RESEARCH ARTICLE SUMMARY

PROTEOMICS

Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells

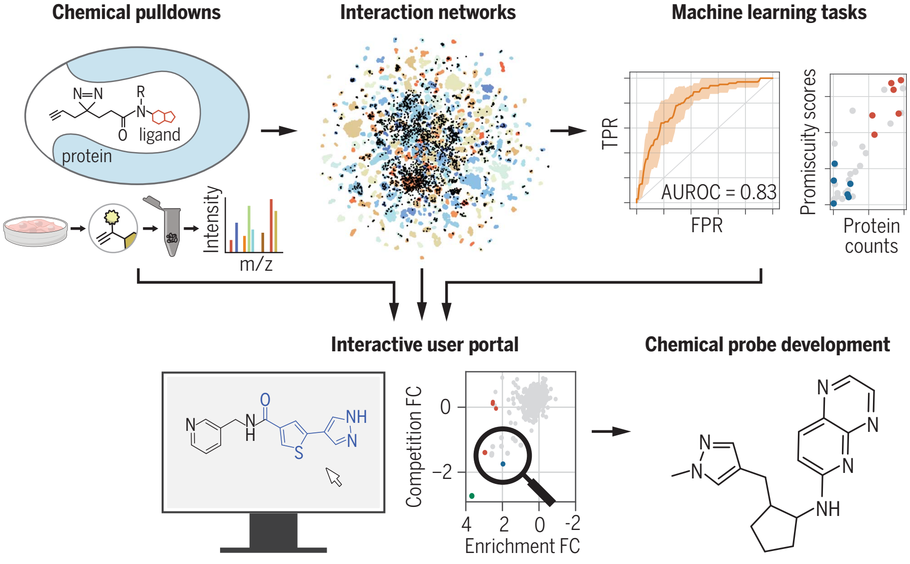
Fabian Offensperger†, Gary Tin†, Miquel Duran-Frigola†, Elisa Hahn, Sarah Dobner, Christopher W. am Ende, Joseph W. Strohbach, Andrea Rukavina, Vincenth Brennsteiner, Kevin Ogilvie, Nara Marella, Katharina Kladnik, Rodolfo Ciuffa, Jaimeen D. Majmudar, S. Denise Field, Ariel Bensimon, Luca Ferrari, Evandro Ferrada, Amanda Ng, Zhechun Zhang, Gianluca Degliesposti, Andras Boeszoermenyi, Sascha Martens, Robert Stanton, André C. Müller, J. Thomas Hannich, David Hepworth, Giulio Superti-Furga, Stefan Kubicek, Monica Schenone, Georg E. Winter*

INTRODUCTION: Chemical modulation of protein function is an important experimental approach to illuminate biological mechanisms and represents the most frequently used strategy to treat human disease. Nevertheless, around 80% of the human proteome lacks annotated small-molecule ligands, thus leaving many proteins, including validated disease targets, outside the reach of mechanistic elucidation and therapeutic innovation.

RATIONALE: To close this gap, unbiased approaches to advance ligand discovery are urgently needed. We set out to determine the proteome-wide binding preferences of more than 400 small-molecule fragments through a chemoproteomics strategy that is based on treatment of intact cells. With these data at hand, we aimed to (i) identify hundreds of fragment-protein interactions and advance selected fragments toward cell-active ligands, (ii) leverage machine learning (ML) binary classifiers to develop models to predict small-molecule behavior in native biological systems, and (iii) build an interactive open-source interface to empower the broad exploration of the data and of all predictive models.

RESULTS: Through this quantitative chemoproteomics strategy, we experimentally determined the interactome of 407 small-molecule fragments. This led to the identification of 47,658 discrete fragment-protein interactions involving more than 2600 proteins, of which 86% previously lacked any annotated ligand. To provide evidence for the translational potential of these starting points, we advanced various hits toward elaborated fragments. With focused synthetic efforts, we develo ligands that (i) engage the E3 ligase ada protein DDB1, (ii) functionally block the human equilibrative nucleoside transporter SLC29A1 (hENT1), or (iii) selectively inhibit a subset of cyclin dependent kinases (CDKs), including the orphan CDK16. In addition to advancing individual fragment-protein hits, we leveraged the depth of the global dataset to develop an ML framework to build models that can predict how fragments interact with native proteins on a proteome-wide scale. This framework included inference of quantitative fragment interactomes, which enabled us to predict to how many proteins a given fragment will bind and whether the bound proteins themselves are chemically broadly accessible or otherwise typically refractory to small-molecule ligands. Moreover, ML models allowed us to capture and predict qualitative interactome signatures. This made it possible for us to investigate and predict whether fragments tend to interact with subsets of proteins of coherent function, such as transporters or RNA-binding proteins. Likewise, ML models allowed us to analyze and predict whether fragments tend to interact with groups of proteins that reside in defined subcellular localizations or compartments, such as lysosomes or mitochondria, which can be indicative of intracellular fragment partitioning and accumulation. Last, we have also provided a platform to develop bespoke ML models that are based on a user-defined input of target proteins, and hence enable the prediction of fragment binding to a custom set of proteins.

CONCLUSION: Our large-scale chemical proteomics survey led to the identification of hundreds of fragment-protein interactions that are poised for future exploration and chemical optimization. Moreover, we found that the generated data is amenable to ML-based models that enable us to predict how chemical matter interacts with native proteomes in intact cells by using their chemical structure as input. To maximize the practical use for the scientific community, all interactomes, enrichment tools, and ML models have been made publicly available for exploration through a web-based application (https://ligand-discovery.ai). Collectively, these data and tools should form a resource to interpret fragment-binding data and expedite ligand discovery efforts. ■



Schematic representation of the ligand discovery approach. Chemoproteomics was used to assess 407 small-molecule fragments. Hundreds of fragment-protein interactions were identified as starting points for probe development. System-level analyses coupled to machine learning enabled prediction of fragment binding and behavior in living cells. An interactive web resource has been provided for data exploration, which also allows the generation and application of bespoke predictive models.

The list of author affiliations is available in the full article online.
***Corresponding author. Email: gwinter@cemm.oeaw.ac.at**
†These authors contributed equally to this work.
Cite this article as F. Offensperger *et al.*, *Science* 384, eadk5864 (2024). DOI: 10.1126/science.adk5864

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adk5864

RESEARCH ARTICLE

PROTEOMICS

Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells

Fabian Offensperger[1]†, Gary Tin[1]†, Miquel Duran-Frigola[1,2]†, Elisa Hahn[1], Sarah Dobner[1], Christopher W. am Ende[3], Joseph W. Strohbach[4], Andrea Rukavina[1], Vincenth Brennsteiner[1], Kevin Ogilvie[3], Nara Marella[1], Katharina Kladnik[1], Rodolfo Ciuffa[1]‡, Jaimeen D. Majmudar[4], S. Denise Field[4], Ariel Bensimon[1], Luca Ferrari[5,6], Evandro Ferrada[1], Amanda Ng[1], Zhechun Zhang[7], Gianluca Degliesposti[1], Andras Boeszoermenyi[1], Sascha Martens[5,6], Robert Stanton[7], André C. Müller[1], J. Thomas Hannich[1], David Hepworth[4], Giulio Superti-Furga[1,8], Stefan Kubicek[1], Monica Schenone[4], Georg E. Winter[1]*

Chemical modulation of proteins enables a mechanistic understanding of biology and represents the foundation of most therapeutics. However, despite decades of research, 80% of the human proteome lacks functional ligands. Chemical proteomics has advanced fragment-based ligand discovery toward cellular systems, but throughput limitations have stymied the scalable identification of fragment-protein interactions. We report proteome-wide maps of protein-binding propensity for 407 structurally diverse small-molecule fragments. We verified that identified interactions can be advanced to active chemical probes of E3 ubiquitin ligases, transporters, and kinases. Integrating machine learning binary classifiers further enabled interpretable predictions of fragment behavior in cells. The resulting resource of fragment-protein interactions and predictive models will help to elucidate principles of molecular recognition and expedite ligand discovery efforts for hitherto undrugged proteins.

Chemical probes provide means to modulate protein function that are complementary to genetic perturbations (*1*). The impact of small molecules on biological circuits is immediate, thus facilitating the delineation of causality (*2*, *3*). Moreover, chemical probes enable dose-resolved modulation, and their effect can transcend antagonism and agonism, for instance by stabilizing protein-protein interactions to generate gain-of-function effects (*2*, *4*). Chemical modulation of protein function has historically represented the backbone of most medicines; despite the rise of other modalities, most Food and Drug Administration (FDA)–approved drugs are small molecules (*5*). However, although the importance of small molecules is undeniable, around 80% of all human proteins still lack chemical ligands, motivating community-wide ligand discovery efforts (*6*, *7*).

The paucity of chemical probes can be rationalized in part by limitations intrinsic to frequently used discovery strategies. Target-centric high-throughput screens (HTS) enable the exploration of large (~10^6) compound libraries but are frequently performed in reductionistic conditions. Phenotypic drug discovery addresses these shortcomings but depends on often time-consuming target-identification strategies (*8*). Fragment-based ligand discovery (FBLD) is an orthogonal strategy that has been successfully employed to overcome limitations of HTS (*9*). Smaller libraries (~10^3) of moderate–molecular weight (MW <300 Da) fragments can reveal ligand-efficient starting points for subsequent medicinal chemistry campaigns. Given the low affinity of fragment hits, FBLD has, until recently, been confined to in vitro setups (*10*). However, pioneering work has outlined the feasibility of conducting FBLD in native biological systems by coupling cellular-fragment treatment with unbiased chemoproteomics. Initial efforts predominantly focused on electrophilic fragments (*11*–*15*). More recently, the concept has been expanded to map reversible fragment-protein interactions. This has been achieved by embedding photoreactive and bioorthogonal reporter groups into "fully functionalized fragments" (FFFs) to facilitate target capture (*16*). Seminal work by Parker and Cravatt has demonstrated that reversible ligand-protein interactions can be detected on previously unliganded proteins (*16*). Although selected examples could be advanced to cell-active probes, the narrow set of 10 to 20 profiled fragments can complicate the separation of specific ligand-protein pairs from promiscuous interactors resulting from labeling biases (*17*, *18*).

In this study, we set out to perform a large cellular FBLD screening campaign, aiming to address three points. First, we needed to see sufficient signal to confidently discern background from genuine fragment-protein interactions (*17*, *18*). Second, and most critically, by mapping the proteome-wide interactome of hundreds of ligands, we wanted to provide actionable chemical starting points to target thousands of proteins lacking annotated ligands. Last, we intended to leverage the dataset to identify, through factorization of the chemoproteomics matrix, patterns of small molecule–protein interactions and to enable machine learning (ML)–based predictions of target classes and cellular behavior of untested fragments. These screening results and ML classifiers jointly constitute a publicly available, interactive resource (https://ligand-discovery.ai) to identify small molecules given a protein target or to predict general interaction properties and potential interactome patterns given a small molecule.

Global analysis of large-scale fragment profiling through quantitative chemoproteomics

To determine the inventory of proteins that can be liganded with a set of fragments ("ligandability maps") at scale, we adapted previously published approaches and designed a library of 407 FFFs that features a diazirine moiety as the photoreactive element, as well as an alkyne handle as the bioorthogonal reporter group (*16*). This set of compounds was selected from a reference library of ~6,000 FFFs, ensuring that they cover a broad chemical space and drug-like properties (Fig. 1, A and B, and fig. S1A). To optimally capture structural and physicochemical properties of our FFF collection and to facilitate ensuing data analysis tasks, we subsequently developed a bespoke FFF-specific descriptor. The descriptor is based on a dense numerical representation of the fragments attached to the diazirine and alkyne groups and trained across a library of ~300,000 FFF compounds [supplementary materials (SM), materials and methods].

To comprehensively determine protein interactors of a given fragment, human embryonic kidney–293T (HEK293T) cells were treated with 50 μM of each FFF, followed by ultraviolet (UV) crosslinking to covalently capture otherwise-reversible binding events. Afterwards, cells were harvested, washed with cold phosphate-buffered saline, lysed, and clicked with biotin-azide for subsequent streptavidin enrichment and several stringent washing steps (*19*). Protein identification was performed by means of on-bead trypsin digestion coupled to mass spectrometry (MS)–based quantitative proteomics with isobaric tags. Implementation

---

[1] CeMM, Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria. [2] Ersilia Open Source Initiative, Cambridge CB1 3DE, UK. [3] Medicine Design, Pfizer Worldwide Research and Development, Groton, CT 06340, USA. [4] Medicine Design, Pfizer, Cambridge, MA 02139, USA. [5] Max Perutz Labs, Vienna Biocenter Campus (VBC), Vienna Biocenter 5, 1030 Vienna, Austria. [6] University of Vienna, Max Perutz Labs, Vienna Biocenter 5, 1030 Vienna, Austria. [7] Molecular Informatics, Machine Learning and Computational Sciences, Early Clinical Development, Pfizer, Cambridge, MA 02139, USA. [8] Center for Physiology and Pharmacology, Medical University of Vienna, 1090 Vienna, Austria.

***Corresponding author. Email: gwinter@cemm.oeaw.ac.at**

†These authors contributed equally to this work.

‡Present address: Alloy Therapeutics, Allschwil, Switzerland.

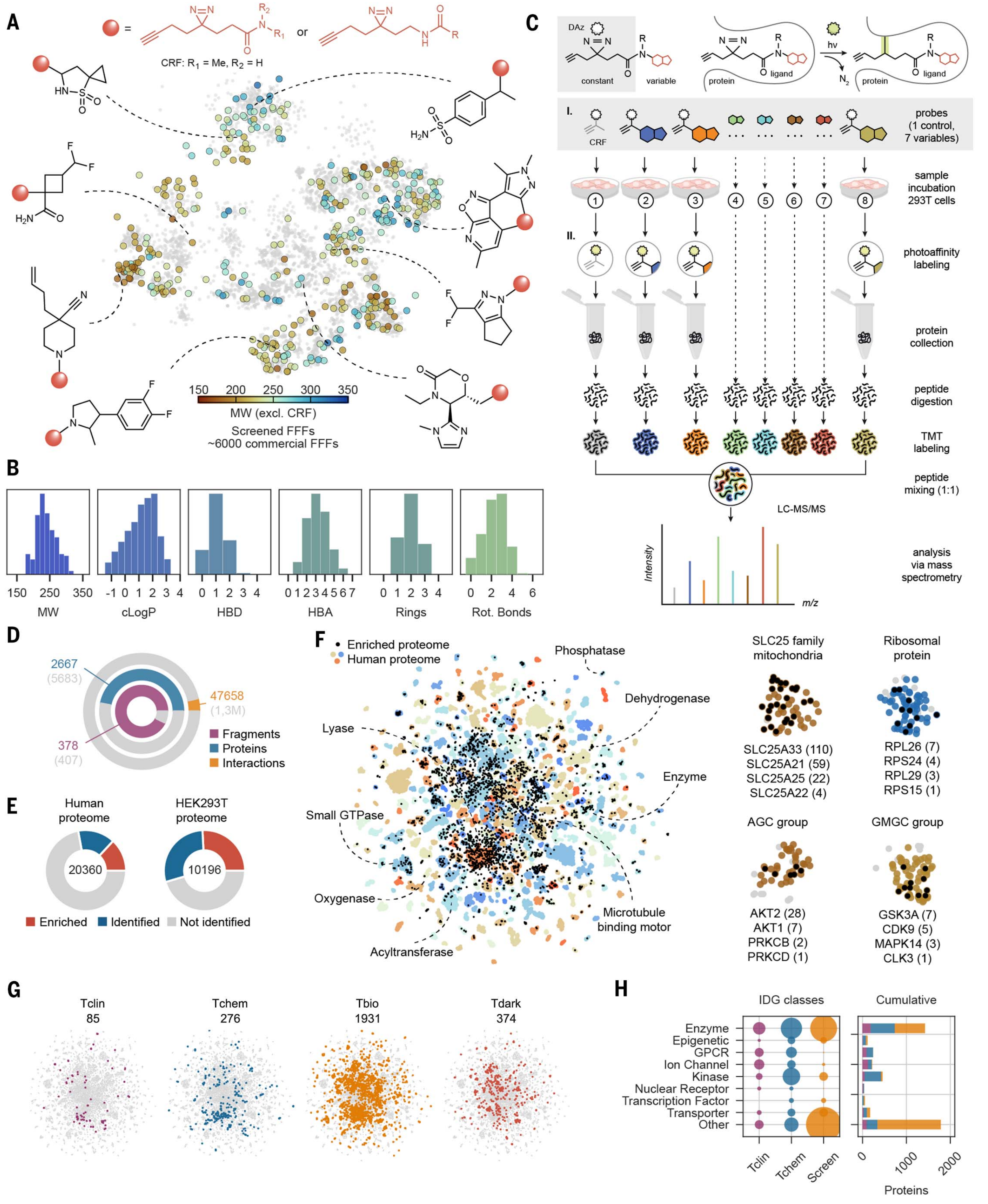


Fig. 1. Establishment and analysis of large-scale ligandability portraits of fully functionalized fragments (FFFs). (**A**) Chemical space covered by the selected 407 fragments. A two-dimensional (2D) uniform manifold approximation and projection (UMAP) visualization of the 512-dimensional FFF descriptor is shown, computed for the ~6000 FFFs from the stock Enamine collection. Screened fragments are highlighted and colored by molecular weight (MW). MW is calculated excluding the diazirine (DAz) and alkyne regions that constitute the constant region fragment (CRF). Chemical structures of eight exemplary FFFs are shown. (**B**) Molecular properties of the selected compounds (excluding the CRF region): MW, Wildman-Crippen calculated LogP (cLogP), hydrogen bond acceptors and donors (HBA, HBD), number of rings, and number of rotatable bonds (Rot. Bonds). (**C**) Schematic depiction of the fragment screening workflow. LC-MS/MS, liquid chromatography tandem mass spectrometry; *m/z*, mass/charge ratio. (**D**) Proportion of enriched fragment-protein interactions with respect to all detected ones (orange), enriched proteins with respect to all detected (blue), and fragments with at least one enriched protein with respect to all assayed FFFs (purple). (**E**) Coverage of the human and HEK293T proteomes. Proportion of enriched (hit) proteins found in at least one fragment are shown in red; remaining proteins that were detected at least once, but not enriched, are shown in blue. (**F**) Proteome-wide view of the liganded proteome. ESM-1b

sequence embeddings were calculated for the functionally annotated human proteome, followed by 2D UMAP projection and *k*-means cluster analysis. Clusters in the 2D map are highlighted with a qualitative color scale; proteins with no clusters assigned are removed. Black dots represent proteins found in at least one ligand interactome, i.e., hit proteins. (Insets, right side) Example clusters of functionally coherent proteins. Black dots again mark hit proteins. Example proteins are indicated with the number of ligands in parentheses. AGC and GMGC refer to kinase groups. (**G**) Ligandability of hit proteins. Enriched proteins stratified by IDG (Illuminating the Druggable Genome) categories (Tclin, Tchem, Tbio, and Tdark). (**H**) Protein-family coverage by approved drug targets (purple), liganded proteins according to external databases (blue), and enriched proteins in our screen (orange). In the left plot, size of the dot is proportional to the number of proteins. Correspondingly, the right panel shows cumulative counts as a bar plot. For a more granular view, see fig. S1C. GPCR, G protein–coupled receptor.

of TMTpro 16plex reagents (TMT, tandem mass tag) enabled the parallel interactome determination of seven fragments in duplicate, comparing protein enrichment with a control reference fragment (CRF; methyl) serving as a null diazirine control (Fig. 1, A and C; see SM, materials and methods for details) (*20*). In our analysis, we detected a total of 5683 distinctive proteins, 2667 (47%) of which are significantly enriched at least fivefold [$\log_2$ fold change (FC) >2.3] over the CRF background control, thus resulting in a total of 47,658 fragment-protein interactions. Together, these hits cover 93% of the screened fragments and a considerable proportion of the expressed proteome in HEK293T cells (Fig. 1, D and E, and fig. S1B). We next asked to what extent our data cover proteins that have different functional and clinical classifications. Projecting all fragment-enriched targets over a display of the human proteome indicates a broad coverage of various protein families, including known ligandable protein families, such as enzymes. Our study also revealed several other protein clusters that are traditionally challenging to ligand discovery efforts, including phosphatases and small guanosine triphosphatases (GTPases) (Fig. 1F). Last, we stratified targets by using as a reference the "Target Development Level" classification of the Illuminating the Druggable Genome (IDG) program (*21*). Although we identified chemical starting points for proteins that are targeted by clinically approved drugs and for other previously liganded proteins (Tclin and Tchem IDG categories, respectively; Fig. 1G) (*21*), our survey also revealed a remarkable number of targets (2305) that previously lacked association with any known chemical matter (Tbio and Tdark; Fig. 1G). Accordingly, in addition to covering druggable protein classes such as kinases, or other enzymes in general, most protein targets liganded by our fragments fell into categories that are underrepresented, such as scaffolding and adapter proteins or RNA splicing factors (Fig. 1H and fig. S1C). Taken together, the proteome-wide binding analysis of more than 400 FFFs has revealed thousands of fragment-protein interactions that can be explored through our interactive resource (https://ligand-discovery.ai).

Quantitative differences in fragment interactomes

To increase confidence that individual fragment-protein interactions represent actionable starting points for ligand discovery and to determine unspecific binding resulting from labeling biases, we next investigated differences in fragment and protein promiscuity. We noticed that fragments spanned a broad range of promiscuity, with some interacting with several hundred proteins, whereas others displayed a much more selective target profile (Fig. 2A). We confirmed the selectivity profile of seven fragments through in-gel fluorescence scanning, thus ascertaining that differential fragment promiscuity is observed also in proteomics-independent readouts (Fig. 2B). We found that in agreement with previous results, protein enrichment was strictly dependent on UV cross-linking, thus confirming that the identified fragment-protein interactions were of reversible nature (*16*). To eliminate potential protein contaminants from our interactomes, we compared our results with previous studies that investigated the labeling bias of diazirine-based photocrosslinkers (*17*, *18*). We surmised that the scale of our survey should allow us to expand the knowledge of background labeling biases that may be introduced by the diazirine reporter group. Indeed, among the 50 most frequently enriched proteins in our dataset, less than half had previously been associated with background labeling (Fig. 2C) (*17*, *18*). Corroborating previous observations, we found that membrane and mitochondrial proteins were among the frequently identified proteins globally (Fig. 2D) (*18*). Overall, the scale of our analysis allowed us to establish a rich labeling background, which can help to inform choice of suitable starting points for FBLD. To maximize the practical utility for the community, we added a functionality to our web resource that allows users to cross-reference their independently generated chemoproteomics data to differentiate specific from unspecific interactors.

For our own selection of follow-up fragment-protein interactions, we initially processed the data by removing all proteins liganded by 10 or more fragments. This yielded a sparse and modular fragment-protein interactome network (Fig. 2E). From this pruned interactome, we wanted to select three diverse fragment-protein interactions for further exploration and optimization. To this goal, we wanted to pick three chemically distinct fragments targeting proteins that span a range of pharmacological knowledge depth, from Tclin to Tbio (Fig. 2F). To focus on strong but specific interactions, we derived an interest score (iScore) that was defined as an enrichment ratio of an interaction over the promiscuity of the corresponding fragment and protein. We focused on the top 5% of interactions (Fig. 2F). On the basis of these criteria, we selected three fragments for validation, follow-up studies, and ensuing chemical optimization, namely C391, C186, and C27, involving interactions with CDK2 and CDK16, DDB1, and SLC29A1, respectively.

We first aimed to validate these interactions in independent experiments at lower fragment concentrations (25 μM instead of 50 μM). Next, we wanted to assess whether these interactions are chemically actionable. To address this, we sought to identify FFF analogs ("fragment elaborates") that feature the variable fragment region and additional substitutions but lack the constant region. To assay fragment elaborates, we tested them in competitive cellular binding assays with the FFF using proteomics and, where available, also in functional experiments.

Selective fragment-protein interactions can be advanced into chemical probes

C391 selectively engages a subset of cyclin dependent kinases

To benchmark our approach, we selected a target class that provides facile access to recombinant binding and inhibition data. We thus followed up on the interactome of fragment C391, a thiophene-pyrazole that elicited pronounced ($\log_2$ FC >4) and selective enrichment of CDK1, CDK2, CDK5, CDK9, and CDK16 (iScore of 2.14 and 2.16 for CDK2 and CDK16, respectively) (Fig. 3A). Closer examination revealed that 11 other CDKs were detected in the respective pulldown, yet most of them failed to be significantly enriched (Fig. 3B). This finding prompted us to investigate whether this scaffold would display a promiscuous pan-kinase binding. For this purpose, and to enable the investigation of other fragments in a similar manner, we developed an interactome–enrichment analysis functionality, which is available through our web resource. This tool takes inspiration from classic gene set enrichment analysis (GSEA) but is geared to detect enrichment of certain annotation terms within the interactome of a particular fragment of interest (*22*). Annotation categories include cellular localization and protein families, among others (SM, materials and methods). Corroborating a bias toward liganding CDKs, we detected a significant enrichment of the CDK

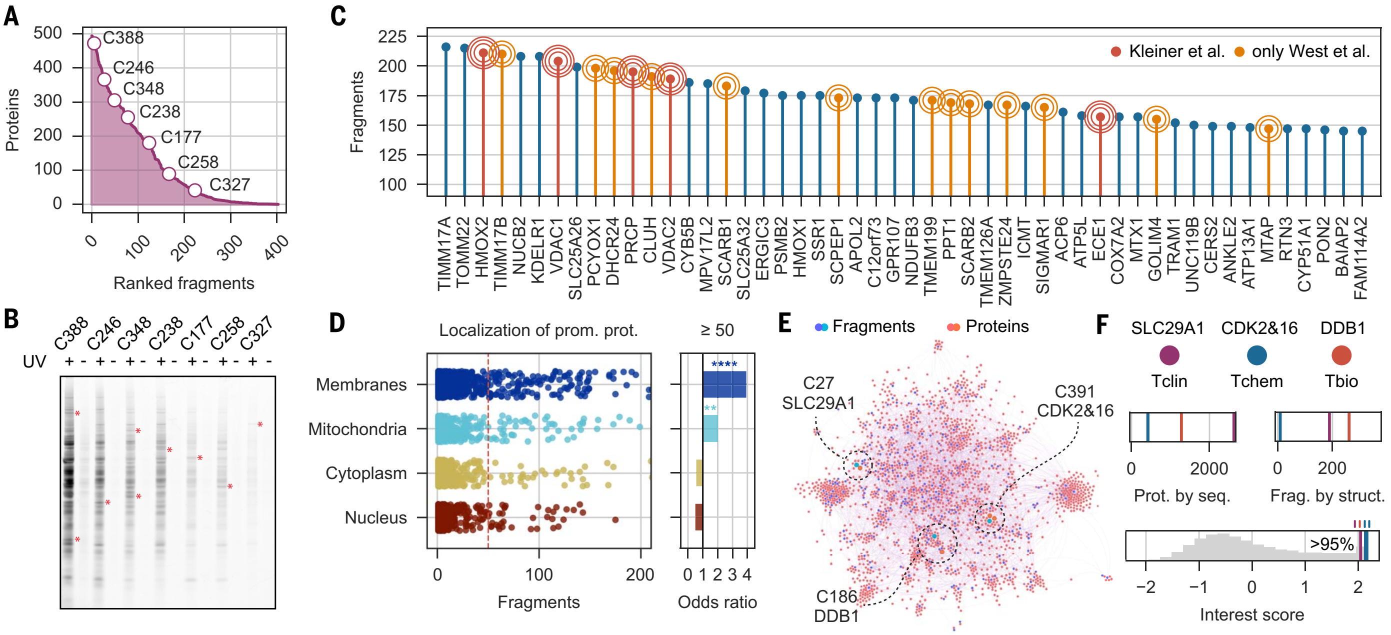


Fig. 2. Quantitative determination of fragment interactomes. (**A**) Number of enriched proteins per fragment. Fragments are ranked by protein counts. (**B**) In-gel fluorescence assay displaying fragments spanning the range of promiscuity shown in (A). Bands with preferential labeling by a certain fragment are labeled with red asterisks. (**C**) Top 50 proteins, according to the number of fragments. "Labeling-bias" proteins reported in Kleiner *et al*. and West *et al*. (*17*, *18*) are highlighted in red. Proteins identified only in West *et al*. (*18*) are shown in orange. (**D**) Cellular localization analysis. Proteins are classified in broad categories, namely membranes, mitochondria, cytoplasm, and nucleus. (Left) Promiscuous proteins (prom. prot.; higher number of fragments) tend to localize in membranes. (Right) This enrichment is quantified with a hypergeometric test under a promiscuity cutoff of 50 fragments. Membrane proteins are four times as likely to be promiscuous. (**E**) Fragment-protein interactome laid out after filtering out proteins with 10 or more fragments. Proteins are shown in red and fragments in blue. (**F**) Selected follow-up fragment-protein interactions. (First row) Proteins are colored by IDG category (Tclin, Tchem, and Tbio). (Second row) The left plot sorts proteins by their sequence similarity (Prot. by seq.), displayed in 1D after hierarchically clustering the 2D coordinates of Fig. 1F and capturing the leaf order of the dendrogram. Likewise, the right plot sorts fragments by chemical structure (Frag. by struct.) according to a hierarchical clustering of Fig. 1A coordinates. Widespread lines in both plots show diversity of selected proteins and fragments. (Third row) An interest score (iScore) for each interaction is defined as the ratio of the median-corrected $\log_2$ FC of each interaction over the geometric mean of the sum of enrichment scores of the fragment and the protein. Interest scores are power-transformed. High interest scores indicate strong and specific enrichment signals.

kinase subfamily in the interactome of C391. By contrast, no enrichment was observed when investigating the larger kinase superfamily (Fig. 3C). In response to this finding, we designed a small set of 10 chemicals elaborated around the C391 core, aiming to find compounds that could outcompete the binding of the FFF C391 to the enriched CDKs. To maintain an unbiased readout over the entire FFF interactome, we again opted for a multiplexed, MS-based, proteomics-based setup that allows us to recapitulate the initial interactome of C391 (comparison between CRF and the respective FFF), while allowing for assessment of the competitive binding pattern of the top six elaborates (fig. S2A), chosen by their predicted permeability. In agreement with the initial fragment screening, we again enriched a set of CDKs with the FFF C391. The different elaborated molecules displayed varied degrees of CDK competition, with C391-E6 showing the most pronounced competition (Fig. 3, D and E). Consistently, we could validate binding of C391-E6 to recombinant CDK2 and CDK16 (LanthaScreen, Thermo Fisher Scientific; fig. S2B and table S6). C391-E6 also potently inhibited the catalytic activity of CDK2 (Z′-LYTE assay, Thermo Fisher Scientific; fig. S2C), which is supported by molecular docking that reveals binding to the active site (fig. S2D).

C186 is a competitive binder of DDB1

Cullin RING E3 ubiquitin ligases (CRLs) are at the center of attention of targeted protein degradation, a pharmacology that is based on small molecules that induce proximity between a protein of interest (POI) and an E3 ligase, thus prompting POI degradation (*23*, *24*). We thus looked for fragments binding to CRL components (*25*–*29*). We observed a strong ($\log_2$ FC >4) enrichment of the CRL adapter protein DDB1 by the FFF C186. This enrichment was highly selective because C186 was the only profiled fragment that displayed statistically significant DDB1 enrichment (iScore of 2.14; Fig. 3F and fig. S2E). In line with a direct binding event, we could validate binding of C186 to recombinant DDB1 by in-gel fluorescence scanning (Fig. 3G). As before, we turned to additional chemoproteomics experiments to identify C186 derivatives capable of competing DDB1 from the FFF C186, revealing C186-E5 as the strongest (but also most promiscuous) competitor (Fig. 3, H and I, and fig. S2, E and F). Competitive DDB1 engagement by C186-E5 was further validated with Western blot analysis (fig. S2G). We and others have previously shown that small molecules can induce the degradation of cyclin K by recruiting the CDK12/13:CycK complex directly to DDB1, thus providing evidence that direct recruitment to DDB1 can offer a means to targeted protein degradation (*30*, *31*). Future experiments will address whether C186 binding to DDB1 has similar neomorphic features, or if C186 elaborates could be furnished into a DDB1-recruiting heterobifunctional degrader, as has recently been shown for a covalent fragment (*32*).

Discovery of an SLC29A1 inhibitor

The solute carrier (SLC) superfamily represents the biggest family of transporters (>400 members), and approximately one-quarter of SLCs are associated with human disease (*33*). Although certain SLCs are targets of approved drugs, the majority remain poorly studied (*34*). Individual SLCs are among the frequently enriched proteins in our dataset (Fig. 2, C and D)

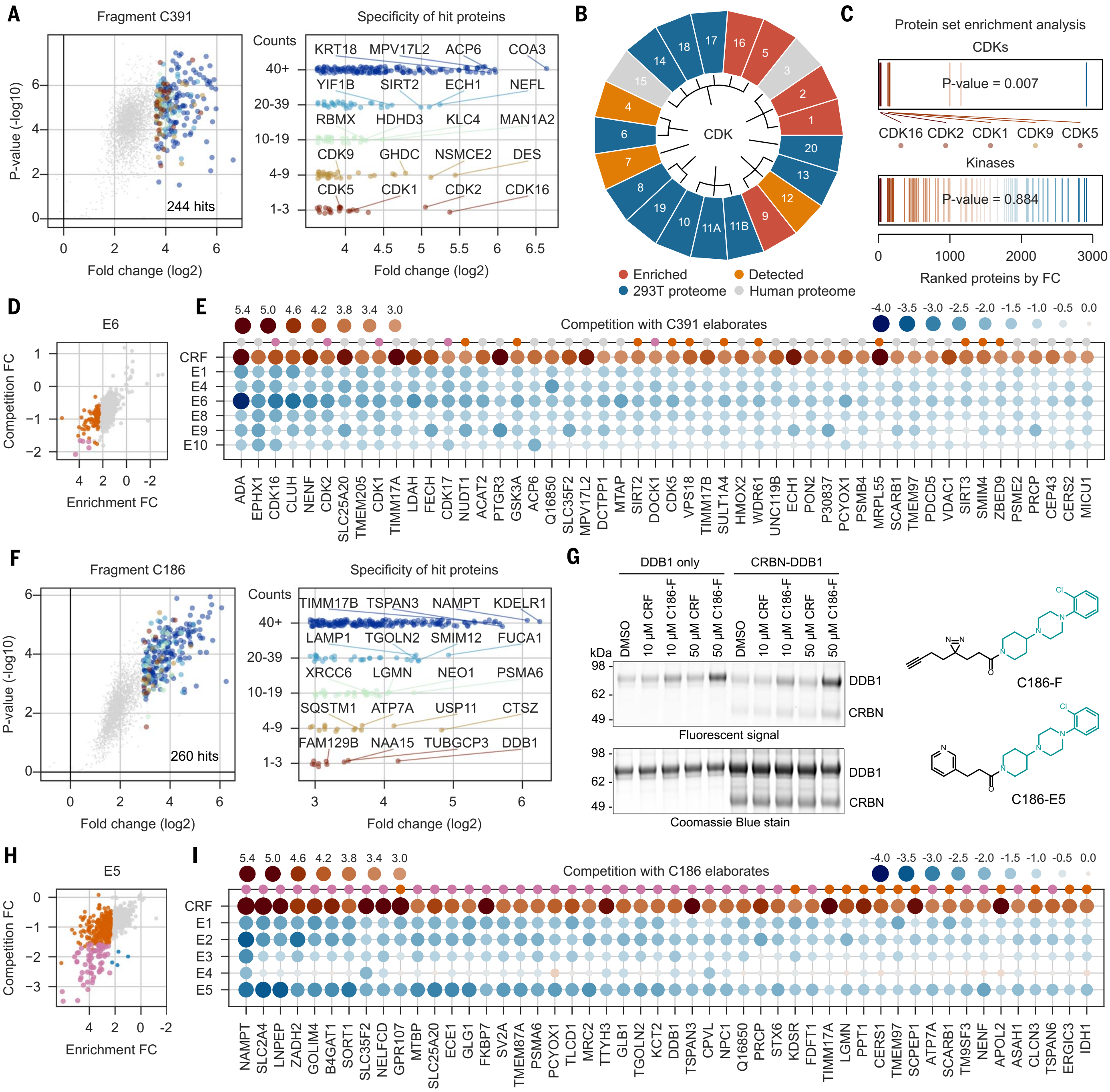


Fig. 3. Identification and elaboration of fragment hits for CDKs and DDB1. (**A**) (Left) Volcano plot related to FFF C391. Hit proteins are highlighted according to our hit-selection criteria (SM, materials and methods). (Right) Enriched (hit) proteins are stratified by their level of specificity, measured as the number of interacting FFFs (counts). (**B**) Cyclin dependent kinases (CDKs) organized in a circular plot according to a high-level phylogeny. The plot shows CDKs enriched in C391 (red), CDKs detected but not enriched (orange), remaining CDKs expressed and measured in HEK293T (blue), and the remaining CDKs in the human proteome (gray). (**C**) Protein set enrichment analysis of CDKs (upper) and the kinase superfamily (lower). Proteins are ranked by $\log_2$ FC. (**D**) Competition experiment for elaborate C391-E6. *x* axis shows the $\log_2$ FC (right to left); *y* axis shows the competition FC measured with respect to the FFF. Enriched and competed proteins are highlighted in light purple. Enriched but not competed proteins are highlighted in red. (**E**) Heatmap representing the top 50 proteins highlighted in the competition experiment. The first row corresponds to the FFF (i.e., the fragment with the CRF). The rest correspond to the derivatives. (Top left) The brown scale and size of the first row denotes enrichment $\log_2$ FC of the FFF. (Top right) The blue scale and size of the other rows indicate the elaborate competition FC. Small purple and red dots at the top show proteins highlighted in (D). (**F**) Volcano plot for C186, analogous to (A). (**G**) Recombinant purified DDB1 or DDB1-CRBN was incubated with the indicated concentration of CRF or C186-F. Photocrosslinking and click reaction with a picolyl-azide-sulfo-Cy3 as fluorophore followed by in-gel fluorescence scanning confirmed direct interaction between C186-F and DDB1. DMSO, dimethyl sulfoxide. (**H**) Competition scatterplot for C186, analogous to (D). (**I**) Competition heatmap for C186, analogous to (E).

(*17*, *18*). We reasoned that, when properly controlled for, fragments that selectively bind to SLCs would nevertheless provide suitable starting points for ensuing ligand optimization. We decided to test this hypothesis by investigating fragment C27, which we found to strongly ($\log_2$ FC >3.5) and selectively enrich the human equilibrative nucleoside transporter SLC29A1 (hENT1), which is involved in cellular uptake of purine and pyrimidine nucleosides (iScore of 2.04; Fig. 4A) (*35*). It is also a target of clinical relevance including cardiopathy, renal disorders, and hypertension (*36*). In addition, a recent study highlighted that SLC29A1 inhibition also increases extracellular levels of inosine, which forms the molecular basis in brown adipose tissue differentiation (*37*). Selective enrichment of SLC29A1 by C27 hence provided motivation to identify functional elaborates through competitive chemoproteomics experiments as described before. First, we confirmed the enrichment of SLC29A1 by the FFF C27 over CRF and also discovered that several analogs, including C27-E9, were capable of competing this interaction (Fig. 4, B and C). Strong competition could also be observed with the known SLC29A1 inhibitor nitrobenzylthioinosine (NBMPR) (Fig. 4C). Second, we confirmed enrichment of SLC29A1 by competitive C27 pulldowns coupled to Western blot analysis (Fig. 4D). Competition could be observed with several derivatives and known SLC29A1 inhibitors (NBMPR and dipyridamole) but not with an acetylated, non-CRF–containing version of C27 (fig. S3A). To enable further design and derivatization of C27 elaborates, we

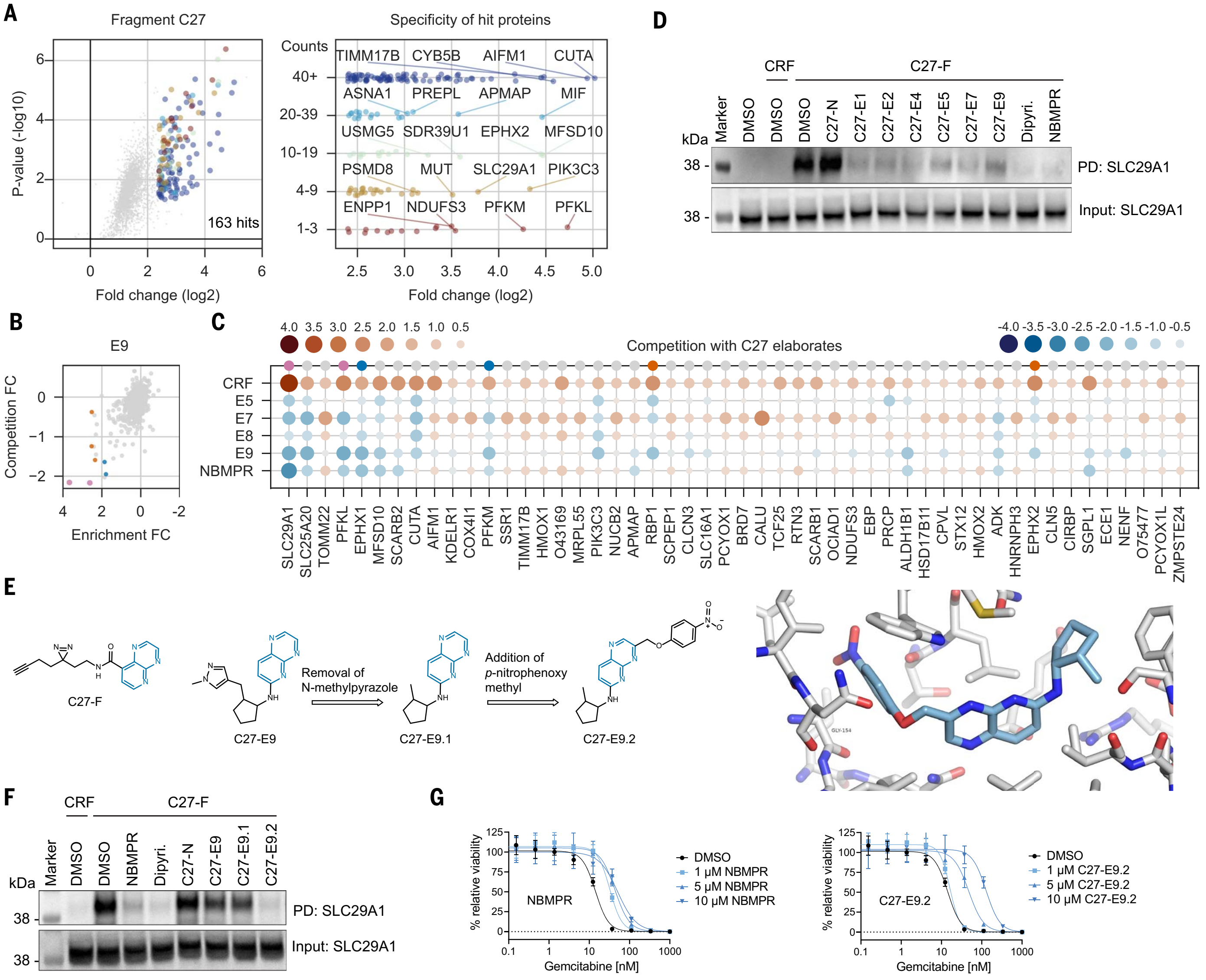


Fig. 4. Discovery of an SLC29A1 inhibitor. (**A**) Volcano plot for fragment C27 (as shown in Fig. 3A). (**B**) Scatter plot for the competitive chemoproteomics experiment for C27-E9 (as shown in Fig. 3D). Competed proteins that lack sufficient enrichment are displayed in blue. (**C**) Heatmap showing the top 50 competed protein targets and competition ratios for C27-E5, C27-E7, C27-E8, C27-E9, and NBMPR (as shown in Fig. 3E). (**D**) Competitive pulldown with C27 elaborates coupled to Western blot analysis. PD: pulldown; Dipyri., dipyridamole. (**E**) Chemical structures of C27-F and the development of fragment C27-E9 over C27-E9.1 to C27-E9.2. Docking model of C27-E9.2 (blue) in SLC29A1 (Protein Data Bank ID: 6OB6) (*35*); residues within a radius of 5 Å are shown in gray. (**F**) Competitive pulldown as shown in (D) with elaborated competitors. (**G**) Functional uptake assay for SLC29A1. KBM7 cells were cotreated with the indicated concentrations of test compound (NBMPR or C27-E9.2) and the indicated concentration of gemcitabine. Cell viability was determined after 48 hours of treatment. A nonlinear fit function [four parameter logistic (4PL) sigmoidal curve fit] was applied with error bars indicating standard deviations. Triplicate analysis.

turned to in silico iterative docking studies using NBMPR in its crystal structure with SLC29A1 as a template (Protein Data Bank ID: 6OB6) (*35*). This approach revealed that the pyridopyrazine core of C27-E9 aligned with the adenine core of NBMPR, whereas the *N*-methylpyrazolmethylcyclopentamine elaboration overlapped with the ribose moiety (fig. S3B). The *p*-nitrobenzyl ring of NBMPR was shown to interact with a hydrophobic pocket present only in SLC29A1 and not in other members of the SLC29 family, which harbors Gly154 rather than serine in the analogous position (*35*). To access this pocket, we designed C27-E9 to have the same *p*-nitrobenzyl extension at position 3 of the pyridopyrazine core, while removing the *N*-methylpyrazole for synthetic feasibility to furnish C27-E9.1 and C27-E9.2. Docking C27-E9.2 into SLC29A1 indeed suggested that it acquires a binding mode similar to that of NBMPR (Fig. 4E). In competitive pulldown experiments, C27-E9.1 was equipotent with the parental C27-E9, but C27-E9.2 displayed enhanced competition comparable to NBMPR (Fig. 4F). SLC29A1 is also the main transporter of a variety of toxic nucleoside anticancer drugs, such as cytarabine and gemcitabine (*38*, *39*). We thus set out to develop an uptake assay in which SLC29A1 inhibition can be measured through shifts in cellular viability upon cotreatment with gemcitabine (fig. S3C). In agreement with competitive pulldown experiments, C27-E9 and C27-E9.1 exhibited minimal rescue to gemcitabine-induced cytotoxicity, whereas C27-E9.2 showed a dose-dependent effect in the low-micromolar range, comparable to NBMPR (Fig. 4G and fig. S3D).

Having established that we could prioritize individual fragment-protein interactions to furnish competitive and often functionally consequential ligands, we next wanted to evaluate how the global dataset of tens of thousands of fragment-protein interactions could be leveraged to discover patterns or general principles of how fragments interact with the proteome in native biological systems. In this context, two properties stood out: first, ligand promiscuity, second, and interrelatedly, the association of a specific FFF (and, by extension, specific chemical properties) with certain cellular locations, protein families, and protein functions.

Fragment promiscuity prediction

We have already observed that analyzed fragments fall on a broad spectrum of selectivity and promiscuity profiles. To model this differentially promiscuous behavior, we first leveraged the aforementioned FFF descriptors and coupled them to a fast, lightweight, and fully automated ML algorithm for binary classification (Fig. 5A and SM, materials and methods).

In brief, we first labeled screened fragments as promiscuous (1) or nonpromiscuous (0), according to thresholds in protein-interaction counts. Then, we used a transformer-based ML model (TabPFN) to map a compound's FFF descriptor to a classification score (0 or 1) (*40*). TabPFN is a fully learned model that approximates Bayesian inference and requires no hyperparameter tuning, making it straightforward to obtain performant ML classifiers based on our chemoproteomics profiling data. Through this approach, promiscuity models can also inform on the specificity of the bound proteins. For instance, a relatively specific fragment might only ligand "frequent-hitter" proteins, whereas a more promiscuous fragment could be enriched in binding to proteins that are very rarely enriched. (fig. S4A). Accuracy of our ML-based promiscuity prediction models was measured in 80:20 train-test splits and yielded satisfactory AUROC (area under the receiver operating characteristic) scores of ~0.8 (Fig. 5B and fig. S4A). A promiscuity prediction module is available through our ligand discovery resource. Next, we wanted to validate the ML models experimentally and further investigate their interpretability, aiming to connect promiscuity predictions with a set of tangible and explainable molecular features. To accomplish this, we comprehensively predicted the binding promiscuity of >5500 FFFs (Fig. 1A and table S3). We implemented a simple tree-based method to explain ML model outcomes using a set of 100 interpretable physicochemical properties selected from an initial pool of >1600 features on the basis of their importance to the model outcome (*41*). We laid out these top 100 features in a 10 × 10 tile according to their intercorrelation (fig. S4B), resulting in easily recognizable regions related to lipophilicity, size, aromaticity, and other attributes (*41*). Importance of each property was estimated with Shapley value scores, which quantify the impact of a given property on the overall prediction outcome, revealing outsized contributions of LogP and the number of aromatic sp^2 carbons (Fig. 5C) (*42*). Ranking the virtually assayed 5680 FFFs by predicted promiscuity scores, we acquired six predicted promiscuous and six predicted nonpromiscuous FFFs for experimental validation, aiming to cover a range of chemical properties and LogP values (table S3). Both in-gel fluorescence (Fig. 5D) and chemoproteomics readouts are in strong agreement with the predicted scores (Fig. 5, E and F, and fig. S4C). The ML models could correctly label challenging cases, such as promiscuous fragments with modest LogP, as well as nonpromiscuous fragments in the high LogP range (Fig. 5G).

Functional and spatial target signatures

Enrichment analysis of fragment interactomes can reveal the preferential binding of fragments to certain protein families. The fragment profile of C391, for instance, pointed to ligand binding to CDKs (Fig. 3, B and C). Moreover, enrichment of proteins corresponding to a specific cellular compartment could provide evidence for intracellular fragment partitioning. For example, we found that fragment C310 strongly interacts with the autophagy receptors SQSTM1 (p62), NBR1, and TAX1BP1, thus making it, in theory, an interesting candidate to modulate autophagic clearance (*43*, *44*). However, ensuing protein set enrichment analysis of the C310 interactome revealed a pleiotropic enrichment for other proteins localized in the autophagosome and lysosome (fig. S5, A to C). Based on this interactome analysis, we surmised that the enrichment of autophagy receptors through C310 was likely an indirect consequence of fragment partitioning rather than the consequence of a characteristic molecular recognition. Indeed, neither recombinant SQSTM1, NBR1, nor TAX1BP1 could be labeled by in-gel fluorescence scanning assays (fig. S5D), whereas confocal microscopy furnished clear evidence of lysosomal accumulation of C310 that also extended to accumulation in autophagosomes (fig. S5, E and F). Hence, proteomics-based interactome profiling allowed us to reveal intracellular fragment partitioning.

Motivated by these findings, we have provided extensive functionalities to analyze FFF enrichments and to build ML models for protein sets of choice in our ligand discovery resource. For illustrative purposes, and to offer a bird's eye view of our interactome, we aimed at segregating FFFs on the basis of 10 coherent and frequently observed groups of enriched proteins. We focused on gene ontology (GO) molecular functions (MF) and cellular component (CC) terms, opting for a simple analysis in which only one or a few interactome signatures are assigned to each FFF. We first pruned FFF interactomes, retaining only nonpromiscuous proteins accounting for MF and CC enrichment signals. This favored subsequent MF and CC coherence of the emerging signatures and narrowed down the analysis to a subset of 970 informative proteins. Next, we applied nonnegative matrix factorization (NMF) to decompose the sparse fragment-protein matrix into a fragment-signature (W; 407 × 10) and a signature-protein (H; 10 × 970) matrix. W denotes the association of each fragment to each signature (Fig. 6A and table S4), and H indicates the protein composition of the signatures (table S5). This procedure revealed a global structure in our FFF profiles, displaying groups of 20 to 40 (5 to 10%) fragments with relatively specific signature patterns (Fig. 6, B and C). Subsequent unbiased enrichment analysis allowed us to assign biological meaning to the respective signatures. For instance, signature 7 was clearly enriched in tubulins (TUBB), signature 1 was characterized by RNA binding proteins, and signature 2 consisted mainly of lysosomal proteins (Fig. 6D). Next, we asked whether our ML modeling framework could be employed to predict interactome signatures for

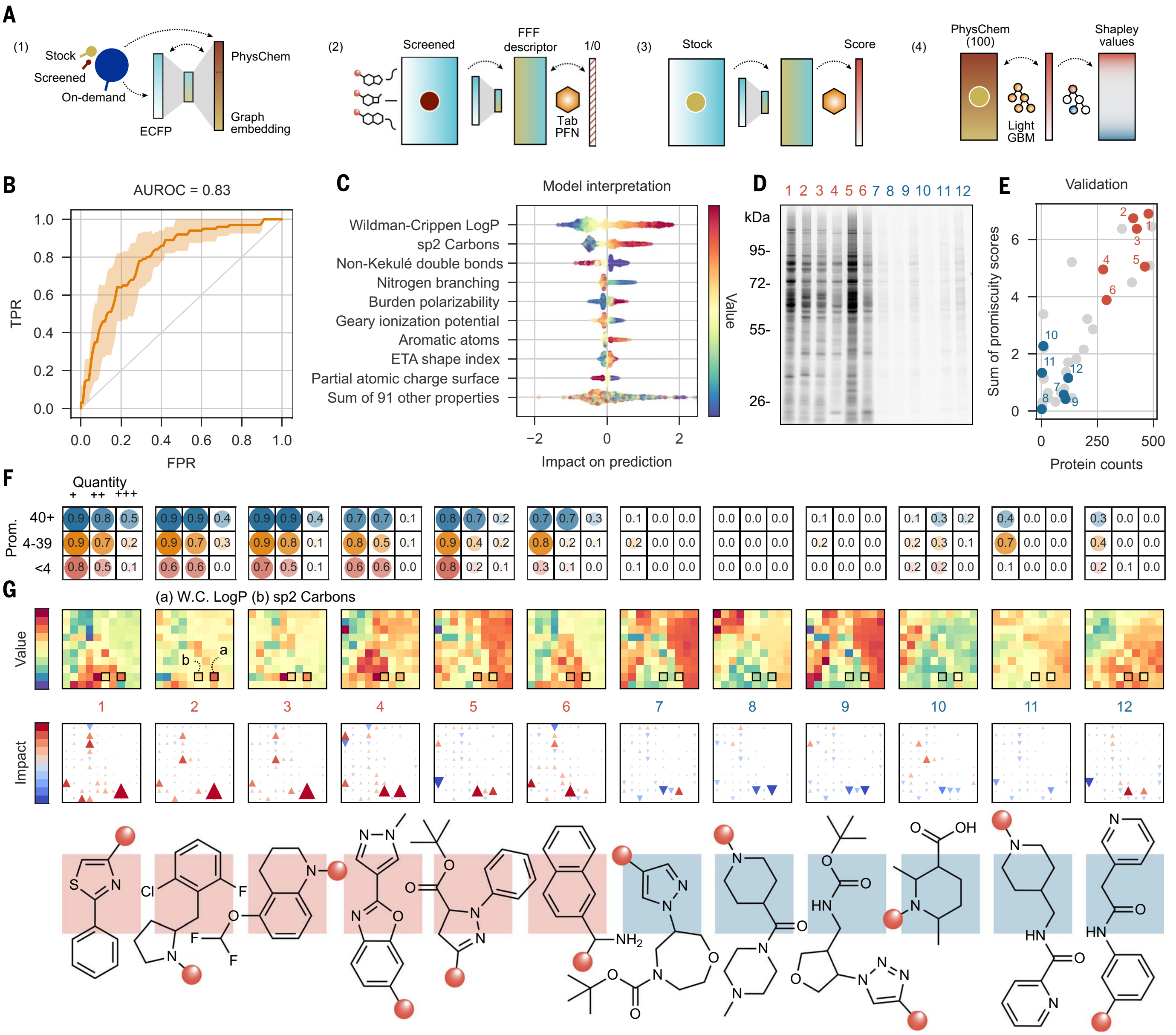


Fig. 5. Fragment promiscuity prediction. (**A**) Scheme of the ML methodology, consisting of a transfer-learning approach in which (1) the FFF chemistry is captured in an FFF descriptor based on a large compound collection (our screened FFFs, Enamine stock FFFs, and Enamine REAL on-demand FFFs; >250,000 in total), which is, in turn, (2) the input layer for binary (1/0) tasks such as promiscuous/nonpromiscuous classification. ML model training is done with screening data (407 compounds). Then, the trained pipeline can be applied prospectively, for example, (3) across Enamine stock FFFs (~6000), to obtain a classification score. For interpretability, (4) surrogate modeling can be done using human-readable molecular features and a tree-based algorithm such as LightGBM (light gradient boosting machine) coupled to Shapley value analysis. We built these interpretable models based on the prescreened Enamine stock library (SM, materials and methods). PhysChem, physicochemical properties; ECFP, extended-connectivity fingerprints. (**B**) Binary classification performance in 10 stratified 80:20 train-test splits. The current task corresponds to a promiscuity cutoff of 100 proteins. The average ROC curve across splits is shown, shaded with the standard deviation interval. TPR, true positive rate; FPR, false positive rate. (**C**) Shapley values observed for 20% (test) Enamine stock fragments. Molecular features are ranked by absolute sum of Shapley values (impact on prediction) across evaluated compounds. The color scale indicates the feature value (high is red; low is blue). Thus, high (red) LogP values have a positive impact on the prediction score, meaning that they contribute to promiscuity. ETA, extended topochemical atom. (**D**) In-gel fluorescence assay for six predicted promiscuous (red) and six predicted nonpromiscuous (blue) FFFs. (**E**) Correlation between predicted promiscuity (*y* axis) and observed promiscuity (*x* axis) in prospective chemoproteomics assays for 35 FFFs. The predicted promiscuity score is the sum of all scores across all individual promiscuity predictors. FFFs presented in (H) are highlighted in red and blue. (**F**) Individual promiscuity (Prom.) predictor scores for the 12 highlighted FFFs, arranged in a 3 × 3 grid as follows. Rows stratify proteins by specificity, measured as the number of fragments per protein [bottom, less than 4 FFFs (1%); middle, between 4 and 39 FFFs; top, 40 or more FFFs (10%)]. Columns count the number of proteins in each of the specificity categories, + being "at least a few" proteins and +++ being "many." Thus, the top right of each FFF shows fragments that bind to many (+++; 200) nonspecific (40+ FFFs) proteins. (Bottom row) Levels are +, 5; ++, 10; and +++, 50. (Middle row) Levels are +, 10; ++, 50; and +++, 100. (Upper row) Levels are +, 50; ++, 100; and +++, 200. (**G**) Interpretability of predictions. 100 molecular features are organized in a 10 × 10 tile, such that neighboring positions correspond to

correlated features (SM, materials and methods). (First row) Cells are colored by molecular property value (high is red; low is blue). As shown, selected compounds have diverse feature profiles. Two relevant features, namely Wildman-Crippen (W.C.) LogP and number of sp^2 carbons, are highlighted to ease readability (details provided in fig. S4B). (Second row) Shapley values of the features are shown, highlighting the most relevant features for the predicted outcome. Big red triangles pointing upwards denote a positive impact; big blue triangles pointing downwards denote a negative impact. Small or inexistent triangles indicate negligible impact. (Bottom row) Structures of the 12 compounds.

novel FFFs, thus anticipating their behavior in native biological systems. For 80:20 train-test splits, 6 of the 10 signatures could be predicted with AUROC accuracy >0.7, of which 3 showed AUROC >0.8 (Fig. 6E). Signature 2, corresponding to lysosomal proteins, stood out as a highly predictable trait (AUROC >0.9) that we could validate experimentally for 13 additionally acquired fragments by confocal microscopy (fig. S5G). Lysosomal accumulation is a well-described phenomenon displayed by many small molecules, particularly those hydrophobic and weakly basic in nature, including clinically approved drugs (*45*–*47*). Consistently, profiled FFFs with interactomes that enrich signature 2 indeed predominantly feature a basic (tertiary) amine (as for instance C310, fig. S5A), as do most of the predicted fragments (table S3). Together, these results provide evidence that chemoproteomics profiling can correctly capture small-molecule behavior in intact cellular systems. Moreover, they demonstrate that higher-order fragment behavior—such as partitioning into cellular compartments—is reflected in differential interactome signatures and that associated ML strategies can predict which fragments will share such features and hence will behave similarly in a cellular context. Motivated by data around the well-known phenomenon of lysosomal partitioning, we next set out to explore the further predictive power of other signatures.

To this aim, we collected 16 proteomics profiles available in the public domain, alongside a prospective validation set of 35 newly purchased Enamine stock FFFs (*16*, *18*). We then applied our ML inference pipeline to predict signature association on the basis of FFF chemical structure alone. Globally, we observed a strongly significant coincidence (empirical P value $<10^{-5}$) between predicted signatures and top-ranked proteins in the FFF interactomes (SM, materials and methods, and fig. S6, B and C). Removing signature 2 from the analysis (arguably the most trivial task) still resulted in a significant trend (fig. S6C). In Fig. 6F and fig. S7, we depict some illustrative cases. For example, JN00026 and JN00033 fragments from West *et al.* were successfully predicted to enrich transmembrane transporters (signature 6), whereas three of the FFFs from Parker *et al.* (PRK4, PRK6, and PRK7) enriched proteins related to RNA binding (signature 1) (*16*, *18*). Other notable results include compound V0019, predicted to enrich signature 1 and signature 3 (proteasome complex); PRK6, a compound predicted to interact with intrinsically disordered nuclear proteins (signature 5); and V0051, an FFF enriching several mitochondrial envelope proteins (signature 9).

The predictive capabilities of the ligand discovery resource go beyond the predefined set of signatures, extending to thousands of potential ML modeling tasks. To enable the assembly of bespoke ML models, we added an "on-the-fly" modeling asset that can build models based on user-defined list of proteins (or subsets thereof). For example, focusing on SLCs, we could predict the tendency of 34 newly screened compounds to interact with this transporter superfamily (Fig. 6G). Our ML models also assigned higher scores to known small-molecule SLC binders as compared with matched, randomly chosen control compounds (Fig. 6H). Collectively, this provides evidence that models generated on the basis of our chemoproteomics data empower the prediction of fragment behavior and engagement in a native cellular environment.

Discussion

Our experiments yielded a diverse catalog of fragment-protein interactomes, substantially expanding the ligandable proteome. Among proteins liganded through our chemoproteomics survey were more than 2000 proteins that previously lacked known ligands (*21*). Careful investigation of background labeling and removal of promiscuous proteins and fragments informed a pruned interactome of strong and selective fragment-protein interactions that revealed actionable starting points for ligand discovery, as exemplified by the successful development of compounds targeting SLC29A1, DDB1, and CDK2 with focused synthetic efforts.

The scale of our study has furthermore empowered an analysis of ligand binding that goes beyond individual proteins toward enriched, biologically coherent patterns of protein targets in the overall interactome of a fragment. These interactome signatures can, for instance, be reflective of binding to specific protein families or of partitioning of the fragment into specific cellular compartments. To evaluate the predictive power of the recorded fragment interactomes, we built ML models that can inform on principles of fragment behavior in native biological systems. We focused our efforts on (i) predicting fragment promiscuity, (ii) revealing higher-order functional signatures across fragment interactomes, and (iii) providing a flexible framework for modeling lower-order annotations, such as protein families or localization terms. Global features such as fragment promiscuity could be predicted and prospectively validated at high accuracy (AUROC >0.8). For the more nuanced interactome signatures, we found that certain signatures can be predicted with similar accuracy. We benchmarked this concept on the well-validated concept of lysosomal accumulation, but we can also extend it to other, nonobvious aspects of fragment behavior, including binding to protein families such as SLCs. Given the interest in drugs that partition into subcellular structures such as nuclear condensates, we anticipate that the presented models offer a rational framework for future ligand design efforts (*48*). Along similar lines, we foresee that integrating the presented proteome-wide binding data with RNA-binding properties will further illuminate our understanding of small-molecule behavior in complex biological systems (*49*, *50*). Additionally, we expect that our models will enable the rational design of bespoke chemical libraries to target specific protein families, such as SLCs and other transporters, thus further streamlining the development of novel medicines for this important but frequently neglected protein class (*33*).

Despite the depth of our survey, several limitations need to be considered. Even though we designed the more than 400 FFFs to be highly diverse, these efforts are insufficient to cover the chemical space in a saturating manner (*51*, *52*). Another limitation lies in the assayed proteomic space. Because all experiments were conducted in HEK293T cells, we lack data for proteins expressed in a tissue-restricted manner. The proteomics-based readout that we used has certain inherent technical limitations and biases, including detectability issues for low-abundance proteins. Moreover, in contrast to recent iterations, our workflow does not allow determination of the fragment binding site on the target protein (*53*). This can potentially complicate fragment elaboration and ML applications. Likewise, fragment optimization can be stymied by low stoichiometry of target engagement, which complicates the identification of competitive elaborated fragments. This might explain why certain signatures were more challenging to predict despite being functionally coherent (e.g., signature 7, tubulin binding). With additional ligandability maps and the advent of other proteome-wide strategies to assay ligand-protein interactions (*54*–*56*), we anticipate that increasingly fine-grained interactome signatures will be detectable and predictable and

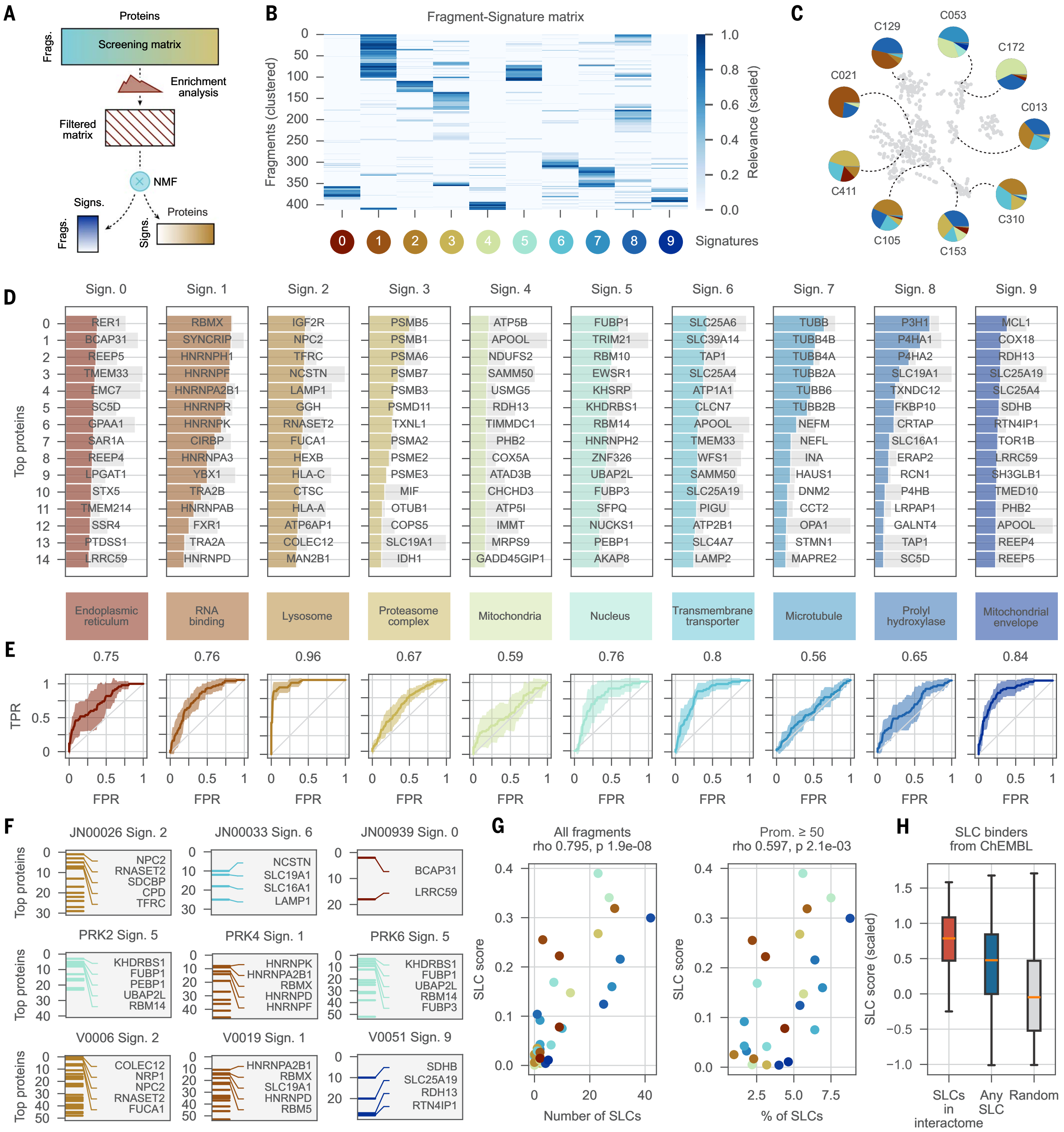


Fig. 6. Identification and predictability of protein-interactome signatures. (**A**) Scheme of the signature modeling pipeline. The full screening matrix is filtered by MF and CC protein set enrichment analysis to retain, for each FFF, only proteins that are at the leading edge of an enrichment signal, resulting in biologically coherent fragment interactomes. The filtered, sparse matrix is decomposed with NMF into a 407 × 10 FFF-signature matrix and a 10 × 970 signature-protein matrix. Frags., fragments; Signs., signatures. (**B**) View of the fragment-signature matrix, hierarchically clustered. Association between fragments and signatures is depicted with the blue color scale (relevance score). (**C**) UMAP projection of the FFF space in 2D, based on interactome profiles. Proximal FFFs (dots) have similar interactome profiles. Nine illustrative fragments are highlighted, with the pie chart representing their association to signatures. Colors match the legend shown in (B). (**D**) Top 15 relevant proteins per signature. The colored portion of the bar shows the association of each protein to the signature of interest, and the gray portion measures association to other signatures. Therefore, bars with small gray areas indicate specific associations. The colored boxes below each signature provide a high-level description of that signature (fig. S6A). (**E**) ROC curves for the ML model validations, performed 10 times in 80:20 train-test splits. The line is the mean curve, and the shade spans the standard deviation. (**F**) Prospective validation examples. The *y* axis ranks top

proteins for each fragment; the horizontal lines mark proteins associated to the signature of interest (top 30). (**G**) Global SLC prediction score (*y* axis) correlated to number (left) and proportion (right) of SLCs in the interactome of 34 newly tested FFFs. The SLC score corresponds to the weighted average of AUROC across SLCs (SM, materials and methods). Dots are colored by MW as in Fig. 1A. (**H**). SLC scores of small-molecule SLC binders from ChEMBL compared with >5000 random compounds from ChEMBL (MW, 50 to 500). The blue box corresponds to compounds annotated to any SLC (n > 3000); the red box corresponds to compounds annotated to SLCs in our interactome (n = 45).

hence further increase the efficiency of wet- and dry-lab cycles. An ultimate goal of these efforts could be the prediction of target proteins from an input chemical structure alone to accelerate small-molecule discovery in the currently unliganded proteome.

Methods summary

Full details of the methods are described in the supplementary materials and are summarized here. Chemoproteomics experiments were performed with a library of 407 chemically diverse chemical fragments from Enamine's catalog (harboring a diazirine moiety for photocrosslinking and an alkyne functionality for click reaction and enrichment) in HEK293T cells and analyzed by TMT 16plex-based MS (*16*, *57*). Fragment-protein interactions of interest were followed up with competitive fingerprinting experiments. Briefly, cells were pretreated with a 20-fold excess of elaborated structures followed by fragment treatment and subsequent processing through our regular chemoproteomics pipeline. Validation experiments included competitive chemical pulldown experiments with Western blot analysis, docking studies (*58*–*63*), production (*64*) and in vitro labeling of recombinant proteins with fragment probes, confocal microscopy to investigate subcellular localization of fragments, and functional cytotoxic uptake assays for SLC29A1 performed using cell viability as readout. In-gel fluorescence analysis of fragment-labeled proteome was done with TAMRA-azide to survey the promiscuity of fragments and was used as an orthogonal follow-up to promiscuity predictions (*16*).

Protein set enrichment analysis was achieved with Molecular Signatures Database (MSigDB) GO MF and CC annotation datasets, as well as PANTHER families and IDG categories, among others (*22*, *65*, *66*). Interactome signature discovery relied on using the Evolutionary Scale Modeling (ESM)–1b protein sequence embeddings and Bioteque CC embeddings (*67*, *68*). For interactome signature discovery, GO terms were used to retain functionally coherent proteins per fragments, followed by NMF for fragment-protein matrix decomposition. The ML methodology included pretraining of the FFF descriptor with a blend of topological and physicochemical descriptors, (*41*, *69*, *70*) binary classification with TabPFN models (*40*, *71*), and interpretability of promiscuity predictions using Shapley value analysis (*42*, *72*, *73*). A fully automated ML modeler is provided as part of the ligand discovery web resource.

REFERENCES AND NOTES

1. C. H. Arrowsmith *et al.*, The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536–541 (2015). doi: 10.1038/nchembio.1867; pmid: 26196764
2. S. L. Schreiber *et al.*, Advancing biological understanding and therapeutics discovery with small-molecule probes. *Cell* **161**, 1252–1265 (2015). doi: 10.1016/j.cell.2015.05.023; pmid: 26046436
3. C. J. Gerry, S. L. Schreiber, Chemical probes and drug leads from advances in synthetic planning and methodology. *Nat. Rev. Drug Discov.* **17**, 333–352 (2018). doi: 10.1038/nrd.2018.53; pmid: 29651105
4. S. L. Schreiber, The Rise of Molecular Glues. *Cell* **184**, 3–9 (2021). doi: 10.1016/j.cell.2020.12.020; pmid: 33417864
5. A. Mullard, 2022 FDA approvals. *Nat. Rev. Drug Discov.* **22**, 83–88 (2023). doi: 10.1038/d41573-023-00001-3; pmid: 36596858
6. T. I. Oprea *et al.*, Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018). doi: 10.1038/nrd.2018.14; pmid: 29472638
7. S. Müller *et al.*, Target 2035 - update on the quest for a probe for every protein. *RSC Med. Chem.* **13**, 13–21 (2021). doi: 10.1039/D1MD00228G; pmid: 35211674
8. F. Vincent *et al.*, Phenotypic drug discovery: Recent successes, lessons learned and new directions. *Nat. Rev. Drug Discov.* **21**, 899–914 (2022). doi: 10.1038/s41573-022-00472-w; pmid: 35637317
9. B. Lamoree, R. E. Hubbard, Current perspectives in fragment-based lead discovery (FBLD). *Essays Biochem.* **61**, 453–464 (2017). doi: 10.1042/EBC20170028; pmid: 29118093
10. D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, H. Jhoti, Twenty years on: The impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* **15**, 605–619 (2016). doi: 10.1038/nrd.2016.109; pmid: 27417849
11. K. M. Backus *et al.*, Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534**, 570–574 (2016). doi: 10.1038/nature18002; pmid: 27309814
12. S. M. Hacker *et al.*, Global profiling of lysine reactivity and ligandability in the human proteome. *Nat. Chem.* **9**, 1181–1190 (2017). doi: 10.1038/nchem.2826; pmid: 29168484
13. H. S. Hahm *et al.*, Global targeting of functional tyrosines using sulfur-triazole exchange chemistry. *Nat. Chem. Biol.* **16**, 150–159 (2020). doi: 10.1038/s41589-019-0404-5; pmid: 31768034
14. M. E. Abbasov *et al.*, A proteome-wide atlas of lysine-reactive chemistry. *Nat. Chem.* **13**, 1081–1092 (2021). doi: 10.1038/s41557-021-00765-4; pmid: 34504315
15. L. Boike, N. J. Henning, D. K. Nomura, Advances in covalent drug discovery. *Nat. Rev. Drug Discov.* **21**, 881–898 (2022). doi: 10.1038/s41573-022-00542-z; pmid: 36008483
16. C. G. Parker *et al.*, Ligand and Target Discovery by Fragment-Based Screening in Human Cells. *Cell* **168**, 527–541.e29 (2017). doi: 10.1016/j.cell.2016.12.029; pmid: 28111073
17. P. Kleiner, W. Heydenreuter, M. Stahl, V. S. Korotkov, S. A. Sieber, A Whole Proteome Inventory of Background Photocrosslinker Binding. *Angew. Chem. Int. Ed.* **56**, 1396–1401 (2017). doi: 10.1002/anie.201605993; pmid: 27981680
18. A. V. West *et al.*, Labeling Preferences of Diazirines with Protein Biomolecules. *J. Am. Chem. Soc.* **143**, 6691–6700 (2021). doi: 10.1021/jacs.1c02509; pmid: 33876925
19. J. Gao, A. Mfuh, Y. Amako, C. M. Woo, Small Molecule Interactome Mapping by Photoaffinity Labeling Reveals Binding Site Hotspots for the NSAIDs. *J. Am. Chem. Soc.* **140**, 4259–4268 (2018). doi: 10.1021/jacs.7b11639; pmid: 29543447
20. J. Li *et al.*, TMTpro reagents: A set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020). doi: 10.1038/s41592-020-0781-4; pmid: 32203386
21. T. K. Sheils *et al.*, TCRD and Pharos 2021: Mining the human proteome for disease biology. *Nucleic Acids Res.* **49** (D1), D1334–D1346 (2021). doi: 10.1093/nar/gkaa993; pmid: 33156327
22. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi: 10.1073/pnas.0506580102; pmid: 16199517
23. A. Hanzl, G. E. Winter, Targeted protein degradation: Current and future challenges. *Curr. Opin. Chem. Biol.* **56**, 35–41 (2020). doi: 10.1016/j.cbpa.2019.11.012; pmid: 31901786
24. J. W. Harper, B. A. Schulman, Cullin-RING Ubiquitin Ligase Regulatory Circuits: A Quarter Century Beyond the F-Box Hypothesis. *Annu. Rev. Biochem.* **90**, 403–429 (2021). doi: 10.1146/annurev-biochem-090120-013613; pmid: 33823649
25. T. Ishida, A. Ciulli, E3 Ligase Ligands for PROTACs: How They Were Found and How to Discover New Ones. *SLAS Discov.* **26**, 484–502 (2021). doi: 10.1177/2472555220965528; pmid: 33143537
26. M. Schapira, M. F. Calabrese, A. N. Bullock, C. M. Crews, Targeted protein degradation: Expanding the toolbox. *Nat. Rev. Drug Discov.* **18**, 949–963 (2019). doi: 10.1038/s41573-019-0047-y; pmid: 31666732
27. B. P. Belcher, C. C. Ward, D. K. Nomura, Ligandability of E3 Ligases for Targeted Protein Degradation Applications. *Biochemistry* **62**, 588–600 (2023). doi: 10.1021/acs.biochem.1c00464; pmid: 34473924
28. X. Zhang, V. M. Crowley, T. G. Wucherpfennig, M. M. Dix, B. F. Cravatt, Electrophilic PROTACs that degrade nuclear proteins by engaging DCAF16. *Nat. Chem. Biol.* **15**, 737–746 (2019). doi: 10.1038/s41589-019-0279-5; pmid: 31209349
29. X. Zhang *et al.*, DCAF11 Supports Targeted Protein Degradation by Electrophilic Proteolysis-Targeting Chimeras. *J. Am. Chem. Soc.* **143**, 5141–5149 (2021). doi: 10.1021/jacs.1c00990; pmid: 33783207
30. C. Mayor-Ruiz *et al.*, Rational discovery of molecular glue degraders via scalable chemical profiling. *Nat. Chem. Biol.* **16**, 1199–1207 (2020). doi: 10.1038/s41589-020-0594-x; pmid: 32747809
31. M. Słabicki *et al.*, The CDK inhibitor CR8 acts as a molecular glue degrader that depletes cyclin K. *Nature* **585**, 293–297 (2020). doi: 10.1038/s41586-020-2374-x; pmid: 32494016
32. M. Meyers, S. Cismoski, A. Panidapu, B. Chie-Leon, D. K. Nomura, Targeted Protein Degradation through Recruitment of the CUL4 Complex Adaptor Protein DDB1. *ACS Chem. Biol.* **19**, 58–68 (2024). doi: 10.1021/acschembio.3c00487; pmid: 38192078
33. A. César-Razquin *et al.*, A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478–487 (2015). doi: 10.1016/j.cell.2015.07.022; pmid: 26232220
34. W. W. Wang, L. Gallo, A. Jadhav, R. Hawkins, C. G. Parker, The Druggability of Solute Carriers. *J. Med. Chem.* **63**, 3834–3867 (2020). doi: 10.1021/acs.jmedchem.9b01237; pmid: 31774679
35. N. J. Wright, S. Y. Lee, Structures of human ENT1 in complex with adenosine reuptake inhibitors. *Nat. Struct. Mol. Biol.* **26**, 599–606 (2019). doi: 10.1038/s41594-019-0245-7; pmid: 31235912
36. R. C. Boswell-Casteel, F. A. Hays, Equilibrative nucleoside transporters—A review. *Nucleosides Nucleotides Nucleic Acids* **36**, 7–30 (2017). doi: 10.1080/15257770.2016.1210805; pmid: 27759477
37. B. Niemann *et al.*, Apoptotic brown adipocytes enhance energy expenditure via extracellular inosine. *Nature* **609**, 361–368 (2022). doi: 10.1038/s41586-022-05041-0; pmid: 35790189
38. R. Maréchal *et al.*, Human equilibrative nucleoside transporter 1 and human concentrative nucleoside transporter 3 predict survival after adjuvant gemcitabine therapy in resected pancreatic adenocarcinoma. *Clin. Cancer Res.* **15**, 2913–2919 (2009). doi: 10.1158/1078-0432.CCR-08-2080; pmid: 19318496
39. M. Sundaram *et al.*, Topology of a human equilibrative, nitrobenzylthioinosine (NBMPR)-sensitive nucleoside transporter (hENT1) implicated in the cellular uptake of adenosine and anti-cancer drugs. *J. Biol. Chem.* **276**, 45270–45275 (2001). doi: 10.1074/jbc.M107169200; pmid: 11584005

40. N. Hollmann, S. Müller, K. Eggensperger, F. Hutter, TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. arXiv:2207.01848v6 [cs.LG] (2022).
41. H. Moriwaki, Y. S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018). doi: 10.1186/s13321-018-0258-y; pmid: 29411163
42. S. M. Lundberg *et al.*, From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020). doi: 10.1038/s42256-019-0138-9; pmid: 32607472
43. J. N. S. Vargas, M. Hamasaki, T. Kawabata, R. J. Youle, T. Yoshimori, The mechanisms and roles of selective autophagy in mammals. *Nat. Rev. Mol. Cell Biol.* **24**, 167–185 (2023). doi: 10.1038/s41580-022-00542-2; pmid: 36302887
44. D. Takahashi, H. Arimoto, Selective autophagy as the basis of autophagy-based degraders. *Cell Chem. Biol.* **28**, 1061–1071 (2021). doi: 10.1016/j.chembiol.2021.05.006; pmid: 34087173
45. S. Niessen *et al.*, Proteome-wide Map of Targets of T790M-EGFR-Directed Covalent Inhibitors. *Cell Chem. Biol.* **24**, 1388–1400.e7 (2017). doi: 10.1016/j.chembiol.2017.08.017; pmid: 28965727
46. A. M. Kaufmann, J. P. Krise, Lysosomal sequestration of amine-containing drugs: Analysis and therapeutic implications. *J. Pharm. Sci.* **96**, 729–746 (2007). doi: 10.1002/jps.20792; pmid: 17117426
47. B. Zhitomirsky, Y. G. Assaraf, Lysosomes as mediators of drug resistance in cancer. *Drug Resist. Updat.* **24**, 23–33 (2016). doi: 10.1016/j.drup.2015.11.004; pmid: 26830313
48. I. A. Klein *et al.*, Partitioning of cancer therapeutics in nuclear condensates. *Science* **368**, 1386–1392 (2020). doi: 10.1126/science.aaz4427; pmid: 32554597
49. Y. Tong *et al.*, Programming inactive RNA-binding small molecules into bioactive degraders. *Nature* **618**, 169–179 (2023). doi: 10.1038/s41586-023-06091-8; pmid: 37225982
50. B. M. Suresh *et al.*, A general fragment-based approach to identify and optimize bioactive ligands targeting RNA. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 33197–33203 (2020). doi: 10.1073/pnas.2012217117; pmid: 33318191
51. G. Karageorgis, D. J. Foley, L. Laraia, H. Waldmann, Principle and design of pseudo-natural products. *Nat. Chem.* **12**, 227–235 (2020). doi: 10.1038/s41557-019-0411-x; pmid: 32015480
52. L. Laraia, H. Waldmann, Natural product inspired compound collections: Evolutionary principle, chemical synthesis, phenotypic screening, and target identification. *Drug Discov. Today. Technol.* **23**, 75–82 (2017). doi: 10.1016/j.ddtec.2017.03.003; pmid: 28647090
53. J. M. Wozniak *et al.*, Enhanced mapping of small-molecule binding sites in cells. *Nat. Chem. Biol.* (2024). doi: 10.1038/s41589-023-01514-z
54. I. Piazza *et al.*, A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172**, 358–372.e23 (2018). doi: 10.1016/j.cell.2017.12.006; pmid: 29307493
55. M. M. Savitski *et al.*, Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784 (2014). doi: 10.1126/science.1255784; pmid: 25278616
56. S. Klaeger *et al.*, The target landscape of clinical kinase drugs. *Science* **358**, eaan4368 (2017). doi: 10.1126/science.aan4368; pmid: 29191878
57. V. Dvorak *et al.*, Paralog-dependent isogenic cell assay cascade generates highly selective SLC16A3 inhibitors. *Cell Chem. Biol.* **30**, 953–964.e9 (2023). doi: 10.1016/j.chembiol.2023.06.029; pmid: 37516113
58. O. Trott, A. J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010). doi: 10.1002/jcc.21334; pmid: 19499576
59. G. M. Morris *et al.*, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009). doi: 10.1002/jcc.21256; pmid: 19399780
60. S. E. Dixon-Clarke *et al.*, Structure and inhibitor specificity of the PCTAIRE-family kinase CDK16. *Biochem. J.* **474**, 699–713 (2017). doi: 10.1042/BCJ20160941; pmid: 28057719
61. G. Kontopidis *et al.*, Differential binding of inhibitors to active and inactive CDK2 provides insights for drug design. *Chem. Biol.* **13**, 201–211 (2006). doi: 10.1016/j.chembiol.2005.11.011; pmid: 16492568
62. N. M. O'Boyle *et al.*, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011). doi: 10.1186/1758-2946-3-33; pmid: 21982300
63. R. Guerois, J. E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002). doi: 10.1016/S0022-2836(02)00442-4; pmid: 12079393
64. E. Turco *et al.*, Reconstitution defines the roles of p62, NBR1 and TAX1BP1 in ubiquitin condensate formation and autophagy initiation. *Nat. Commun.* **12**, 5212 (2021). doi: 10.1038/s41467-021-25572-w; pmid: 34471133
65. P. D. Thomas *et al.*, PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022). doi: 10.1002/pro.4218; pmid: 34717010
66. A. Lachmann, Z. Xie, A. Ma'ayan, blitzGSEA: Efficient computation of gene set enrichment analysis through gamma distribution approximation. *Bioinformatics* **38**, 2356–2357 (2022). doi: 10.1093/bioinformatics/btac076; pmid: 35143610
67. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021). doi: 10.1073/pnas.2016239118; pmid: 33876751
68. A. Fernández-Torras, M. Duran-Frigola, M. Bertoni, M. Locatelli, P. Aloy, Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat. Commun.* **13**, 5304 (2022). doi: 10.1038/s41467-022-33026-0; pmid: 36085310
69. Y. Rong *et al.*, Self-supervised graph transformer on large-scale molecular data. arXiv:2007.02835v2 [q-bio.BM] (2020).
70. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework" in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, 4 to 8 August 2019 (Association for Computing Machinery, 2019), pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.
71. J. T. Vogelstein *et al.*, Supervised dimensionality reduction for big data. *Nat. Commun.* **12**, 2872 (2021). doi: 10.1038/s41467-021-23102-2; pmid: 34001899
72. C. Wang, Q. Wu, M. Weimer, E. Zhu, FLAML: A Fast and Lightweight AutoML Library. arXiv:1911.04706v3 [cs.LG] (2019).
73. W. X. Shen *et al.*, Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* **3**, 334–343 (2021). doi: 10.1038/s42256-021-00301-6
74. M. Duran-Frigola, N. Marella, Ligand Discovery chemoproteomics data, machine learning assets and apps, Version v0.0.2, Zenodo (2024); https://doi.org/10.5281/zenodo.10838062.

ACKNOWLEDGMENTS

We acknowledge J. Shipps (Pfizer) and A. Ender (CeMM) for administrative leadership of this collaboration. Moreover, we are grateful to all members of the Winter, Kubicek, and Superti-Furga labs for helpful discussions and editorial contributions. We thank all members of the Molecular Discovery Platform at CeMM for critical support. We acknowledge the contributions of R. Mnatsakanyan in implementing the described workflow. We thank A. Bergthaler for sharing pLEX305_2xHA-P2A-Puro vector. Moreover, we thank the imaging facility at the Medical University of Vienna for assistance with high-content microscopy studies. We thank C. Parker for helpful discussions around competitive pulldown assays as well as in-gel fluorescence scanning experiments. We thank B. Cravatt for helpful advice on fragment elaboration and interpretation of lysosomal partitioning of small molecules. We are grateful for advice from members of the Resolute consortium (www.re-solute.eu) on targeting of SLCs. **Funding:** We acknowledge support from the Vienna Science and Technology Fund (WWTF), grant LS21-015, and the Austrian Science Fund (FWF), grant FWF SFB F79, for work on autophagy and autophagy receptors (Winter and Martens lab). **Author contributions:** F.O. performed chemoproteomics experiments and led validation studies; G.T. performed chemoproteomics experiments and performed and supervised synthetic chemistry efforts; M.D-F. analyzed data, led all ML efforts, and wrote the manuscript; E.H. performed chemoproteomics experiments; S.D. performed chemoproteomics experiments and validation assays; C.W.A.E. and J.W.S. supervised and contributed to elaborate library design; A.R. operated the mass spectrometer, performed postdigestion sample prep, and analyzed data; V.B. analyzed and managed data; K.O. performed chemical syntheses; N.M. managed data and built software infrastructure; K.K. performed experiments; R.C. analyzed data and contributed to writing of the manuscript; J.D.M. and S.D.F. contributed to experimental design; A.B. performed experiments; L.F. produced recombinant autophagy receptors; E.F. performed docking studies; A.N. analyzed colocalization experiments; Z.Z. contributed to shape-data analyses and database infrastructure; G.D. built database infrastructure; A.B. performed docking studies; S.M. supervised recombinant protein work with autophagy receptors; R.S. contributed to shape-data analyses and database infrastructure; A.M. and J.T.H. supervised mass spectrometry analysis; D.H. and G.S.F. supervised and conceptualized the project; S.K. supervised the study and conceptualized the project; M.S. supervised and conceptualized the project; and G.E.W. supervised and conceptualized the project; wrote the manuscript, and has overall responsibility for the conducted research. **Competing interests:** M.D-F. is a founder and lead scientist at the Ersilia Open Source Initiative. S.K., G.S-F., and G.E.W. are scientific founders and shareholders of Proxygen and Solgate. G.E.W. is on the scientific advisory board of Nexo Therapeutics. A.B. is a founder and shareholder of Solgate. S.M. is a member of the scientific advisory board of Casma Therapeutics. C.W.A.E., J.W.S., K.O., J.D.M., S.D.F., Z.Z., R.S., D.H., and M.S. are employees and shareholders of Pfizer. The presented research has been funded by Pfizer. **Data and materials availability:** Proteomics measurements related to this study have been deposited in PRIDE (PXD041587). All subsequent data, code, and ML models are publicly available at https://github.com/ligand-discovery, archived at Zenodo (*74*). Executable notebooks for data analysis and paper figures can be found in the "primary-paper-figures" and "interactome-signatures" repositories. ML model training data and scripts for pretraining, binary classification, and interpretability are available in the "fragment-embedding," "mini-automl," and "mini-xai" repositories, respectively. Code to deploy interactive applications can be found at "multi-apps," including "protein-profile-explorer," "protein-set-enrichment-analysis," "fragment-predictor," and "on-the-fly-modeling." Online versions of these apps are accessible at https://ligand-discovery.ai. **License information:**

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adk5864
Materials and Methods
Figs. S1 to S8
Tables S1 to S7
MDAR Reproducibility Checklist

Submitted 31 August 2023; accepted 22 March 2024
10.1126/science.adk5864