputs the classification metric of the potential for the warhead to target the Cys residue. Once the warhead group is selected, noncovalent scaffold design can be initiated. This could involve various lead compound identification approaches including virtual screening ***(1, 2)***. SILCS-Pharm ***(27, 28)*** and SILCS-MC docking near the chosen Cys residue may be used to evaluate the druggability of the region employing the previously calculated SILCS FragMaps and exclusion map. $LGFE_{Eq}$ and $LGFE_{Rx}$ values as well as predicted binding poses may then subsequently be used to guide further optimization of the noncovalent scaffold.

When applying the workflow, the most time-consuming part is running the SILCS simulations. This typically requires 1-3 days for most proteins using 10 GPUs ***(76)***. Once the FragMaps, Cys probability map and exclusion map are available, they can be used repeatedly, allowing for rapid calculation of all the SILCS metrics described above for the warhead fragments as well as putative covalent ligands. When lead compounds are available, optimization of noncovalent scaffold maybe undertaken using the same FragMaps. New ideas can be quickly tested using SILCS-MC on thousands of designs through the rapid calculation of $LGFE_{Eq}$, $LGFE_{Rx}$ and $Prob_{Cys}$ in minutes. Thus, the full SILCS-Covalent workflow is very efficient while providing reliable predictions as discussed in the previous sections.

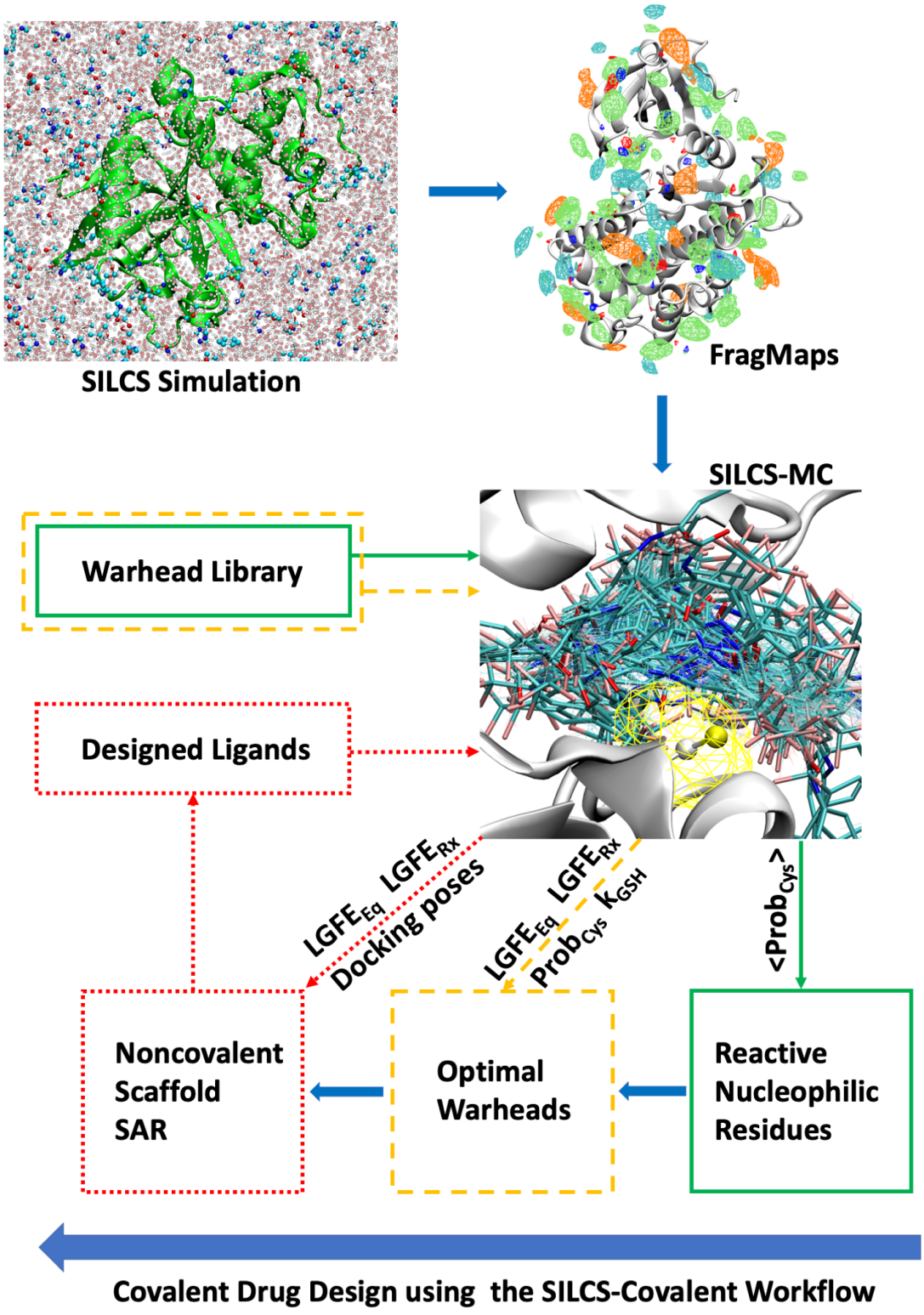

Figure 7. Workflow of the SILCS-Covalent drug design process. The workflow starts by running SILCS simulations from which the FragMaps as well as Cys sulfur probability distributions are generated (Apolar FragMaps in green, hydrogen bond donor and acceptor FragMap in blue and red, respectively, and positively charged and negatively charged FragMaps in cyan and orange, respectively). SILCS-MC is

then conducted for a warhead library or designed ligands around the specific Cys residue (Cys sulfur probability map used to calculate the $Prob_{Cys}$ metric in yellow). Potential reactive Cys residues are first identified as indicated by the green solid line marked route. Next, optimal warhead groups are identified for the reactive Cys, as indicated by the orange dashed lines. The last step is to design noncovalent scaffolds and establish a structure-activity relationship (SAR) as indicated by red dotted lines. Different metrics are employed at different steps as shown for each route.

4. Summary

In this study, we explored the use of SILCS in covalent drug discovery, motivated by the broad utility and computational efficiency of the technology. Our protocol uses unbiased and reaction-competent docking, offering the computational simplicity of a method that may be used for both covalent and noncovalent drug design. Through SILCS-MC docking of a previously curated warhead library, reaction-competent binding poses were collected, and the warhead contact probabilities with nucleophilic Cys residue sulfurs were evaluated. The average contact probability over all warhead fragments targeting each Cys residue in the tested proteins was verified to be able to identify tractable Cys residues for covalent binding. Using three metrics from SILCS-MC as well as an experimental intrinsic reactivity descriptor, six ML classification models were trained to predict warhead fragment reactivity targeting the six kinases, from which a RF based ML model was identified to perform the best and is suggested for practical use. Application of the ML model on a non-kinase protein GAPDH verified that the ML model was able to identify effective warheads from the warhead library. Further study using three proteins shows both noncovalent, equilibrium binding affinity, $LGFE_{Eq}$, and reaction-competent binding affinity, $LGFE_{Rx}$, are in moderate correlations with experimental binding free energies. This suggests the utility of these affinity metrics for noncovalent scaffold design. The combination of the ability to identify potentially reactive Cys residues, warheads with the potential to target those residues and calculated binding affinities for ligand optimization, represents an integrated workflow SILCS-Covalent for covalent drug discovery.

The ability of the current protocol for covalent drug design and optimization of the noncovalent scaffold of the ligands needs to be fully explored using larger data sets. However, the current results are

promising as the three analyzed sets of compounds cover a wider range of activity with noncovalent scaffold regions of the molecules encompass a range of chemical structures. In addition, the SILCS BML optimization approach was shown to increase the predictability of the SILCS metrics, which may be performed once experimental data on a system is generated. The moderate correlations between SILCS-MC metrics with experimental data together with the ability of BML to further improve predictions, suggests the potential utility of SILCS for covalent drug design.

Supporting Information. Confusion matrixes for the holdout tests of the six tested ML methods, predicted energy metrics and experimental binding data for the three test sets, selected Cys residues for ML model development, RF ML model predictions for GAPDH, details about the selection of SILCS metrics, ML model training hyperparameters and description of performance metrics, 2D structures of all ligands from the three test sets, correlation plots between SILCS-MC energy metrics with experimental data before and after BML reweighting, correlation analysis for the two LGFE metrics. This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements. This work was supported by NIH grant GM131710, University of Maryland Center for Biomolecular Therapeutics (CBT), Samuel Waxman Cancer Research Foundation, and the Computer-Aided Drug Design (CADD) Center at the University of Maryland, Baltimore.

Conflict of interest: A.D.M. is Co-founder and CSO of SilcsBio LLC.

References:

1. Yu, W.; MacKerell, A. D. Computer-aided drug design method. In: Sass, P. (eds) Antibiotics. Methods in Molecular Biology, vol 1520, **2017**, Humana Press, New York, NY, USA, pp 85-106.
2. Yu, W.; Weber, D. J.; MacKerell, A. D. Computer-aided drug design: an update. In: Sass, P. (eds) Antibiotics. Methods in Molecular Biology, vol 2601. **2023**, Humana Press, New York, NY, USA, pp 123-152.
3. Sutanto, F.; Konstantinidou, M.; Dömling, A. Covalent inhibitors: a rational approach to drug discovery. *RSC Med. Chem.* **2020**, *11*, 876-884.

4. Boike, L.; Henning, N. J.; Nomura, D. K. Advances in covalent drug discovery. *Nat. Rev. Drug Discov.* **2022**, *21*, 881-898.
5. Flower, R. The development of COX2 inhibitors. *Nat. Rev. Drug Discov.* **2003**, *2*, 179–191.
6. Singh, J.; Petter, R.; Baillie, T.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov.* **2011**, *10*, 307-317.
7. De Vita, E. 10 years into the resurgence of covalent drugs. *Future Med. Chem.* **2020**, *13*, 193-210.
8. Baillie ,T. A. Targeted covalent inhibitors for drug design. *Angew. Chem. Int. Ed.* **2016**, *55*, 13408-13421.
9. Burger, J. A.; Tedeschi, A.; Barr, P. M.; Robak, T.; Owen, C.; Ghia, P.; Bairey, O.; Hillmen, P.; Bartlett, N. L.; Li, J.; Simpson, D.; Grosicki, S.; Devereux, S.; McCarthy, H.; Coutre, S.; Quach, H.; Gaidano, G.; Maslyak, Z.; Stevens, D. A.; Janssens, A.; Offner, F.; Mayer, J.; O'Dwyer, M.; Hellmann, A.; Schuh, A.; Siddiqi, T.; Polliack, A.; Tam, C. S.; Suri, D.; Cheng, M.; Clow, F.; Styles, L.; James, D. F.; Kipps, T. J.; RESONATE-2 Investigators. Ibrutinib as initial therapy for patients with chronic lymphocytic leukemia. *New Engl. J. Med.* **2015**, *373*, 2425-2437.
10. Hammond, J.; Leister-Tebbe, H.; Gardner, A.; Abreu, P.; Bao, W.; Wisemandle, W.; Baniecki, M.; Hendrick, V. M.; Damle, B.; Simón-Campos, A.; Pypstra, R. Rusnak, J. M.; EPIC-HR Investigators. Oral nirmatrelvir for high-risk, nonhospitalized adults with COVID-19. *New Engl. J. Med.* **2022**, *386*, 1397-1408.
11. Kumalo, H. M.; Bhakat, S.; Soliman, M. E. S. Theory and applications of covalent docking in drug discovery: merits and pitfalls. *Molecules* **2015**, *20*, 1984-2000.
12. Oyedele, A. Q. K.; Ogunlana, A. T.; Boyenle, I. D.; Adeyemi, A. O.; Rita, T. O.; Adelusi, T. I.; Abdul-Hammed, M.; Elegbeleye, O. E.; Odunitan, T. T. Docking covalent targets for drug discovery: stimulating the computer-aided drug design community of possible pitfalls and erroneous practices. *Mol. Divers.* **2022**, DOI: 10.1007/s11030-022-10523-4.
13. Zhu, K.; Borrelli, K. W.; Greenwood, J. R.; Day, T.; Abel, R.; Farid, R. S.; Harder, E. Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. *J. Chem. Inf. Model.* **2014**, *54*, 1932-1940.
14. London, N.; Miller, R.; Krishnan, S.; Uchida, K.; Irwin, J. J.; Eidam, O.; Gibold, L.; Cimermančič, P.; Bonnet, R.; Schoichet, B. K.; Taunton, J. Covalent docking of large libraries for the discovery of chemical probes. *Nat. Chem. Biol.* **2014**, *10*, 1066-1072.
15. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
16. Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561-1573.
17. Wei, L.; Chen, Y.; Liu, J.; Rao, L.; Ren, Y.; Xu, X.; Wan, J. Cov_ DOX: a method for structure prediction of covalent protein- ligand bindings. *J. Med. Chem.* **2022**, *65*, 5528-5538.
18. Scarpino, A.; Ferenczy, G. G.; Keserű, G. M. Comparative evaluation of covalent docking tools. *J. Chem. Inf. Model.* **2018**, *58*, 1441-1458.
19. Marino, S. M.; Gladyshev, V. N. Analysis and functional prediction of reactive cysteine residues. *J. Biol. Chem.* **2012**, *287*, 4419-4425.
20. Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525-537.
21. Awoonor-Williams, E.; Rowley, C. N. Evaluation of methods for the calculation of the pKa of cysteine residues in proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662-4673.
22. Liu, R.; Verma, N.; Henderson, J. A.; Zhan, S.; Shen, J. Profiling MAP kinase cysteines for targeted covalent inhibitor design. *RSC Med. Chem.* **2022**, *13*, 54-63.
23. Guvench, O.; MacKerell, A. D., Jr. Computational Fragment-Based Binding Site Identification

by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435.
24. Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. Sampling of organic solutes in aqueous and heterogeneous environments using oscillating excess chemical potentials in grand canonical-like monte carlo-molecular dynamics simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2281-2290.
25. Raman, E. P.; Yu, W.; Guvench, O.; MacKerell, A. D. Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J. Chem. Inf. Model.* **2011**, *51*, 877-896.
26. Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell, A. D. Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. *J. Chem. Inf. Model.* **2013**, *53*, 3384-3398.
27. Yu, W.; Lakkaraju, S.; Raman, E. P.; MacKerell, A., Jr. Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling. *J. Comput. Aided Mol. Des.* **2014**, *28*, 491-507.
28. Yu, W.; Lakkaraju, S. K.; Raman, E. P.; Fang, L.; MacKerell, A. D. Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules. *J. Chem. Inf. Model.* **2015**, *55*, 407-420.
29. Yu, W.; Jo, S.; Lakkaraju, S. K.; Weber, D. J.; MacKerell, A. D. Exploring protein-protein interactions using the site-identification by ligand competitive saturation methodology. *Proteins: Struct. Funct. Bioinf.* **2019**, *87*, 289-301.
30. MacKerell, A. D.; Jo, S.; Lakkaraju, S. K.; Lind, C.; Yu, W. Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (SILCS-Hotspots). *Biochim. Biophys. Acta Gen. Subj.* **2020**, *1864*, 129519.
31. Goel, H.; Hazel, A.; Yu, W.; Jo, S.; MacKerell, A. D. Application of site-identification by ligand competitive saturation in computer-aided drug design. *New J. Chem.* **2022**, *46*, 919-932.
32. Ustach, V. D.; Lakkaraju, S. K.; Jo, S.; Yu, W.; Jiang, W.; MacKerell, A. D. Optimization and evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a tool for target-based ligand optimization. *J. Chem. Inf. Model.* **2019**, *59*, 3018-3035.
33. Goel, H.; Hazel, A.; Ustach, V. D.; Jo, S.; Yu, W.; MacKerell, A. D. Rapid and accurate estimation of protein-ligand relative binding affinities using site-identification by ligand competitive saturation. *Chem. Sci.* **2021**, *12*, 8844-8858.
34. Petri, L.; Egyed, A.; Bajusz, D.; Imre, T.; Hetényi, A.; Martinek, T.; Ábrányi-Balogh, P.; Keserű, G. M. An electrophilic warhead library for mapping the reactivity and accessibility of tractable cysteines in protein kinases. *Eur. J. Med. Chem.* **2020**, *207*, 112836.
35. 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.* **2022**, *50*, D445-D450.
36. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer Jr, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535-542.
37. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296– W303.
38. Jo, S.; Cheng, X.; Lee, J.; Kim, S.; Park, S. J.; Patel, D. S.; Beaven, A. H.; Lee, K. I.; Rui, H.; Park, S.; Lee, H. S.; Roux, B.; MacKerell, A. D.; Klauda, J. B.; Qi, Y.; Im, W. CHARMM-GUI 10 years for biomolecular modeling and simulation. *J. Comput. Chem.* **2017**, *38*, 1114-1124.
39. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Res.* **2022**, *50*, D439-D444.
40. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool,

K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589.
41. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.
42. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257-3273.
43. Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-Induced forces between two hydrophilic groups. *J. Phys. Chem.* **1994**, *98*, 2198–2202.
44. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701-1718.
45. Levitt, M.; Lifson, S. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **1969**, *46*, 269-279.
46. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690.
47. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511-519.
48. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A: At. Mol. Opt. Phys.* **1985**, *31*, 1695.
49. Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182-7190.
50. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671-690.
51. Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* **2012**, *33*, 2451-2468.
52. Broyden, C. G. The convergence of a class of double-rank minimization algorithms. *J. Inst. Math. Appl.* **1970**, *6*, 76-90.
53. Awoonor-Williams, E.; Isley, W. C.; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N. Quantum chemical methods for modeling covalent modification of biological thiols. *J. Comput. Chem.* **2019**, *41*, 427-438.
54. *Molecular Operating Environment (MOE)*, Chemical Computing Group Inc., https://www.chemcomp.com : Montreal, Canada.
55. Guo, X.; Zhang, Y. CovBinderInPDB: A structure-based covalent binder database. *J. Chem. Inf. Model.* **2022**, *62*, 6057-6068.
56. Caldwell, R.; Liu-Bujalski, L.; Qiu, H.; Mochalkin, I.; Jones, R.; Neagu, C.; Goutopoulos, A.; Grenningloh, R.; Johnson, T.; Sherer, B.; Gardberg, A.; Follis, A. V.; Morandi, F.; Head, J. Discovery of a novel series of pyridine and pyrimidine carboxamides as potent and selective covalent inhibitors of BTK. *Bioorg. Med. Chem. Lett.* **2018**, *28*, 3419-3424.
57. Falgueyret, J. P.; Oballa, R. M.; Okamoto, O.; Wesolowski, G.; Aubin, Y.; Rydzewski, R. M.; Prasit, P.; Riendeau, D.; Rodan, S. B.; Percival, M. D. Novel, nonpeptidic cyanamides as potent and

reversible inhibitors of human cathepsins K and L. *J. Med. Chem.* **2001**, *44*, 94-104.
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; V;erplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825-2830.
59. Kaoud, T. S.; Johnson, W. H.; Ebelt, N. D.; Piserchio, A.; Zamora-Olivares, D.; Van Ravenstein, S. X.; Pridgen, J. R.; Edupuganti, R.; Sammons, R.; Cano, M.; Warthaka, M.; Harger, M.; Tavares, C. D. J.; Park, J.; Radwan, M. F.; Ren, P.; Anslyn, E. V.; Tsai, K. Y.; Ghose, Ranajeet.; Dalby, K. Modulating multi-functional ERK complexes by covalent targeting of a recruitment site in vivo. *Nat. Commun.* **2019**, *10*, 5232.
60. Sammons, R. M.; Ghose, R.; Tsai, K. Y.; Dalby, K. N. Targeting ERK beyond the boundaries of the kinase active site in melanoma. *Mol. Carcinog.* **2019**, *58*, 1551-1570.
61. Rao, S.; Gurbani, D.; Du, G.; Everley, R. A.; Browne, C. M.; Chaikuad, A.; Tan, L.; Schröder, M.; Gondi, S.; Ficarro, S. B.; Sim, T.; Kim, N. D.; Berberich, M. J.; Knapp, S.; Marto, J. A.; Westover, K. D.; Sorger, P. K.; Gray, N. S. Leveraging compound promiscuity to identify targetable cysteines within the kinome. *Cell Chem. Biol.* **2019**, *26*, 818-829.E9.
62. Gógl, G.; Törő, I.; Reményi, A. Protein-peptide complex crystallization: a case study on the ERK2 mitogen-activated protein kinase. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *D69*, 486-489.
63. Cohen, M. S.; Zhang, C.; Shokat, K. M.; Taunton, J. Structural bioinformatics-based design of selective, irreversible kinase inhibitors. *Science* **2005**, *308*, 1318-1321.
64. Serafimova, I. M.; Pufall, M. A.; Krishnan, S.; Duda, K.; Cohen, M. S.; Maglathlin, R. L.; McFarland, J. M.; Miller, R. M.; Frödin, M.; Taunton, J. Reversible targeting of noncatalytic cysteines with chemically tuned electrophiles. *Nat. Chem. Biol.* **2012**, *8*, 471-476.
65. Andersen, J. L.; Gesser, B.; Funder, E. D.; Nielsen, C. J. F.; Gotfred-Rasmussen, H.; Rasmussen, M. K.; Toth, R.; Gothelf, K. V.; Arthur, J. S. C.; Iversen, L.; Nissen, P. Dimethyl fumarate is an allosteric covalent inhibitor of the p90 ribosomal S6 kinases. *Nat. Commun.* **2018**, *9*, 4344.
66. Chan, A. I.; McGregor, L. M.; Jain, T.; Liu, D. R. Discovery of a covalent kinase inhibitor from a DNA-encoded small-molecule library × protein library selection. *J. Am. Chem. Soc.* **2017**, *139*, 10192-10195.
67. Forster, M.; Chaikuad, A.; Bauer, S. M.; Holstein, J.; Robers, M. B.; Corona, C. R.; Gehringer, M.; Pfaffenrot, E.; Ghoreschi, K.; Knapp, S.; Laufer, S. A. Selective JAK3 inhibitors with a covalent reversible binding mode targeting a new induced fit binding pocket. *Cell Chem. Biol.* **2016**, *23*, 1335-1340.
68. Boggon, T. J.; Li, Y.; Manley, P. W.; Eck, M. J. Crystal structure of the Jak3 kinase domain in complex with a staurosporine analog. *Blood* **2005**, *106*, 996-1002.
69. Klaeger, S.; Heinzlmeir, S.; Wilhelm, M.; Polzer, H.; Vick, B.; Koenig, P. A.; Reinecke, M.; Ruprecht, B.; Petzoldt, S.; Meng, C.; Zecha, J.; Reiter, K.; Qiao, H.; Helm, D.; Koch, H.; Schoof, M.; Canevari, G.; Casale, E.; Depaolini, S. R.; Feuchtinger, A.; Wu, Z.; Schmidt, T.; Rueckert, L.; Becker, W.; Huenges, J.; Garz, A. K.; Gohlke, B. O.; Zolg, D. P.; Kayser, G.; Vooder, T.; Preissner, R.; Hahne, H.; Tõnisson, N.; Kramer, K.; Götze, K.; Bassermann, F.; Schlegl, J.; Ehrlich, H. C.; Aiche, S.; Walch, A.; Greif, P. A.; Schneider, S.; Felder, E. R.; Ruland, J.; Médard, G.; Jeremias, I.; Spiekermann, K.; Kuster, B. The target landscape of clinical kinase drugs. *Science* **2017**, *358*, eaan4368.
70. Mons, E.; Jansen, I. D. C.; Loboda, J.; van Doodewaerd, B. R.; Hermans, J.; Verdoes, M.; van Boeckel, C. A. A.; van Veelen, P. A.; Turk, B.; Turk, D.; Ovaa, H. The alkyne moiety as a latent electrophile in irreversible covalent small molecule inhibitors of Cathepsin K. *J. Am. Chem. Soc.* **2019**, *141*, 3507-3514.
71. Tsuboi, K.; Bachovchin, D. A.; Speers, A. E.; Spicer, T. P.; Fernandez-Vega, V.; Hodder, P.; Rosen, H.; Cravatt, B. F. Potent and selective inhibitors of glutathione S-transferase omega 1 that impair cancer drug resistance. *J. Am. Chem. Soc.* **2011**, *133*, 16605-16616.
72. Ramkumar, K.; Samanta, S.; Kyani, A.; Yang, S.; Tamura, S.; Ziemke, E.; Stuckey, J. A.; Li, S.; Chinnaswamy, K.; Otake, H.; Debnath, B.; Yarovenko, V.; Sebolt-Leopold, J. S.; Ljungman, M.; Neamati, N. Mechanistic evaluation and transcriptional signature of a glutathione S-transferase omega 1

inhibitor. *Nat. Commun.* **2016**, *7*, 13084.
73. Menon, D.; Innes, A.; Oakley, A. J.; Dahlstrom, J. E.; Jensen, L. M.; Brüstle, A.; Tummala, P.; Rooke, M.; Casarotto, M. G.; Baell, J. B.; Nguyen, N.; Xie, Y.; Cuellar, M.; Strasser, J.; Dahlin, J. L.; Walters, M. A.; Burgio, G.; O'Neill, L. A. J.; Board, P. G. GSTO1-1 plays a pro-inflammatory role in models of inflammation, colitis and obesity. *Sci. Rep.* **2017**, *7*, 17832.
74. Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. USA.* **2003**, *100*, 1603-1608.
75. Galbiati, A.; Zana, A.; Conti, P. Covalent inhibitors of GAPDH: From unspecific warheads to selective compounds. *Eur. J. Med. Chem.* **2020**, *207*, 112740.
76. Zhao, M.; Kognole, A. A.; Jo, S.; Tao, A.; Hazel, A.; MacKerell, A. D. GPU-Specific Algorithms for Improved Solute Sampling in Grand Canonical Monte Carlo Simulations. **2023**, manuscript under review.

TOC