nature reviews chemistry

https://doi.org/10.1038/s41570-025-00737-z

Roadmap

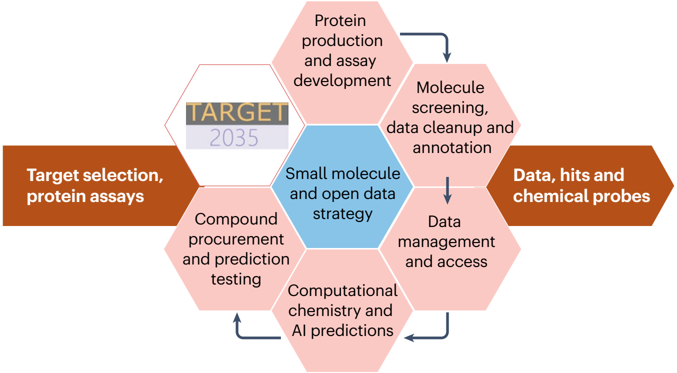
Check for updates

Protein–ligand data at scale to support machine learning

Aled M. Edwards [1], Dafydd R. Owen [2] & The Structural Genomics Consortium Target 2035 Working Group*

Abstract

Target 2035 is a global initiative that aims to develop a potent and selective pharmacological modulator, such as a chemical probe, for every human protein by 2035. Here, we describe the Target 2035 roadmap to develop computational methods to improve small-molecule hit discovery, which is a key bottleneck in the discovery of chemical probes. Large, publicly available datasets of high-quality protein–small-molecule binding data will be created using affinity-selection mass spectrometry and DNA-encoded chemical library screening. Positive and negative data will be made openly available, and the machine learning community will be challenged to use these data to build models and predict new, diverse small-molecule binders. Iterative cycles of prediction and testing will lead to improved models and more successful predictions. By 2030, Target 2035 will have identified experimentally verified hits for thousands of human proteins and advanced the development of open-access algorithms capable of predicting hits for proteins for which there are not yet any experimental data.


Protein production and assay development
Molecule screening, data cleanup and annotation
TARGET 2035
Target selection, protein assays
Small molecule and open data strategy
Data, hits and chemical probes
Compound procurement and prediction testing
Data management and access
Computational chemistry and AI predictions

Sections

[1] Structural Genomics Consortium, University of Toronto and University Health Network, Toronto, Ontario, Canada. [2] Pfizer Research and Development, Cambridge, MA, USA. *A list of authors and their affiliations appears at the end of the paper. e-mail: aled.edwards@utoronto.ca; dafydd.owen@pfizer.com

Introduction

Chemical probes – potent, selective, cell-active small molecules targeting specific proteins – constitute some of the most impactful research tools in the life sciences arsenal, as evidenced by citations and impact on drug discovery[1,2]. The broader availability of chemical probes for all human proteins would greatly advance our understanding of the human proteome, as well as help prioritize potential new drug targets. In 2009, the Structural Genomics Consortium (SGC) launched a programme to assemble and invent chemical probes for human proteins related to cell signalling, protein homeostasis and epigenetics. The programme successfully developed and collected new chemical probes for over 200 unique proteins from the academic and industrial communities. The impact of these 200 chemical probes has been profound: more than 60,000 samples have been distributed to scientists around the world, they have collectively been cited at least 13,000 times as assessed by searching for the name of the probe in Google Scholar, and the discoveries they have enabled are being tested in more than 85 clinical trials.

The obligatory first step in creating a chemical probe for a new protein (or a proximity pharmacology tool such as proteolysis-targeting chimeras (PROTACs))[3] is to identify a validated, chemically tractable hit. For proteins that belong to precedented classes of drug targets, hits can often be identified quite readily, either by screening focused chemical libraries that are enriched in experimentally verified structural classes[4] or by making computational predictions based on pre-existing experimental data[5–10]. By contrast, for lesser-studied proteins, hit-finding is more challenging and is often rate determining. Currently, hit finding is almost always initiated with an experimental screen of large and diverse chemical libraries followed by time- and cost-intensive hit verification and optimization. Although the available experimental hit-finding approaches have expanded greatly over the past 20 years, there has not been a dramatic improvement in their overall success rates or cost effectiveness[11–13]. This situation underscores the need for a radically different approach in the context of the Target 2035 initiative[14].

Computational methods, particularly machine learning (ML) and artificial intelligence (AI) strategies have the most potential to develop cost-effective hit-finding methods for unprecedented targets[15]. However, the development of hit-finding algorithms is currently limited by the lack of suitable protein–ligand datasets in the public domain[16,17]: the existing chemical bioactivity datasets are either fragmented across databases such as ChEMBL and PubChem, or are not available to the public[18], most have been compiled from non-standardized experimental protocols that introduce noise into training data[19], the datasets are not always prepared for ML/AI analysis, and most lack high-quality data on inactive compounds[20].

With data paucity identified as the greatest hurdle to the development of hit-finding algorithms, our SGC/Target 2035 working group decided that the next phase of the Target 2035 initiative (2025–2030) should organize a programme that (1) systematically generates large experimental protein–small–molecule binding datasets and provides open access to the well-annotated data, and (2) works with the community to train, develop, refine, test and benchmark hit-finding algorithms, to start.

A scientific and operational plan for the initiative, including target selection, data generation and dissemination, benchmarking of ML/AI predictions, success criteria, governance, and funding was discussed in a face-to-face meeting in Frankfurt, Germany, in the autumn of 2023, and in London, UK, in the autumn of 2024. In this Roadmap, we consolidate the outputs from these meetings into an ambitious yet tractable roadmap to provide sufficient experimentally derived data to transform hit finding into a computational endeavour. We also highlight how the Target 2035 open science initiative is structured to provide ample mechanisms for the greater experimental and computational academic and industry communities to contribute and benefit.

In our approach, there are conceptual parallels between our proposed approach and the development of the AlphaFold programs for protein structure prediction. The successful application of ML to protein structure prediction was empowered by massive open data generation by the structural biology and genomics communities, longstanding stewardship of the data by the Protein Data Bank and GenBank teams[21] and an engaged structure prediction community, whose algorithms were benchmarked by the CASP team (Critical Assessment of Protein Structure Prediction)[22] through open challenge competitions. This analogy has limits though. The immense space of intramolecular interactions that is afforded by 20 amino acids and defines the protein-folding paradigm is relatively constrained when compared with the diversity of possible interactions between proteins and ~10^{60} drug-like molecules. Clearly, novel ML strategies will be necessary for a breakthrough in AI-driven drug design, and it is not possible a priori to predict the size and diversity of the protein–ligand datasets that will be required to enable such a breakthrough, or even if it will be possible in the near term. With this caveat, it is nevertheless apparent that high-quality, large-size protein–ligand datasets will be foundational to solving the problem.

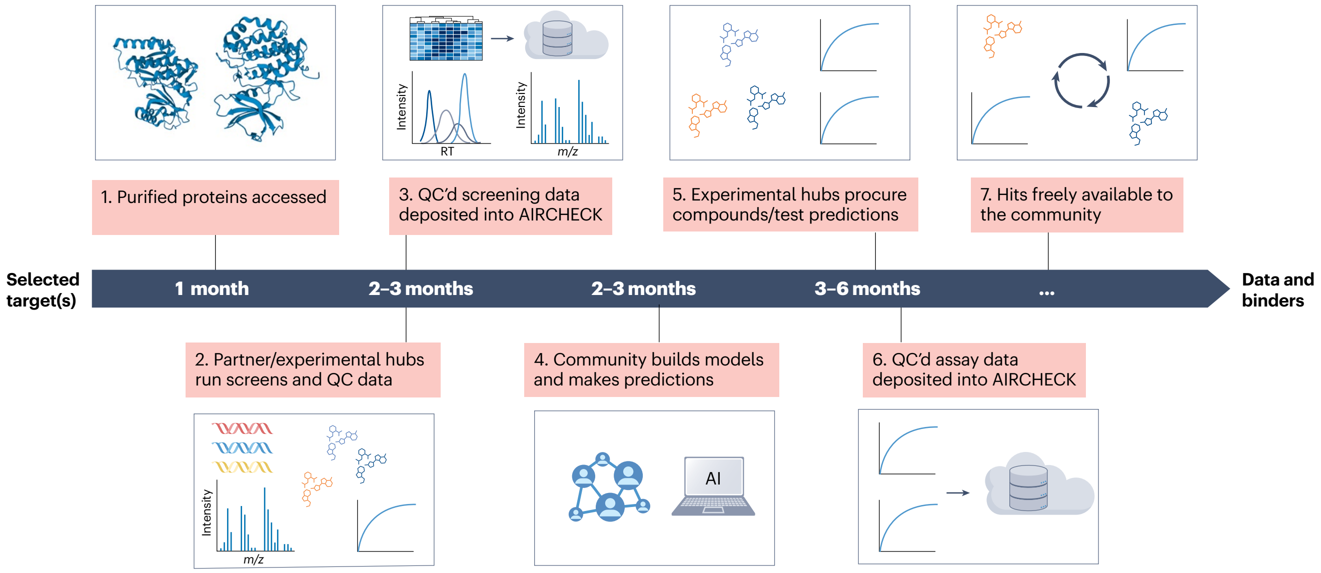
Overview of project workflow

This 5-year project will generate high-quality, open datasets that include binding data for millions to billions of small molecules to more than 2,000 diverse proteins. The data will include results from testing both experimentally derived and computationally predicted hit candidates using orthogonal biophysical and functional assays.

The project workflow is outlined in (Fig. 1) and is described in more detail below. In brief, the project will

(1) Generate purified proteins both within the project and by inviting community members to contribute purified proteins. All purified proteins would be subject to strict quality control.
(2) Generate binding data using affinity-selection mass spectrometry (AS–MS) and DNA-encoded chemical library (DEL) screening, each of which measures the binding of small molecules to purified proteins directly. AS–MS and DEL screening are also performed in a standardized way, and the outputs have associated quality metrics. Candidate small-molecule binders will be tested in secondary screens using orthogonal, high-quality biophysical assays.
(3) Make annotated primary screening data openly available in an ML/AI-ready format via a project database called AIRCHECK (Artificial Intelligence-Ready CHEmiCal Knowledge base; https://aircheck.ai/).
(4) Challenge the ML/AI and computational chemistry communities to make predictions based on the data and organize a series of benchmarking competitions to help advance the methods.
(5) Experimentally test community predictions using biophysical methods.
(6) Share assay data from predicted binders via AIRCHECK.
(7) Share reagents, protocols, binders and data without restrictions on use.

In addition, for as many confirmed binders as feasible, co-crystallization with the cognate target would be attempted, and



Intensity
RT
Intensity
m/z
1. Purified proteins accessed
3. QC'd screening data deposited into AIRCHECK
5. Experimental hubs procure compounds/test predictions
7. Hits freely available to the community
Selected target(s)
1 month
2–3 months
2–3 months
3–6 months
...
Data and binders
2. Partner/experimental hubs run screens and QC data
4. Community builds models and makes predictions
6. QC'd assay data deposited into AIRCHECK
Intensity
m/z
AI

Fig. 1 | Data generation pipeline. The workflow for generating data and binders. (1) Purified proteins are produced in experimental hubs, by partners or by the community. (2) Proteins are screened in project screening laboratories, and data are experimentally annotated in partner laboratories or experimental hubs. (3) Quality-controlled (QC) screening data are deposited into the AIRCHECK database. (4) Computational experts in the project and the community build machine learning (ML)/artificial intelligence (AI) models and make predictions about new or improved binders. (5) Predicted compounds are procured and tested in experimental hubs. (6) The QC'd assay data, including hits and binding data, are deposited into the Artificial Intelligence-Ready CHEmiCal Knowledge base (AIRCHECK). (7) Hits and data are released to the community, freely available for further research and development.

structure–activity relationships explored by testing structural analogues of the confirmed binders, either purchased from vendors or synthesized by collaborating chemists. Ideally, all binders would be tested in functional assays when available.

The project will have two important outcomes. First, it will generate new small-molecule binders for prioritized proteins. Second, it will create a comprehensive, well-annotated dataset to advance computational methods. This second outcome will be achieved by prioritizing data quality, data consistency and data access, and designing the experimental workflow and outputs in partnership with data scientists[23].

Access to diverse high-quality proteins

To generate protein–small-molecule binding datasets of sufficient size and diversity, it will be critical to access and prosecute a structurally diverse set of purified, homogeneous and stable proteins. Given that it is not possible a priori to estimate the number of datasets that will be required to enable computational methods, for planning purposes, we have arbitrarily set a goal to screen a minimum of 2,000 different proteins over 5 years. For perspective, and to attest to the feasibility of the project, this is approximately the number of unique proteins purified within the SGC in the 5-year period between 2007 and 2012.

The selection of which proteins to screen will be guided by the requirement to maximize the structural and functional diversity of the protein targets, as well as by the desire of funders and participants to identify hits for protein targets of their immediate scientific interest. Initially, experimental tractability will be prioritized to establish and optimize project platforms, logistics, data workflows and procedures. Tractable targets constitute those that can be readily purified in sufficient quantities, are known to have suitable biophysical properties, and for which orthogonal assays are either already available or can be readily developed (Fig. 2). The SGC and the wider protein and structural biology communities have successfully produced over a thousand human proteins (or domains thereof) in the past that meet these criteria, and these proteins should be easily and rapidly accessible or able to be repurified. A graphical representation of ~400 proteins already purified at the SGC is included in Supplementary Fig. 1 and a snapshot of the protein database in Supplementary Table 1. As the project progresses, the number of never-before purified proteins and proteins that are more technically challenging to produce will be increased.

Protein production

To ensure protein quality and consistency, protein quality criteria have been established and implemented (an exemplar is shown in Supplementary Fig. 2) and the majority of the proteins will be produced in a handful of geographically distributed protein purification hubs that share methodologies. These hubs will probably be organized around protein families and/or scientific themes. To attract a wider diversity of protein targets, experts in the community would be invited to contribute purified proteins that meet the diversity and quality criteria. The incentives for community members to contribute proteins will be to access high-quality chemical screening capabilities and to be able to pursue any small-molecule hits identified in the screens, without precondition (https://public.thesgc.org/protein_registry/protein_intake.php).

Protein–ligand open data generation

All purified proteins that pass quality control will be screened for binders within large chemical libraries. Screens will be carried out in academic or industry hubs, selected for having track records in high-quality data generation. The distribution of proteins to and among

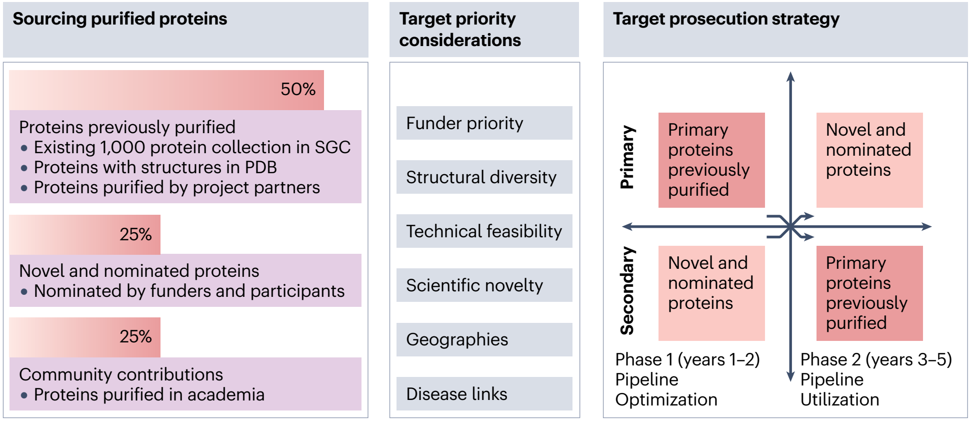

Sourcing purified proteins
50%
Proteins previously purified
• Existing 1,000 protein collection in SGC
• Proteins with structures in PDB
• Proteins purified by project partners
25%
Novel and nominated proteins
• Nominated by funders and participants
25%
Community contributions
• Proteins purified in academia
Target priority considerations
Funder priority
Structural diversity
Technical feasibility
Scientific novelty
Geographies
Disease links
Target prosecution strategy
Primary
Secondary
Primary proteins previously purified
Novel and nominated proteins
Novel and nominated proteins
Primary proteins previously purified
Phase 1 (years 1–2)
Pipeline Optimization
Phase 2 (years 3–5)
Pipeline Utilization

Fig. 2 | Target selection and prioritization. The protein pipeline will comprise proteins that have been produced previously by project participants, new proteins that are nominated by the funders, and proteins contributed by experts in the wider community. Among the proteins, the project will create a Target List that integrates structural and ligand-binding pocket diversity, funder interests and scientific priorities. Proteins produced previously will be prioritized at the outset to focus on logistics and to generate data, and never previously produced proteins will be added as the project progresses. PDB, Protein Data Bank; SGC, Structural Genomics Consortium.

the screening hubs will be centrally coordinated to avoid duplication of effort.

A key strategic decision was to select the data-generating modalities. Platform(s) that screen for direct binding of ligands to purified proteins would be implemented, for the following reasons:

(1) Direct-binding assays eliminate the impractical requirement to develop bespoke functional assays for each protein, including the thousands of human proteins with no known activity. For proteins with known function, a functional assay might aid in the hit verification process, and in the further advancement of the chemical matter.

(2) A single preparation of purified protein can be used both for the primary binding screen used for hit identification as well as for the secondary orthogonal biophysical assays[24] used for hit verification.

(3) Screening campaigns could begin immediately, using the many hundreds of human proteins that have been purified or can be readily purified in high quality and quantity, by the SGC, by industry, and by the wider protein and structural biology academic communities.

After considering many screening platforms, DEL[25–28] and AS–MS[29–31] were chosen. These two biophysical screening methods have been used successfully for a wide variety of proteins, have the potential to generate millions of high-quality data points per screen and have already demonstrated efficient hit-finding results in our hands for diverse proteins. In addition, data generated by these methods have a common experimental design and can be represented in a machine-readable format and aggregated into increasingly large datasets. The large size and high dimensionality of these data also leverage respective analytical techniques that have been extensively developed by the ML/AI community[32,33] and employed in cheminformatics for drug discovery applications[34].

DEL screening

DEL screening is an affinity-mediated technique that has been used for more than two decades as a tool for identifying compounds that bind proteins[35–38]. In this technology, pools of compounds, each covalently attached to an oligonucleotide whose sequence encodes its synthetic history (and therefore the presumed compound identity), are incubated with the protein. Proteins are then captured using an affinity tag and associated library members are separated from non-binders by washing. The DNA encoding the retained library members is then amplified and sequenced, allowing the synthetic history of each compound and their enrichment over the background to be determined. Historically, enriched library compounds were resynthesized off the DNA and tested for binding or activity in an orthogonal assay. The technology allows for probing an enormous chemical library (>1 trillion members), but it has limitations: the presence of DNA induces many false positives, synthesizing the many potential binders off-DNA is time consuming and costly, and the chemical diversity of the library members is restricted by the requirement to use reactions that are compatible with the presence of DNA.

Some of these limitations can be overcome by integrating ML/AI with DEL screening. In this iteration of DEL screening data analysis, the datasets, comprising billions of data points and including both positive and (critically) negative binding data, are used to train ML algorithms and build models to predict the molecular features of a binder[26,28,39]. These algorithms are then used to search for active molecules within the billions of commercially available compounds or compound collections internal to organizations. The compounds are then acquired and tested for binding to the purified proteins using orthogonal binding and/or functional assays. This strategy offers several potential advantages. First, it is faster and less expensive for most investigators to purchase molecules[40] than to synthesize each of the enriched library compounds. Second, predictions are not restricted to the molecules in the DEL but can be made against the large, diverse, and more drug-like chemical space represented in pre-enumerated, synthetically accessible commercial libraries.

This conceptual DEL ML workflow was pioneered by McCloskey et al.[26] using three precedented targets and the scalability and generalizability of the approach have been subsequently confirmed[27,28,41,42] These encouraging results emboldened us to imagine a scaled-up process in which DEL screening datasets from hundreds to thousands of proteins, including detailed protocols and metadata, would be provided to the academic and industry communities without restriction in a standardized, ML-ready format[43,44]. By providing open access to these data (aircheck.ai), the ML/AI community would be enabled to make predictions that can be tested experimentally and to develop methods that can be benchmarked (Fig. 3). In the first datasets in AIRCHECK, the data include a 10:1 ratio of negative to positive training examples, and up to 1 million data points. Negative training examples were proportionally distributed to positive training examples on a

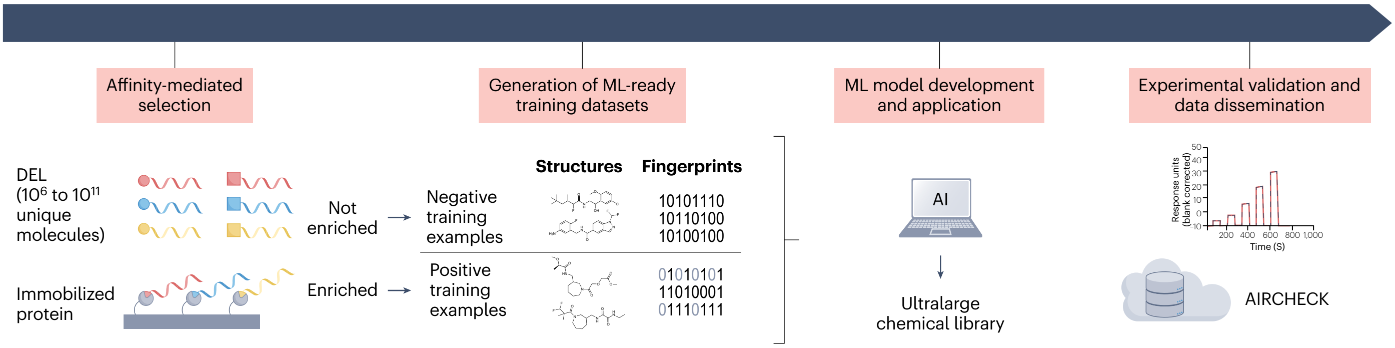

Affinity-mediated selection
Generation of ML-ready training datasets
ML model development and application
Experimental validation and data dissemination
DEL
(10^6 to 10^11 unique molecules)
Immobilized protein
Not enriched
Enriched
Structures
Fingerprints
Negative training examples
10101110
10110100
10100100
Positive training examples
01010101
11010001
01110111
AI
Ultralarge chemical library
Response units (blank corrected)
50
40
30
20
10
0
-10
200
400
600
800
1,000
Time (S)
AIRCHECK

Fig. 3 | Schematic representation of DEL screen output data and ML/AI workflow. Affinity-mediated selection of DNA-encoded chemical library (DEL) members leads to the enrichment of potential binders. Deep sequencing is used to identify the DNA barcodes of enriched and unenriched DEL members. Output data are subsequently translated into chemical structures and their corresponding chemical fingerprints. Both positive (enriched) and negative (not enriched) DEL members are included in open machine learning (ML)-ready datasets. The datasets are used to train ML models that are in turn used to recognize and nominate potential small-molecule binders from ultralarge chemical libraries. These compounds are procured, and their binding is tested experimentally in biophysical and/or biochemical assays. All generated data (ML-ready datasets, including chemical structures and/or their corresponding fingerprints, ML models and ligand validation data) are made public in a purpose-built, cloud-based storage system called Artificial Intelligence-Ready CHEmiCal Knowledge base (AIRCHECK). AI, artificial intelligence.

per-library basis. These data have already been used to build models that have successfully predicted new micromolar binders for the WDR91 protein[45].

Initially, the DEL screens will be carried out in selected organizations that have a track record of success in applying ML to their DEL data[43,44]. Over time, any other company or academic that has robust DEL synthesis and screening infrastructure and that agrees to share relevant data openly and in a standardized, ML-ready format[43,44] would be welcome to join the initiative.

AS–MS

AS–MS has emerged as a robust hit identification approach in the pharmaceutical industry[46]. In this method, pools of mass-differentiated compounds, typically up to 2,000, are first incubated with the protein. The protein and small molecules are then resolved chromatographically, and compounds that co-elute with the protein are subjected to liquid chromatography–mass spectrometry and unambiguously identified by their exact masses. Compound binding is then verified using an orthogonal functional or binding assay(s). The current upper limit of detection for compounds in most AS–MS platforms is an affinity constant in the 1–15 micromolar range[46].

With some notable exceptions[47–50], AS–MS has not been widely adopted as a small-molecule screening platform in academia, in part due to the significant infrastructure that is required, but mostly because cost-effective use of the infrastructure requires a pipeline of purified proteins in multi-milligram quantities. Given the ability to access thousands of purified proteins in these quantities in this project, AS–MS was prioritized as a screening platform (Fig. 4). To optimize screening capacity and throughput, we have elected to implement an off-line AS–MS method that screens affinity-tagged proteins (his, GFP or biotin) against pools of compounds, and then resolves the protein/compound complexes from the non-binding compounds by binding the tagged protein to the corresponding magnetic affinity microbeads[50]. This pipeline was piloted by screening

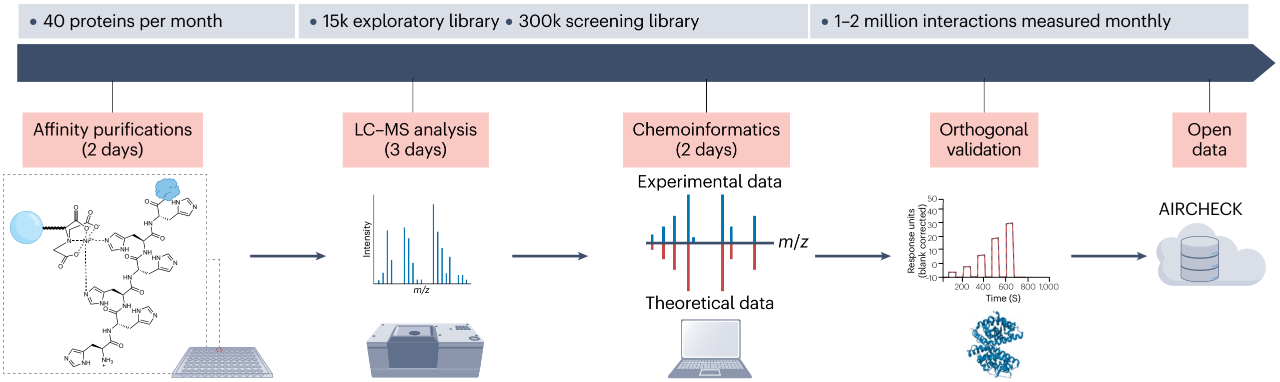

• 40 proteins per month
• 15k exploratory library • 300k screening library
• 1–2 million interactions measured monthly
Affinity purifications (2 days)
LC–MS analysis (3 days)
Chemoinformatics (2 days)
Orthogonal validation
Open data
Intensity
m/z
Experimental data
m/z
Theoretical data
Response units (blank corrected)
50
40
30
20
10
0
-10
200
400
600
800
1,000
Time (S)
AIRCHECK

Fig. 4 | The AS–MS screening workflow. From left to right: protein affinity purification by pooling 500 compounds in each affinity-selection mass spectrometry (AS–MS) sample; liquid chromatography–mass spectrometry (LC–MS) analysis of AS–MS samples; automatic data processing to identify hits; validation of hits using orthogonal biophysical methods (for example, surface plasmon resonance); and uploading AS–MS data to the Artificial Intelligence-Ready CHEmiCal Knowledge base (AIRCHECK) database, which is freely accessible to the whole community.

Table 1 | Data management features

Attribute	Description
The AIRCHECK database	Houses Target 2035 screening datasets Supports machine learning (ML)/artificial intelligence (AI) model development, evaluation and reusability Follows FAIR principles (findable, accessible, interoperable, reusable) Publishes and documents data and computer code for data processing, quality control and normalization Ensures transparency and allows users to scrutinize data transformation and ML/AI models
Standardizing experimental data using controlled vocabulary	Links experimental protocols to assay data via electronic lab notebooks and laboratory information management systems Uses commercial tools to allow uptake by the community Shares database architecture and controlled vocabulary across labs Facilitates integration of data (e.g., protein production, screening hit validation)
Robust versioning	Automatically tracks and documents dataset changes Uses data nutrition labels to visualize and summarize dataset characteristics and updates Transforms datasets for integration into repositories, such as ChEMBL and PubChem
Reusability	Provides comprehensive documentation, including experimental protocols and lab notebooks Offers analysis code and output files from tutorials and workshops, and fully specified ML/AI models Creates educational materials for users Enables users to understand the data and previous analyses
Data release	Releases generated and quality-controlled data immediately or at regular intervals (e.g., quarterly) Aligns data releases with open benchmarking challenges to encourage use and re-use Releases data in the context of chemical probe collaborations for added scientific value
Integrating diverse data	Supports ingestion of data from diverse screening platforms (affinity-selection mass spectrometry, DNA-encoded chemical library ML) Creates multimodal data objects integrating data for a single target from various platforms Tracks processing pipelines and ensures full traceability of data generation (inspired by the ORCESTRA platform for genomics data[70])
Equity and inclusion	Ensures data and computational resources are free to access Cloud implementation allows users with limited resources to run ML/AI methods using free research credits Partners with cloud providers to facilitate resource use for users from low-income countries[65] Develop the Artificial Intelligence-Ready CHEmiCal Knowledge base (AIRCHECK) web application following the Web Accessibility Initiative to maximize inclusion and diversity[71]
Data science	Trains ML models using rigorously processed and curated data Represents data in formats optimized for downstream applications Uses random, chronological or other splitting mechanism to divide data into training, validation and test sets Continuously tests and updates models with new data Enhances predictive accuracy and monitors 'model drift' over time Uses active learning to drive design–make–test–analyse cycles Evaluates prediction uncertainty to inform decision-making and reinforce model reliability

a diverse set of 31 proteins against a small chemical library explicitly optimized for mass spectrometry screening, and binders were discovered for 11 proteins[51].

The primary binding data and metadata from both DEL and AS–MS screens, as well as the results from secondary biophysical assays are now being placed into AIRCHECK without restriction on use. Raw mass spectrometry data will also be made available via Metabolomics Workbench (https://www.metabolomicsworkbench.org/) or a similar vehicle.

Annotation and verification of screening data

With a priority to generate screening datasets for ML/AI applications, particular attention will be paid to data quality, data annotation and data availability – using learnings from the experiences of our industry partners and other public initiatives[52]. Data quality standards will be made openly available and implemented at three key levels: for the protein samples, for the DEL and AS–MS screening outputs, and for the hit annotation.

Proteins

Proteins entering screens must meet established experimental quality criteria and must also be accompanied by key metadata that might influence data interpretation and model building, such as purification conditions or the presence of metal ions.

Screening datasets

Primary AS–MS and DEL screening-derived datasets will be assessed for technical quality against a set of relevant parameters (Supplementary Fig. 3). For public DEL and AS–MS screens that pass quality checks, all the raw screening data will be placed into the public domain.

Secondary annotation of primary screening data

Both experimental screening platforms will generate false positive and false negative hits, and to maintain the quality of the datasets for ML/AI applications, true and false positives must be distinguished using orthogonal assays[53]. This is technically challenging because weaker binding compounds are often insoluble at the concentrations used for many biophysical or functional assays[54,55], which readily leads to artefacts in any single assay. As a result, many candidate binders may have to be tested in several different assays to gain sufficient confidence in their veracity.

Given the technical challenges in analysing weakly binding compounds, it will be critical to agree on how much effort the project should invest in determining if a screening hit is a true binder and to communicate the limitations of each of the assays[54,55] and the resulting data to the modelling community. The strategic decision is how to balance annotating the largest number of true positives in the dataset, which is optimal for model building and also provides practical and valuable insight into the ligandability of a protein, with investing considerable resources in characterizing weakly binding compounds, which reduces the number of proteins that can be screened. The CACHE competition has created a document that explains how to interpret the biophysical binding assays and how to identify potential artefacts[56]. Constant and close discussion between experimentalists and data scientists in the project will minimize misinterpretation or over-interpretation of the screening and hit-characterization data.

For this project, a generous threshold in nominating hits from the initial screen would be implemented. A target affinity threshold of 10 μM (K_D value) would be set for the orthogonal assay, potentially with some target-specific leeway[20]. Ideally, all candidate hits arising from the first orthogonal assay would be tested in an additional assay. The outcome will be a robust and inclusive list of well-annotated positive binders with a K_D of ≤10 μM.

Data consistency. To prioritize data consistency, secondary screening and data annotation will be centralized in well-equipped and experienced academic or commercial laboratories that follow standard operating procedures. Samples will be exchanged regularly among laboratories and tested to monitor and eliminate any inter-laboratory variability. These laboratories will have access to a range of orthogonal assay formats including some form of surface binding assay such as surface plasmon resonance[57] or grating-coupled interferometry[58], and other biophysical methods with reasonable throughput, such as spectral shift, microscale thermophoresis, NMR or thermal shift methodologies[59–61]. One of the complexities of the project is that for many of the novel proteins that will be screened, orthogonal assays will have to be built without the benefit of a positive control binder. If functional assays that confirm target modulation are readily available, they would add another layer of verification to the hit-confirmation process and provide invaluable insight into how to develop the ligand into a chemical probe.

Data management and access

To fully realize the value of the annotated protein–ligand datasets, data management approaches will be treated with equal diligence as the experimental methods. Accordingly, the project will adhere to the data management roadmap recently described by Edfeldt and colleagues[23]. This will include establishing a controlled vocabulary for experimental data, using automation and electronic laboratory notebooks whenever possible, centralizing the database architecture to facilitate data integration and providing comprehensive documentation. Raw data will be provided whenever possible and transparent and reproducible data processing will be performed, including choosing the most relevant data representation, defining the right training and test sets, and providing estimates of prediction uncertainty. The comprehensive data management plan and its attributes are outlined in Table 1.

Benchmarking with experimental feedback

The intention of providing large, consistent and high-quality datasets to the community is to enable the development of computational and ML/AI hit-finding and hit optimization methods. The models will be focused in the near term on predicting binders and optimization strategies for proteins in the screening set and in the longer term to build foundation models of hit discovery and optimization.

To accelerate the development of these methods, the project will partner with organizations, including CASP, DREAM[62] and CACHE[15], that launch benchmarking challenges, including those in which predictions from the community will be tested experimentally and compared. Data used as input to challenges would be kept confidential while challenges are in progress, and a regular cadence of challenges and data

Table 2 | Sample benchmarking challenges

Data	Challenge	Experimental validation
SMILES and/or fingerprint and enrichment metrics of DNA-encoded chemical library (DEL) screening hits and negatives from 4-10B compound library	Train machine learning (ML)/artificial intelligence (AI) models on DEL screening data and use them to predict actives from billions of commercial compounds	Procure and test predicted hits with two orthogonal assays
300k affinity-selection mass spectrometry (AS–MS) compound library (SMILES) and protein target	Predict true and false positives	Compare predictions with screening results, annotated with orthogonal assays
AS–MS screening and orthogonal hit confirmation data for 80% of a 300k compound library	Challenge 1: predict confirmed hits for the remaining 20% hold-out set Challenge 2: if successful, predict novel hits from commercial libraries	For challenge 1: unblind existing data from the hold-out set For challenge 2: procure and test predicted hits with two orthogonal assays
300k AS–MS compound library (SMILES), protein target and annotated screening results (including orthogonal hit verification)	Challenge 1: use target-based and/or receptor-based virtual screening to predict experimental hits Challenge 2: if successful, predict novel hits from commercial libraries	Challenge 1: unblind existing data Challenge 2: procure and test predicted hits with two orthogonal assays
SMILES and/or fingerprint and enrichment metrics of DNA-encoded chemical library (DEL) screening hits and negatives from 4-10B compound libraries against hundreds of targets	Build a foundation model to predict hits from commercial libraries for targets absent from the training set	Procure and test predicted hits with two orthogonal assays
AS–MS screening and orthogonal confirmation data for 80% of >1,000 targets	Predict hits for homologous and/or unrelated targets	Procure and test predicted hits with two orthogonal assays

Table 3 | Community contributions

Project stage	Primary Actors	Community Contribution Opportunities
Target selection	Project participants, funders	Nominate targets
Protein production and assays	Experimental protein production hubs	Provide purified protein (academia, pharma, contract research organizations), protocols, tools, vectors
Screens	Experimental screening hubs	Companies or academic labs provide access to screening technologies and libraries
Experimental testing of predictions	Experimental assay hubs	Specialized contract research organizations, pharma and academic labs conduct orthogonal assays to test predictions for selected targets
Data management	Artificial Intelligence-Ready CHEmiCal Knowledge base (AIRCHECK), cloudservices providers	Cloud providers offer cloud credits, community deposit data
machine learning (ML)/artificial intelligence (AI) models and predictions	Project participants	ML/AI community, academic labs, subject matter experts generate and share predictions and models
Hit optimization and chemical probe generation	Scientific community	Structural Genomics Consortium, pharma and academic labs provide resources or donate high-quality probes

release would be established. The value of benchmarking initiatives in computational biology was clearly established by CASP, which, for over 30 years, has driven and monitored progressive improvements in computational methods[63,64].

Some of the proposed initial benchmarking challenges are listed in Table 2. As the project advances, other types of benchmarking challenges would probably be incorporated, including those that involve combining data from multiple platforms, not only from AS–MS and DEL screening but also from novel hit-finding screening platforms that may arise in the future. A combination of challenges that better represent a typical drug discovery screening pipeline may also have added value, including those that integrate some form of experimental or computational protein structural information. However, as even relatively simple challenges require significant logistics and the associated experimental costs are high, running more elaborate pipelines at the start of the project is probably too ambitious.

Participants will be encouraged to make their models open source and freely available to anyone for use directly from AIRCHECK. To encourage this, the costs of procuring compounds and testing them experimentally, partly or in full, would ideally be defrayed for qualified participants who make their ML/AI models publicly available and with permissive licenses.

From pilots to implementation

Pilot projects have laid foundational elements for this project. The capacities are now in place to

(1) produce more than 2,000 high-quality human proteins, most 'never previously liganded'; (2) screen these proteins against compound libraries using AS–MS and DEL; (3) store and disseminate project data, with a robust data management plan and database architecture; (4) annotate screening data and test predictions; and (5) solicit community contributions and participation.

In the first year of the project, the individual elements will be scaled and integrated to create a data generation plan that balances the shorter-term goal of identifying hits for high-priority proteins with the longer-term goal of generating data that will advance computational hit finding. The most likely screening cascade will involve screening each protein first by AS–MS against an exploratory (~15k) library whose composition will be made openly available. The rationale is that this screen is scalable, yields a direct binding readout, is the most cost effective and will most rapidly identify those proteins that are readily 'ligandable'. The exploratory screen will also flag proteins that have physiochemical properties that render them unsuitable for AS–MS or DEL and will not be screened further. For example, the exploratory screen would flag proteins that appear stable but that in fact have transiently unfolded regions that may bind large numbers of compounds nonspecifically.

Stable and monodisperse proteins that do not yield hits from the AS–MS exploratory screen, or for which greater chemical diversity or large datasets are required, will be channelled into screens with larger chemical libraries, using both AS–MS and DEL. The proposed screening cascade will be reviewed periodically and adjusted to optimize the process or incorporate other screening approaches as needed.

Encouraging community contributions

Active participation of the wider scientific community will be essential to meet the project goals. Robust community engagement will be made feasible only by adopting open science principles within the project. For clarity, this means compounds, data and algorithms developed using project resources will be made available without restriction on use, and without intellectual property constraints. This open science position provides a clarity of purpose and short-circuits what could be prolonged and complex discussions over ownership of compounds and algorithms. In keeping with this position, there will also be no restrictions on subsequent research or commercial use of data, chemical structures and algorithms generated using project resources. With this as background, community contributions in the following areas are envisioned (Table 3).

Protein scientists

Structural biologists, and protein scientists more broadly, often have unique expertise in purifying proteins in their scientific areas of interest. Community members would be encouraged to contribute their purified proteins to the screening process. For the project, this will expand the diversity of the protein–ligand datasets. For the contributing scientist, this could provide open access to hits that they can pursue without restriction in their own laboratories. Already >30 protein scientists have sent proteins to Toronto for AS–MS screening, including from Brazil, the UK, Canada, Germany, Sweden and the USA, and binders for 8 of these community proteins have already been identified, verified by surface plasmon resonance, and shared with the contributor (for example, Wang et al.[51]). Tapping into this diverse community at a larger scale will bring enormous scientific benefit, but will also add logistical burdens, so the project will need to implement this process carefully.

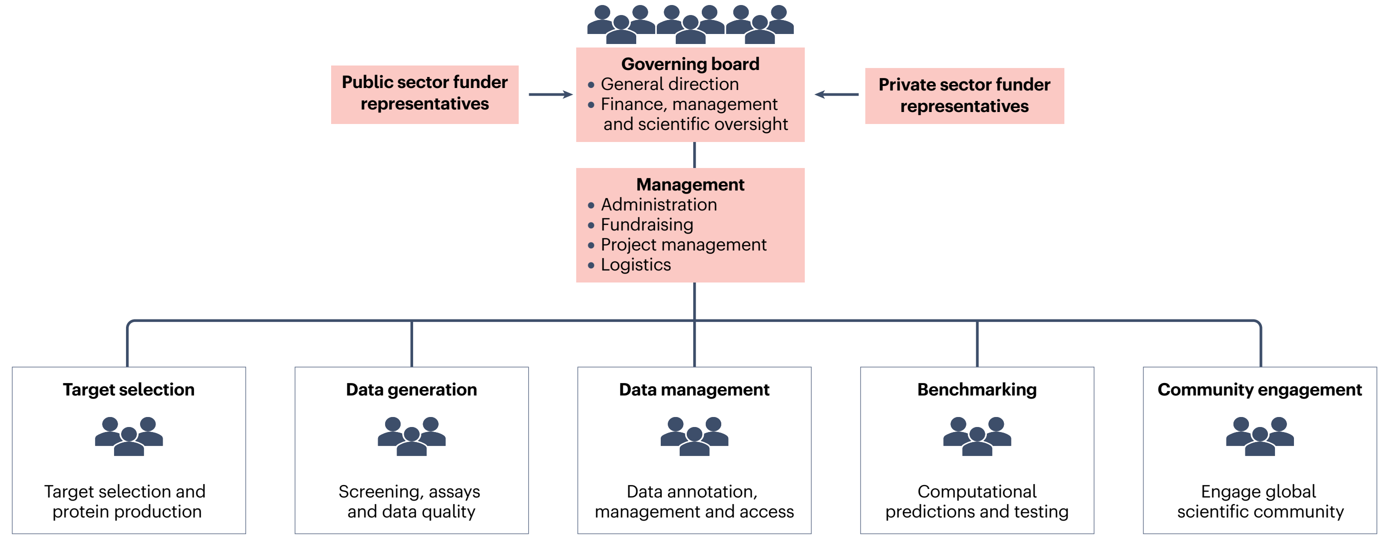

Public sector funder representatives
Governing board
• General direction
• Finance, management and scientific oversight
Private sector funder representatives
Management
• Administration
• Fundraising
• Project management
• Logistics
Target selection
Target selection and protein production
Data generation
Screening, assays and data quality
Data management
Data annotation, management and access
Benchmarking
Computational predictions and testing
Community engagement
Engage global scientific community

Fig. 5 | Project governance. The project is designed as a public–private partnership. The governance structure is designed to ensure efficient operation, strategic alignment and excellence in research. It integrates inputs from both public and private sector funders to direct a multitiered management system composed of specialized committees.

Data generation

Project screening data would be generated initially using the AS–MS and DEL screening platforms in selected hubs. However, there are clear advantages to expanding the number of participating screening laboratories, and the range of data generation technologies. Accordingly, new screening methodologies would be explored on a continual basis. To manage this process, a set of ~25 well-characterized, diverse and ligandable proteins that will have been screened comprehensively through all the initial platforms will serve as a technology test set for new screening hubs or technologies. The project board and its scientific advisers will review all data and provide recommendations about adding new centres or technologies.

Engaging computational scientists worldwide

Each screen will generate multiple GB-scale datasets, which may need to be downloaded and manipulated. The use of cloud resources will ensure the scalability of the AIRCHECK platform while allowing users to easily access the data and the computational resources for ML/AI modelling. It also allows users to leverage education or research credits from large cloud providers to support more equitable, diverse and inclusive access (for example, Google Cloud program for higher education in Africa[65]). Scientists from resource-poor environments will be actively encouraged to participate. We will also facilitate the development of open-source algorithms by collaborating closely with a project-associated global network of computational scientists, called MAINFRAME[66].

Chemists

The synthetic and medicinal chemistry communities will be encouraged (for example, through the SGC's Open Chemistry Networks) to design and/or generate molecules related to the hits to improve the original binders. Testing these compounds within the project may generate preliminary structure–activity relationships and provide confidence that the binder can be advanced. Chemists will also be encouraged to contribute compounds that are theoretically accessible through their chemistries to the emerging virtual screening library of all compounds that are synthetically accessible[67].

Training and networking

The project will be generating data explicitly to promote the development of ML/AI algorithms and as such will be operating at the intersection of experimental, data and computational sciences. This will provide

Table 4 | Metrics

Activity	Metric
Protein production	Number of proteins purified Structural and functional diversity of proteins screened Number of purified proteins contributed to the project from the community Geographic diversity of contributors
Screening	Number of assays developed and verified Reproducibility of assays Number of screens completed Amount of structured, machine-usable data generated Novelty of binders Number of new screening (data generation) technologies assessed
Benchmarking	Number and diversity of participants engaging in challenges Improvements in machine learning (ML) algorithms for binding and affinity prediction Number of freely available improved ML algorithms for binding, selectivity and affinity prediction
General	Extent of follow-on funding accrued to pursue or make use of confirmed hits Publications Number of new collaborators and new funders joining the project Creation of an engaged open community of scientists within Target 2035

Table 5 | Benefits to project participants

Stakeholder	Benefits
Academia	Access to real-world datasets for machine learning (ML)/artificial intelligence (AI) in drug discovery Cross-disciplinary training at the intersection of experiment and machine learning, including industry internships Opportunity to identify ligands for their own projects Freely available chemical starting points for new proteins Access to standardized datasets that can be used to develop and benchmark new methods Funding Collaborate with industry
Governments	Open data to drive economic growth Training of high-quality personnel in AI and drug discovery Catalysing partnerships with the pharma and AI sector Leveraged funding Democratizing early drug discovery
Industry	Protein and screening reagents and protocols Chemical starting points for unprecedented targets Access to and benchmarking of new protein and screening technologies Access to state-of-the-art ML models Access to trained scientists Collaborations with academic experts Leveraged funding
Foundations	Target and leverage consortium resources to problems of relevance to the foundation's mission

an excellent training environment for scientists seeking a working and operational knowledge of the various domains, and programmes for trainees will be established. Regular project meetings that prioritize scientific exchange between the various communities will be established.

Project structure and governance

The project will be structured as a pre-competitive, open science partnership in which compound assay data generated with project resources, including chemical structures of confirmed hits and algorithms, will be made available to the public under a license that requires attribution but that places no restriction on subsequent use. As stated previously, the rationale is pragmatic and evidence-based: pragmatic, in that it would be almost impossible to imagine a seamless cross-sectoral, cross-disciplinary and multinational collaboration that could operate under an agreement that allowed for the protection of potential intellectual property; and evidence- based, in that the development of ML/AI algorithms, in whatever field, advances most rapidly when provided with open data and with a mechanism to benchmark progress transparently[68].

The project needs to involve scientists from both public and private sectors to access the wide range of skill sets and expertise that will be required. It will also involve funding from both public and private sectors to achieve the requisite scale (Fig. 5). The major funders from the public and private sectors will form a governing board that oversees all project activities, including financial, scientific and management. The governing board will also oversee risk management, including any potential security risks associated with the data and the algorithms developed in the project. The governance board will be mandated to balance the needs of private sector funders with those of the public sector and its funding bodies, and also to provide a fair and time-limited mechanism for project or community contributors to pursue selected scientific questions. The governance structure that is currently used by the SGC is suitable because it has been used successfully to govern mission-oriented public–private partnerships of this complexity and scale[69].

A range of outcomes

The long-term aim of this project is to develop efficient computational hit-finding algorithms that can be used to generate freely available, small-molecule binders initially for thousands of proteins, and eventually for all relevant human proteins. However, over the course of the project, intermediate outcomes of considerable value will be generated, and these outcomes should be used as metrics to track and manage the project. Some of the key metrics are listed in Table 4.

A range of benefits to all participants

Open-access public–private partnerships are structures to carry out projects that require skills distributed among a wide range of academic and industry scientists, that tackle problems that span the boundary of public and private interests, and that might otherwise be crippled by intellectual property negotiations. However, in return for ceding their potential intellectual property rights to the public good, funders and participants must feel that they gain more than they lose, directly or indirectly. Table 5 lists some of the benefits that this project will generate for participants.

Published online: 23 July 2025

References

1. Edwards, A. M. et al. Too many roads not taken. *Nature* **470**, 163–165 (2011).
2. Moustakim, M. et al. Target identification using chemical probes. *Methods Enzymol.* **610**, 27–58 (2018).
3. Bond, M. J. & Crews, C. M. Proteolysis targeting chimeras (PROTACs) come of age: entering the third decade of targeted protein degradation. *RSC Chem. Biol.* **2**, 725–742 (2021).
4. Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P. & Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.* **49**, D562–D569 (2021).
5. Bender, B. J. et al. A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).
6. Petrović, D. et al. Virtual screening in the cloud identifies potent and selective ROS1 kinase inhibitors. *J. Chem. Inf. Model.* **62**, 3832–3843 (2022).
7. Alon, A. et al. Structures of the σ2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).
8. Stein, R. M. et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).
9. Ren, F. et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **43**, 63–75 (2025).
10. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
11. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
12. Zhu, T. et al. Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. *J. Med. Chem.* **56**, 6560–6572 (2013).
13. Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276 (2010).
14. Carter, A. J. et al. Target 2035: probing the human proteome. *Drug Discov. Today* **24**, 2111–2115 (2019).
15. Ackloo, S. et al. CACHE (Critical assessment of computational hit-finding experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* **6**, 287–295 (2022).
16. For chemists, the AI revolution has yet to happen. *Nature* **617**, 438 (2023).
17. Mock, M., Edavettal, S., Langmead, C. & Russell, A. AI can help to speed up drug discovery — but only if we give it the right data. *Nature* **621**, 467–470 (2023).
18. Martin, E. J. et al. All-assay-max2 pQSAR: activity predictions as accurate as four-concentration IC$_{50}$s for 8558 novartis assays. *J. Chem. Inf. Model.* **59**, 4450–4459 (2019).
19. Landrum, G. A. & Riniker, S. Combining IC$_{50}$ or Ki values from different sources is a source of significant noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).

20. Martin, E. J. & Zhu, X. W. Collaborative profile-QSAR: a natural platform for building collaborative models among competing companies. *J. Chem. Inf. Model.* **61**, 1603–1616 (2021).
21. Zardecki, C., Dutta, S., Goodsell, D. S., Voigt, M. & Burley, S. K. RCSB Protein Data Bank: a resource for chemical, biochemical, and structural explorations of large and small biomolecules. *J. Chem. Educ.* **93**, 569–575 (2016).
22. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–v (1995).
23. Edfeldt, K. et al. A data science roadmap for open science organizations engaged in early-stage drug discovery. *Nat. Commun.* **15**, 5640 (2024).
24. Thorne, N., Auld, D. S. & Inglese, J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **14**, 315–324 (2010).
25. Clark, M. A. et al. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.* **5**, 647–654 (2009).
26. McCloskey, K. et al. Machine learning on DNA-encoded libraries: a new paradigm for hit finding. *J. Med. Chem.* **63**, 8857–8866 (2020).
27. Li, A. S. M. et al. Discovery of nanomolar DCAF1 small molecule ligands. *J. Med. Chem.* **66**, 5041–5060 (2023).
28. Ahmad, S. et al. Discovery of a first-in-class small-molecule ligand for WDR91 using DNA-encoded chemical library selection followed by machine learning. *J. Med. Chem.* **66**, 16051–16061 (2023).
29. Kelly, M. A., McLellan, T. J. & Rosner, P. J. Strategic use of affinity-based mass spectrometry techniques in the drug discovery process. *Anal. Chem.* **74**, 1–9 (2002).
30. Prudent, R., Annis, D. A., Dandliker, P. J., Ortholand, J. Y. & Roche, D. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nat. Rev. Chem.* **5**, 62–71 (2021).
31. Gesmundo, N. J. et al. Nanoscale synthesis and affinity ranking. *Nature* **557**, 228–232 (2018).
32. L'Heureux, A., Grolinger, K., Elyamany, H. F. & Capretz, M. A. M. Machine learning with big data: challenges and approaches. *IEEE Access.* **5**, 7776–7797 (2017).
33. Najafabadi, M. M. et al. Deep learning applications and challenges in big data analytics. *J. Big Data* **2**, 1 (2015).
34. Lo, Y. C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 15–38-1546 (2018).
35. Brenner, S. & Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl Acad. Sci. USA* **89**, 5381–5383 (1992).
36. Melkko, S., Dumelin, C. E., Scheuermann, J. & Neri, D. Lead discovery by DNA-encoded chemical libraries. *Drug Discov. Today* **12**, 456–471 (2007).
37. Gironda-Martínez, A., Donckele, E. J., Samain, F. & Neri, D. DNA-encoded chemical libraries: a comprehensive review with succesful stories and future challenges. *ACS Pharmacol. Transl. Sci.* **4**, 1265–1279 (2021).
38. Peterson, A. A. & Liu, D. R. Small-molecule discovery through DNA-encoded libraries. *Nat. Rev. Drug Discov.* **22**, 699–722 (2023).
39. Lim, K. S. et al. Machine learning on DNA-encoded library count data using an uncertainty-aware probabilistic loss function. *J. Chem. Inf. Model.* **62**, 2316–2331 (2022).
40. Tingle, B. I. et al. ZINC-22 — a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
41. Ackloo, S. et al. A target class ligandability evaluation of WD40 repeat-containing proteins. *J. Med. Chem.* **68**, 1092–1112 (2024).
42. Han, S. et al. Highly selective novel heme oxygenase-1 hits found by DNA-encoded library machine learning beyond the DEL chemical space. *ACS Med. Chem. Lett.* **15**, 1456–1466 (2024).
43. SGC and HitGen announce research collaboration focused on DNA-encoded library based drug discovery. *HitGen* https://www.hitgen.com/en/news-details-319.html (2023).
44. X-chem and structural genomics consortium enter into collaboration to unlock the human proteome and promote open science. *X-Chem* https://www.x-chemrx.com/about/news/x-chem-and-structural-genomics-consortium-enter-into-collaboration-to-unlock-the-human-proteome-and-promote-open-science/ (2023).
45. Wellnitz, J. et al. Enabling open machine learning of DNA encoded library selections to accelerate the discovery of small molecule protein binders. Preprint at https://doi.org/10.26434/chemrxiv-2024-xd385 (2024).
46. Prudent, R., Lemoine, H., Walsh, J. & Roche, D. Affinity selection mass spectrometry speeding drug discovery. *Drug Discov. Today* **28**, 103760 (2023).
47. Xin, Y. et al. Affinity selection of double-click triazole libraries for rapid discovery of allosteric modulators for GLP-1 receptor. *Proc. Natl Acad. Sci. USA* **120**, e2220767120 (2023).
48. Liu, J. et al. The omega-3 hydroxy fatty acid 7(S)-HDHA is a high-affinity PPARα ligand that regulates brain neuronal morphology. *Sci. Signal.* **15**, eabo1857 (2022).
49. Zhang, P. et al. Development of an α-klotho recognizing high-affinity peptide probe from in-solution enrichment. *JACS Au* **4**, 1334–1344 (2024).
50. Muchiri, R. N. & van Breemen, R. B. Affinity selection–mass spectrometry for the discovery of pharmacologically active compounds from combinatorial libraries and natural products. *J. Mass Spectrom.* **56**, e4647 (2021).
51. Wang, X. et al. Enantioselective protein affinity selection mass spectrometry (EAS-MS). Preprint at https://doi.org/10.1101/2025.01.17.633682 (2025).
52. Paillard, G. et al. The ELF Honest Data Broker: informatics enabling public–private collaboration in a precompetitive arena. *Drug Discov. Today* **21**, 97–102 (2016).
53. Quancard, J. et al. The European Federation for Medicinal Chemistry and Chemical Biology (EFMC) best practice initiative: hit generation. *ChemMedChem* **18**, e202300002 (2023).
54. Giannetti, A. M., Koch, B. D. & Browner, M. F. Surface plasmon resonance based assay for the detection and characterization of promiscuous inhibitors. *J. Med. Chem.* **51**, 574–580 (2008).
55. Rich, R. L. & Myszka, D. G. Grading the commercial optical biosensor literature — class of 2008: 'The Mighty Binders'. *J. Mol. Recognit.* **23**, 1–64 (2010).
56. Understanding SPR data. *Critical Assessment of Computational Hit-Finding Experiments (CACHE)* https://cache-challenge.org/sites/default/files/downloadable/forms/understanding_SPR_data.pdf (2024).
57. Wood, R. W. XLII. On a remarkable case of uneven distribution of light in a diffraction grating spectrum. *Lond. Edinb. Dubl. Phil. Mag. J. Sci.* **4**, 396–402 (1902).
58. Kartal, Ö., Andres, F., Lai, M. P., Nehme, R. & Cottier, K. waveRAPID — a robust assay for high-throughput kinetic screens with the creoptix WAVEsystem. *SLAS Discov.* **26**, 995–1003 (2021).
59. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2**, 2212–2221 (2007).
60. Sparks, R. P. & Fratti, R. in *Methods in Molecular Biology* (ed. Fratti, R.) 1860, 191–198 (2019).
61. Langer, A. et al. A new spectral shift-based method to characterize molecular interactions. *Assay Drug Dev. Technol.* **20**, 83–94 (2022).
62. Meyer, P. & Saez-Rodriguez, J. Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. *Cell Syst.* **12**, 636–653 (2021).
63. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
64. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
65. Manoharan, F. Google cloud expands higher education credits to 8 countries in Africa. *Google Cloud* https://cloud.google.com/blog/topics/public-sector/google-cloud-expands-higher-education-credits-8-countries-africa/ (2022).
66. MAchine learning Innovation Network For Research to Advance MEdicinal chemistry. *MAINFRAME* https://www.aircheck.ai/mainframe (2025).
67. Bedart, C. et al. The pan-Canadian chemical library: a mechanism to open academic chemistry to high-throughput virtual screening. *Sci. Data* **11**, 597 (2024).
68. Burley, S. K. & Berman, H. M. Open-access data: a cornerstone for artificial intelligence approaches to protein structure prediction. *Structure* **29**, 515–520 (2021).
69. Edwards, A. Reproducibility: team up with industry. *Nature* **531**, 299–301 (2016).
70. Mammoliti, A. et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. *Nat. Commun.* **12**, 5797 (2021).
71. Accessibility principles. *Web Accessibility Initiative (WAI)* https://www.w3.org/WAI/fundamentals/accessibility-principles/ (2024).

Acknowledgements

The SGC is a registered charity (no. 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute (OGI-196), Janssen, Merck KGaA (aka EMD in Canada and USA), Pfizer, Takeda and the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement no. 875510. This work was also funded by the Member States of the European Molecular Biology Laboratory. A.A.A. is an ISCIII–Miguel Servet Fellow supported by the Instituto de Salut Carlos III grant CP23/00115 and by the Spanish Ministry of Science and Innovation (MCIN/AEI) (PID2022- 136344OA-I00); CERCA Program/Generalitat de Catalunya, and FEDER funds/European Regional Development Fund (ERDF) — a way to Build Europe.

Competing interests

D.-A.C., K.S. and D.R.O. are shareholders in Pfizer Inc. The Cernak Lab's research has been supported by MilliporeSigma, Johnson & Johnson, Relay Therapeutics, Merck & Co., Inc., SPT Labtech, National Defense Medical Center, Shanghai University of Traditional Chinese Medicine, Ministry of Education Taiwan, and Entos, Inc. T.C. has consulted for the University of Dundee Drug Discovery Unit, Scorpion Therapeutics, Relay Therapeutics, Amgen, Genentech, Janssen, Pfizer, Vertex, MilliporeSigma, the US Food & Drug Administration, Gilead, AbbVie, Corteva, Syngenta, Firmenich, Biogen, Bayer, UCB Biopharma, National Taiwan University, AstraZeneca, Grunenthal, and Iambic Therapeutics (previously known as Entos, Inc.). He holds equity in Scorpion Therapeutics and is a co-founder and equity holder at Iambic Therapeutics. B.H.-K. is a co-Founder of the MAQC (Massive Analysis and Quality Control) Society and part of the Scientific Advisory Board of: Consortium de recherche biopharmaceutique (CQDM), Quebec, Canada, Break Through Cancer, Commonwealth Cancer Consortium, United States, Canadian Institute of Health Research–Institute of Genetics, Canada, Cancer Grand Challenges, United Kingdom, Shriners Children, United States. He is part of the Executive Committee of the Terry Fox Digital Health and Discovery Platform, Canada and in the Board of Directors of AACR International–Canada, The American Association for Cancer Research, United States. D.W.Y. is co-founder and shareowner of Deliver Therapeutics. I.V.H. is part of the Board of Directors of TenAces Biosciences. A.K. serves on the SAB of Cilcare, Sulfateq BV and Heartbeat.bio. J.C.M. may hold stock options in Astrazeneca. A.M.-F. is the Board Chair of SGC and Conscience. She is also a shareholder for Bayer AG and an external consultant for Nuvisan ICB GmbH. N.B.-B. is on the SAB for Oxford Vacmedix and holds shares of Exact Sciences. A.T. is co-founder of Predictive LLC. A.S.D. holds stocks in DANAHER. F.K. is a shareholder in Evotec

SE. A.F.-M. possess Bayer AG shares. A.A.A. is a consultant to Darwin Health and has received grant funding from Vivan Therapeutics and AtG Therapeutics. D.P. holds stock in Novartis.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41570-025-00737-z.

Peer review information *Nature Reviews Chemistry* thanks Brian Shoichet, Brent Stockwell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

The Structural Genomics Consortium Target 2035 Working Group

Aled M. Edwards[1], Dafydd R. Owen[2], Leili Zhang[3], Damian W. Young[4], Timothy M. Willson[5], James Wellnitz[5,6], Yanli Wang[7], Jarrod Walsh[8], Erik Vernet[9], Alexander Tropsha[5,6], Claudia Tredup[10,11], Matthew H. Todd[12], Amelia Tjaden[10,11], Sven Thamm[13], Michael Sundström[14], Andreas Steffen[15], Shaun Stauffer[16], Lucas Rodrigo de Souza[17], Min Shen[18], Kristof Schütt[19], Lovisa Holmberg Schiavone[20], Matthieu Schapira[1], Santha Santhakumar[1], Kumar Saikatendu[21], Emma Rivers[8], Dušan Petrović[22], Hui Peng[1], John P. O'Donnell[23], Susanne Müller-Knapp[10,11], Anke Mueller-Fahrnow[24], Maxwell R. Morgan[1], Florian Montel[13], Juan Carlos Mobarec[25], Maurice Michel[26,27], Sofia Melliou[1], Uta Lessel[13], Andrew R. Leach[28], Oliver Krämer[29], Florian Krieger[30], Stefan Knapp[10,11], Anthony D. Keefe[31], Aimo Kannt[32], Scott A. Johnson[33], Sandra Häberle[34], Emily Rose Holzinger[35], Ingo V. Hartung[36], Rachel J. Harding[1], Thomas Hanke[10,11], Levon Halabelian[1], Benjamin Haibe-Kains[1], Judith Günther[37], Marie-Aude Guié[31], Claudia Gordijo[1], Opher Gileadi[14], Luca Foschini[38], Amaury Fernández-Montalván[13], Ola Engkvist[39,40], Madison M. Edwards[1], Katharina Duerr[41], David Drewry[5], Dengfeng Dou[42], Snezana Djordjevic[12], Alejandra Solache Diaz[43], Sergio Martinez Cuesta[44], Rafael Counago[5], Wendy D. Cornell[3], Jesse A. Coker[16], Djork-Arné Clevert[15], Timothy Cernak[45], Nicola A. Burgess-Brown[12], Peter J. Brown[5], Mario H. Bengtson[17], Frances M. Bashore[5], Dalia Barsyte-Lovejoy[1], Arrash J. Baghaie[31], Alison D. Axtman[5], Cheryl Arrowsmith[1], Albert A. Antolin[46] & Suzanne Ackloo[1]

[3]IBM Accelerated Discovery Research, Yorktown Heights, NY, USA. [4]Collaborative Drug Discovery (CDD) Baylor College of Medicine One Baylor Plaza, Houston, TX, USA. [5]Structural Genomics Consortium, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [6]Division of Chemical Biology and Medicinal Chemistry UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [7]National Institutes of Health, Bethesda, MD, USA. [8]Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK. [9]Research & Early Development, Novo Nordisk A/S, Måløv, Denmark. [10]Institute of Pharmaceutical Chemistry, Johann Wolfgang Goethe University, Frankfurt, Germany. [11]Buchmann Institute for Molecular Life Sciences and Structural Genomics Consortium (SGC), Frankfurt, Germany. [12]Structural Genomics Consortium, School of Pharmacy, University College London, London, UK. [13]Discovery Research, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany. [14]Structural Genomics Consortium, Department of Medicine, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden. [15]Machine Learning and Computational Sciences, Pfizer Research and Development, Berlin, Germany. [16]Center for Therapeutics Discovery, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [17]Center for Molecular Biology and Genetic Engineering (CBMEG), Universidade Estadual de Campinas (UNICAMP), Campinas/SP. Center for Medicinal Chemistry (CQMED), Universidade Estadual de Campinas (UNICAMP), Campinas/SP, Brazil. [18]National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD, USA. [19]Pfizer Research and Development, Machine Learning and Computational Sciences, Berlin, Germany. [20]Protein Science, Structural Biology and Biophysics, Discovery Sciences, Research and Development, AstraZeneca, Gothenburg, Sweden. [21]Takeda Pharmaceuticals, San Diego, CA, USA. [22]Digital Life Sciences, Nuvisan ICB GmbH, Berlin, Germany. [23]Research & Development, Pharmaceuticals, Bayer AG, Monheim, Germany. [24]Nuvisan Innovation Campus Berlin GmbH, Berlin, Germany. [25]Protein, Structure and Biophysics, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK. [26]Science for Life Laboratory, Department of Oncology and Pathology, Karolinska Institute, Stockholm, Sweden. [27]Center for Molecular Medicine, Karolinska Institute and Karolinska Hospital, Stockholm, Sweden. [28]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. [29]Discovery Research, Boehringer Ingelheim International GmbH, Ingelheim, Germany. [30]Evotec SE, Hamburg, Germany. [31]X-Chem Inc., Waltham, MA, USA. [32]Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Frankfurt, Germany. [33]Bristol Myers Squibb, San Diego, CA, USA. [34]Structural Genomics Consortium Frankfurt, Goethe University Frankfurt, Frankfurt, Germany. [35]Bristol Myers Squibb, Cambridge, MA, USA. [36]Discovery and Development Technologies, Merck KGaA, Darmstadt, Germany. [37]Bayer AG, Drug Discovery Sciences, Berlin, Germany. [38]Sage Bionetworks, Seattle, WA, USA. [39]Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden. [40]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden. [41]OMass Therapeutics Ltd, ARC, Oxford, UK. [42]HitGen Inc., Chengdu, China. [43]Abcam, Biomedical Campus, Cambridge, UK. [44]Data Sciences and Quantitative Biology, Discovery Sciences, R&D BioPharmaceuticals, AstraZeneca, Cambridge, UK. [45]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA. [46]proCURE Department, Oncobell Program, Catalan Institute of Oncology (ICO) and Bellvitge Biomedical Research Institute (IDIBELL), Hospital Duran y Reynals, L'Hospitalet del Llobregat, Barcelona, Spain.