CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding

Suzanne Ackloo[1], Rima Al-awar[2,3], Rommie E. Amaro[4,5], Cheryl H. Arrowsmith[1], Hatylas Azevedo[6], Robert A. Batey[7], Yoshua Bengio[8], Ulrich A.K. Betz[9], Cristian G. Bologa[10], John D. Chodera[11], Wendy D. Cornell[12], Ian Dunham[13,14], Gerhard F. Ecker[15], Kristina Edfeldt[16], Aled M. Edwards[1*], Michael K. Gilson[17,18], Claudia R. Gordijo[1], Gerhard Hessler[19], Alexander Hillisch[20]*, Anders Hogner[21], John J. Irwin[22], Johanna M. Jansen[23], Daniel Kuhn[24], Andrew R. Leach[13,14], Alpha A. Lee[25,26], Uta Lessel[27], John Moult[28,29], Ingo Muegge[30], Tudor I. Oprea[10,31], Benjamin G. Perry[32], Patrick Riley[33], Kumar Singh Saikatendu[34], Vijayaratnam Santhakumar[1], Matthieu Schapira[1,3], Cora Scholten[35], Matthew H. Todd[36], Masoud Vedadi[1,3], Andrea Volkamer[37], Timothy M. Willson[38]

*Corresponding authors: aled.edwards@utoronto.ca and alexander.hillisch@bayer.com

[1] Structural Genomics Consortium, University of Toronto, Toronto, Ontario, Canada
[2] Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[3] Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada
[4] Department of Chemistry and Biochemistry, UC San Diego, La Jolla, CA, USA
[5] Drug Design Data Resource, University of California, San Diego, La Jolla, CA, USA
[6] Aché Laboratórios Farmacêuticos, Guarulhos, São Paulo, Brazil
[7] Department of Chemistry, University of Toronto, Toronto, Ontario, Canada
[8] Mila, University of Montreal, Québec, Canada
[9] Merck Healthcare KGaA, Darmstadt, Germany
[10] Department of Internal Medicine, University of New Mexico School of Medicine, University of New Mexico Albuquerque, Albuquerque, NM, USA
[11] Computational and Systems Biology Program Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[12] Healthcare & Life Sciences Research, IBM TJ Watson Research Center, New York, USA
[13] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK
[14] Open Targets, Wellcome Genome Campus, Hinxton, UK
[15] Department of Pharmaceutical Sciences, University of Vienna, Vienna, Austria
[16] Structural Genomics Consortium, Department of Medicine, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden
[17] Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, CA, USA
[18] Drug Design Data Resource, University of California, San Diego, La Jolla, CA, USA
[19] Sanofi-Aventis Deutschland GmbH, R&D, Integrated Drug Discovery, Frankfurt am Main, Germany
[20] Research and Development, Bayer AG, Pharmaceuticals, Wuppertal, Germany
[21] Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
[22] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA

[23] Novartis Institutes for BioMedical Research, Emeryville, CA, USA
[24] Merck Healthcare KGaA, Computational Chemistry & Biologics, Darmstadt, Germany
[25] PostEra Inc, San Franciso, CA, USA
[26] Department of Physics, University of Cambridge, Cambridge, UK
[27] Boehringer Ingelheim Pharma GmbH & Co. KG, Medicinal Chemistry, Biberach an der Riss, Germany
[28] Institute for Bioscience and Biotechnology Research, Rockville, MD, USA
[29] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA
[30] Alkermes, Inc., Waltham, MA, USA
[31] University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, USA
[32] Drugs for Neglected Diseases initiative, Geneva, Switzerland
[33] Relay Therapeutics, Boston, MA, USA
[34] Global Research Externalization, Takeda California, Inc., San Diego, CA, USA
[35] Bayer AG, Open Innovation – Public Private Partnerships, Pharmaceuticals, Berlin, Germany
[36] School of Pharmacy, University College London, London, UK
[37] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Berlin, Germany
[38] Structural Genomics Consortium, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ABSTRACT

Computational approaches in drug discovery and development hold great promise, with artificial intelligence methods undergoing widespread contemporary use, but the experimental validation of these new approaches is frequently inadequate. We are initiating Critical Assessment of Computational Hit-finding Experiments (CACHE) as a public benchmarking project that aims to accelerate the development of small molecule hit-finding algorithms by competitive assessment. Compounds will be identified by participants using a wide range of computational methods for dozens of protein targets selected for different types of prediction scenarios, as well as for their potential biological or pharmaceutical relevance. Community-generated predictions will be tested centrally and rigorously in an experimental hub(s), and all data, including the chemical structures of experimentally tested compounds, will be made publicly available without restrictions. The ability of a range of computational approaches to find novel compounds will be evaluated, compared, and published. The overarching goal of CACHE is to accelerate the development of computational chemistry methods by providing rapid and unbiased feedback to those developing methods, with an ancillary and valuable benefit of identifying new compound-protein binding pairs for biologically interesting targets. The initiative builds on the power of crowd sourcing and expands the open science paradigm for drug discovery.

INTRODUCTION

The past decades have witnessed continuous incremental improvements in the computational methods used to facilitate small molecule drug discovery and development. However, the field is currently witnessing a revived optimism, fueled by continuous leaps in computational power, increased accessibility to diverse chemical space, improved physics-based methods, and the emerging potential of newer machine learning and AI approaches. It is fair to say that the question is not whether *in silico* design will grow as a part of drug discovery, but how much and how quickly these methods will reduce the experimental discovery cycle. This optimism must be tempered by the fact that considerable additional progress is needed. Today, no algorithm can select, design, or rank potent, drug-like small molecule protein binders consistently. Predicting pharmacokinetic, pharmacodynamic and toxicity (ADMET) properties is even more difficult, as this involves the interactions of a candidate drug with the entire biological system, rather than a single protein target.

Significant advances in computational methods can be gained through blinded benchmarking exercises, as evidenced by community progress in developing computational methods to predict protein structure from primary sequence. In 1993 when the "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) exercise[1] (https://predictioncenter.org/) was launched, humans were often better at predicting protein structures than were computational methods. Leveraging high-quality and openly available experimental information in a well-organized database (the Worldwide Protein Data Bank [www.wwpdb.org]), *in silico* approaches gradually became increasingly capable of predicting protein structure. Very recently machine learning algorithms have been able to predict the structures of many (but not all) globular proteins as accurately as can be determined experimentally [2,3].

In computational chemistry and specifically in protein-targeted small molecule discovery, organizing benchmarking exercises similar to CASP has occurred [4-12], but none are currently operational. In addition, other than the TDT and DREAM benchmarking initiatives [7,8,12] which included a prospective arm to its prediction challenge, there has been no concerted effort to provide experimental testing of predictions. This includes no opportunity to fund the synthesis and quality control of predicted compounds and to test their binding rigorously under standard conditions. Commercial sensitivities also complicate small molecule binding benchmarking. A large fraction of the experimental data suitable for benchmarking *in silico* binding predictions are generated within the pharma industry and kept confidential, rather than being released for general use. In addition, significant advances in computational chemistry technologies are taking place within companies, and massive private investment is flowing into new companies for the development of artificial intelligence (AI) methods. These companies are also likely reluctant to share their methods in any detail, or see them put to the test publicly. Finally, unlike the prediction of experimental protein structures, which is a narrowly-scoped problem with a relatively unambiguous solution, benchmarking the prediction of useful small molecules against a target is much more difficult – multiple possible solutions differ both quantitatively and qualitatively.

In the past few years, a combination of scientific and technical advances has made it possible to conceptualize a benchmarking exercise that can overcome some of these limitations. Among the most important is the creation of ultra-large libraries of chemicals that can be described *in silico* and procured "on-demand" [13,14]. This reduces significantly the cost associated with accessing chemical matter to test predictions. There has also been a dramatic increase in the availability of computational resources, which facilitates data sharing and democratizes the ability to make predictions. Hand in hand with the improvements in infrastructure has been the development of new computational algorithms, both machine learning and physics-based. Recently, all these advances were mobilized within community-wide efforts to identify new chemical starting points for SARS-CoV-2 drug discovery [15].

The past years have also seen increased community acceptance of the idea that public and private sectors can collaborate pre-competitively in areas that were once considered commercially sensitive. The "open-access, open-source, open-data" paradigm is now accepted as an accelerator of biomedical science. Critically, this paradigm has expanded to include the notion that placing chemical matter, including advanced molecules such as chemical probes, in the public domain without complex and rate-limiting intellectual property agreements provides immense scientific value [16].

Based on this new landscape, we propose to create a public-private partnership called Critical Assessment of Computational Hit-finding Experiments (CACHE) to benchmark computational approaches to finding small molecule hits for proteins. Modelled after CASP, CACHE will organize hit-finding challenges against selected targets and participants will use various computational methods to predict hits. However, unlike CASP, which was able to piggy-back on experiments being done in the structural biology community, CACHE must have an experimental arm testing predictions prospectively. Each challenge will typically include two testing iterations to enable refinement and forward application of predictive models. Upon completion of a hit-finding challenge, all data generated by CACHE, including all chemical structures, will be publicly available without restrictions on use.

THE GENESIS OF THE CACHE CONCEPT

In November 2020, prompted by recent developments and interest in computational methods, including deep learning, as well as the challenges in identifying the best performing methods, ~80 scientists from industry, academia, and funding agencies met virtually to consider potential areas of drug discovery that might benefit from coordinated benchmarking. Of the many areas that were identified, the group prioritized hit-finding – identification of a small molecule that binds a targeted protein with high enough affinity to qualify as a credible starting point for a drug discovery project – as particularly suitable and practical, and an excellent area to begin. To advance the idea, a set of ~30 representatives developed a draft concept for CACHE in four working groups, which focused on: 1) Target selection and prioritization; 2) Virtual library construction; 3) Measuring outcomes; and 4) Governance. Here we present these groups' ideas for the CACHE project.

THE CACHE CONCEPT

CACHE will present and organize a variety of "hit-finding" challenges to the community. As a part of this, and as described in detail below, CACHE will identify suitable protein targets, curate the virtual chemical libraries, define success parameters for predictions, and solicit predictions for hit compounds. For evaluation, CACHE will purchase or otherwise procure the predicted compounds, experimentally measure the binding of the compounds to their intended target, calculate other key properties of the active compounds, and share the outcomes openly with the community (FIG. 1). We envision that CACHE, like CASP, will organize multiple rounds of challenges, providing on-going opportunities for algorithm developers to improve and test their methods.

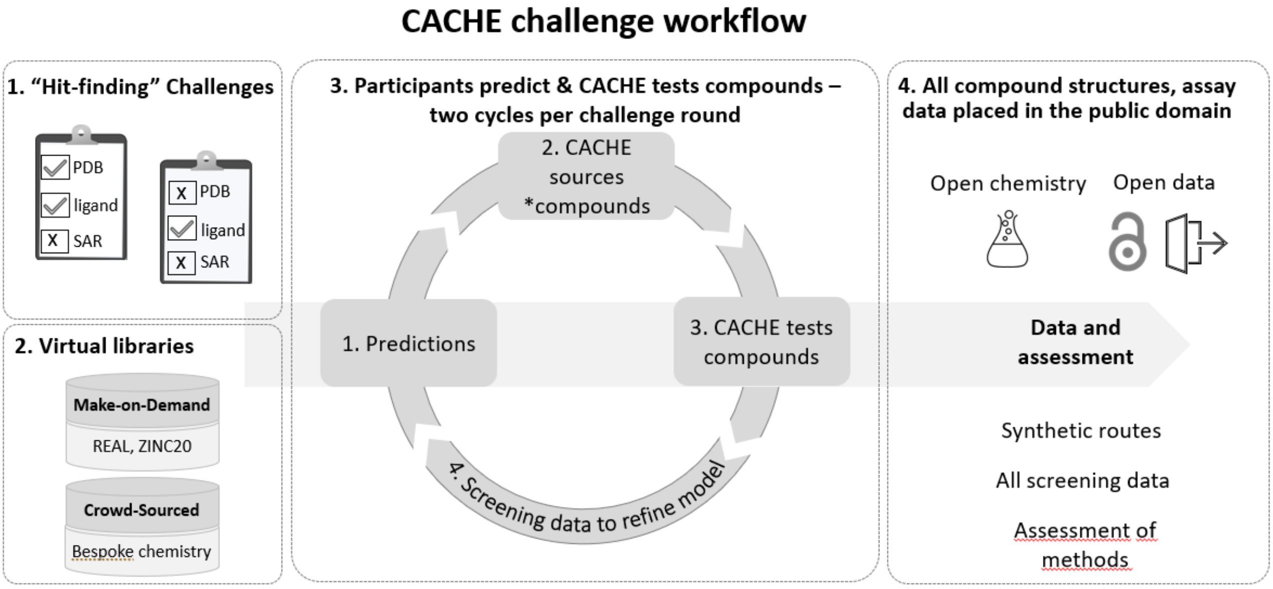


Fig. 1. CACHE challenge workflow. 1. "Hit-finding" challenges: CACHE presents a variety of "hit-finding" challenges to the community, including assessment criteria. **2. Virtual libraries:** CACHE will establish and host two virtual libraries; a make-on-demand library (REAL, ZINC20) and a library comprising compounds synthetically accessible by chemists in academia or industry (bespoke chemistry). **3. Participants predict chemical matter & CACHE experimentally tests compounds:** Each participant will have opportunity to make two cycles of predictions per round. CACHE will procure and assay the predicted compounds. At this stage, structures of compounds will be made available to all participants, but screening data will be provided only to the specific participant and competition management, in order to serve as a starting point for an additional cycle of predictions. **4. Compounds, data placed in the public domain:** Once the second cycle is complete, the data package, including all structures and screening data, as well as an assessment of each compound, will be made available to all, without restriction.

CACHE CHALLENGES & TARGET SELECTION

CACHE will organize hit-finding challenges that represent the range of common scenarios encountered in hit-finding depending on available target information (FIG. 2, Panel B). A Target Selection Group will select targets appropriate for each of these 5 scenarios. Specifically, the Target Selection Group will define the acceptance criteria for targets in each scenario, bioinformatically create a "long list" of targets that meet these criteria, and create a mechanism(s) for the community, including the funders of CACHE, to prioritize from this list those targets to be included in the launched benchmarking challenges.

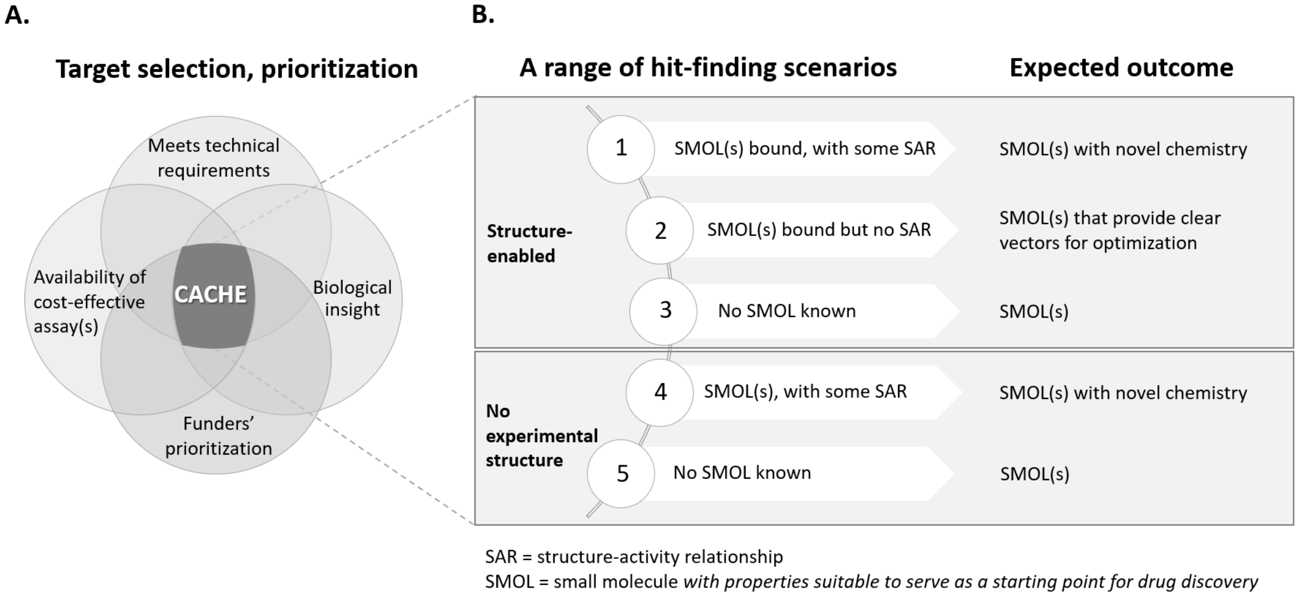

A.
B.
Target selection, prioritization
Meets technical requirements
Availability of cost-effective assay(s)
CACHE
Biological insight
Funders' prioritization
A range of hit-finding scenarios
Expected outcome
Structure-enabled
1 SMOL(s) bound, with some SAR
SMOL(s) with novel chemistry
2 SMOL(s) bound but no SAR
SMOL(s) that provide clear vectors for optimization
3 No SMOL known
SMOL(s)
No experimental structure
4 SMOL(s), with some SAR
SMOL(s) with novel chemistry
5 No SMOL known
SMOL(s)
SAR = structure-activity relationship
SMOL = small molecule *with properties suitable to serve as a starting point for drug discovery*

Fig. 2. Target selection consideration and classes of CACHE challenges. A. Targets will be selected from a long list of proteins that represent a range of scenarios of varying technical difficulty, are experimentally enabled (for example, there must be a robust binding assay) and, where possible, represent opportunities to make new biological or medical discoveries. Funders can prioritize targets within each challenge. **B.** The five potential "hit-finding" scenarios that address key technical questions in computational chemistry. SMOL: small molecule.

The most important technical requirement for inclusion of a potential target into the list will be the availability of two cost-effective direct binding assays that can provide rapid, validated, high-quality experimental feedback. From this list of targets that meet technical requirements, CACHE and its funders will select the final targets using a prioritization scheme that maximizes the structural diversity of the target proteins, and ideally that also takes into account the opportunity to discover new biological insights, so that CACHE outputs benefit not only the computational and pharmaceutical communities. We anticipate that a funder (such as a disease-focused charity) would consider CACHE as an attractive funding opportunity through the mobilization of a wide global network of computational chemists to focus on their priority target(s) (FIG. 2). We also imagine that, in lieu of providing direct support, funders, foundations, or companies might also offer in-kind support for CACHE, for example, by offering to experimentally test all predictions for a given target, or providing access to computational resources, assay reagents, or laboratory equipment. Ideally, for each of the 5 scenarios, CACHE will have the resources to pursue at least 3 different targets at a time.

CACHE WILL FACILITATE AND ENCOURAGE PARTICIPATION

CACHE will provide virtual compound libraries

To enable rapid and cost-effective testing of predictions, CACHE will establish a well-defined and robust "Core" Make-on-Demand (MoD) virtual library comprising compounds that are readily accessible from commercial vendors, at reasonable cost. A combination of Enamine REAL* and ZINC20 [14] might comprise the core of this library.

CACHE will annotate compounds in the library with predicted physical properties (e.g., cLogP, PSA, Fsp3, etc.) that will be assessed in the success criteria (see next section). The aim is to enable each participant to select individual subsets and/or apply relevant filtering which he/she sees fit, while ensuring any such pre-filtering or sub-set restrictions can be accounted for in any subsequent evaluation and comparison of approaches. CACHE will also create "Subsets" within the initial library as may be required to account for the needs of specific CACHE participants (e.g., a 1% diversity set or a 10% diversity set might be preferred when examining computationally intensive approaches, etc.). The libraries will evolve, adding more compounds as they become commercially available or accessible, and creating additional library subsets as feedback on the performance is collected.

CACHE will also explore mechanisms to provide participants access to a virtual library (VL) containing new chemistry. The concept here is to ask synthetic chemists within academic or industry to contribute to a "New Chemical Space" VL by adding compounds that they would be willing to synthesize "on demand" in a timely manner with their emerging synthetic chemistry protocols, and using their own resources. The purpose of including these compounds is to access structurally unique scaffolds or sub-libraries. This library could serve as a counterpoint to the "parallel synthesis" MoD library and would also provide a mechanism for academic chemists to get their new chemistry screened by a large number of computational methods against a wide range of biologically interesting targets, and in turn for CACHE to sample new and exciting compounds.

At regular and defined intervals over the course of the CACHE benchmarking exercises, the CACHE virtual libraries team will evaluate the impact of library choice, composition and nature (diversity, size) on both virtual screening capabilities and on general screening success and recommend changes accordingly.

CACHE will provide robust experimental testing of predictions

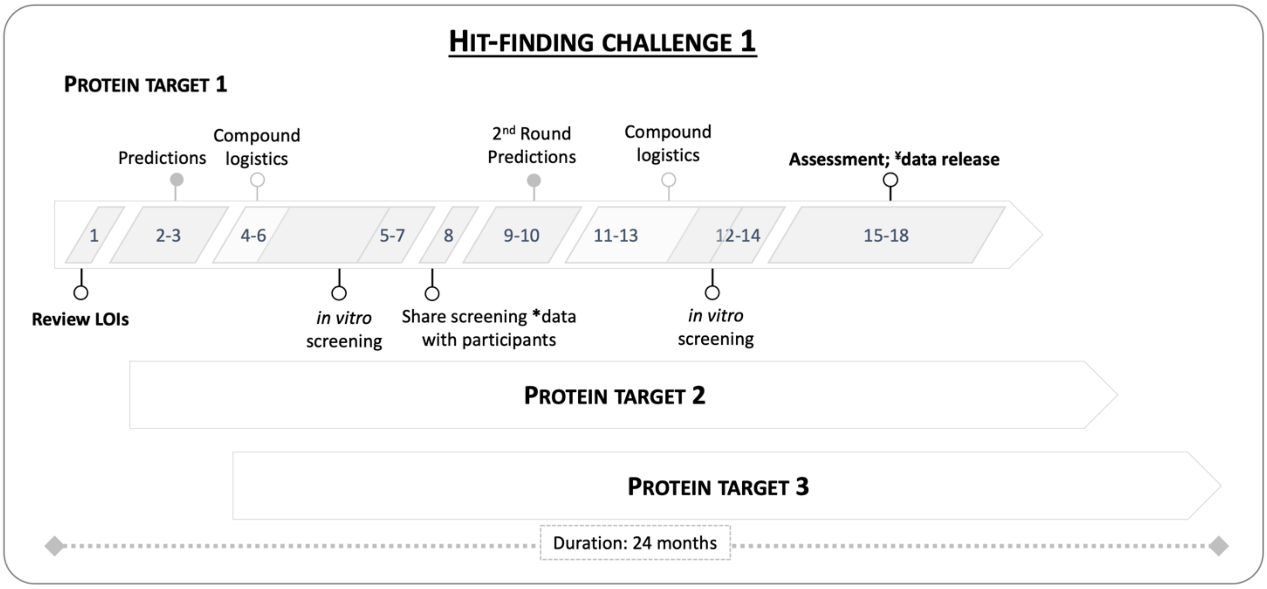
At the core of the CACHE initiative will be an experimental hub (or hubs) that will provide rapid, high-quality testing of the predictions. Predicted compounds will be submitted to the experimental hub, which will procure the compounds and assay them using a binding assay selected to be most appropriate for the protein target. Each compound will be assayed at a single concentration in duplicate, and each positive re-tested in dose-response mode, as well as in an orthogonal biophysical

* https://enamine.net/compound-collections/real-compounds/real-space-navigator

assay, which is critical. Feedback will be given first to the predictor(s), who will have the opportunity to submit a new set of predictions.

Each CACHE challenge round will take ~18 months, with two cycles of predictions per round

The CACHE challenge will involve two cycles of predictions in order to give participants the opportunity to incorporate learnings from the first round into their designs. The timing and sequence of the proposed challenge round is shown in FIG. 3. Challenges will be staggered in order to avoid overwhelming the experimental hub. As part of each challenge, participants will be asked to make predictions from a small library constituting the combined list of predicted compounds contributed to the first cycle by all participants. Experimental testing of these compounds and then comparing with predictions will facilitate inter-algorithm benchmarking.



*participants are anonymized; ¥ participants with top performing methods are de-anonymized

Fig. 3. The timelines of challenge activities. After reviewing the Letters of Intent (LOI), each complete challenge round will take ~18 months, with the various stages outlined.

CACHE BENCHMARKING

Benchmarking virtual screening methods poses a challenge, because no single measure, or even combination of measures, can be used to unambiguously quantify the success of virtual screens, let alone determine which among many binders is the "best". Although binding affinity will be the main benchmarking criterion, it is relevant to incorporate other properties important for drug discovery, including the calculated lipophilicity and novelty of the predicted structures, and also the experimentally-determined selectivity against one or more homologous "off-targets" (if this is called for in a challenge). Solubility of the compounds will be predicted, but also will be determined experimentally. Insoluble compounds will be flagged because precipitation is a confounder in nearly all binding assays.

CACHE will publish in advance the success criteria (activity, selectivity, aqueous solubility, lipophilicity, novelty, etc.) for each challenge, and how they will be combined into an overall multiobjective score [17,18], such as oralPhysChemScore (oPCS) [19] . CACHE will provide the workflows and scripts that were used to calculate the different descriptors. In one possible scheme (TABLE 1), active compounds will not be ranked *per se*, but rather will be classified into 3 buckets (green, yellow, red) by summing up the traffic light values for each property (column). The scoring scheme used to assess a compound's physical and molecular properties will be similar across the challenges, but the values for potency and selectivity may change depending on the challenge. For example, compounds with weaker affinity might be acceptable for hard targets with no precedent, but higher affinities might be the aim if the challenge is to identify novel chemotypes for precedented targets. As stated above, to facilitate comparison among methods, all predictions from all participants for a given target will be combined into a single small virtual library, and all participants will also be asked to rank these compounds.

Top-scoring molecules will be further analyzed by a panel of experienced medicinal chemists in order to provide additional annotation to the molecules, including opinion on the suitability of the hits to serve as a starting point for potential drug discovery programs. Their reflections will not influence the score, but rather will help contextualize the output.

Table 1. Example CACHE traffic light (TL) scoring scheme for one arbitrary target protein. Scoring schemes will differ for different challenges.

TL value	TL Binding affinity (measured)	TL Sw (measured)	TL logD @ pH 7.5 (measured)	TL MW/MW corr	TL PSA (Å^{2})	TL # rotatable bonds	TL Fsp3	Novelty*
0	<1µM	≥ 50	<3	≤ 400	≤ 120	≤ 7	>0.3	<0.6
1	10-1 µM	10-50	3-4	400-500	120-140	8-10	0.2-0.3	0.6 - 0.8
2	>10 µM	< 10	>4	> 500	> 140	≥ 11	<0.2	>0.8

TL = traffic light, Sw = solubility in water; Fsp3 = fraction of sp3 hybridized carbon atoms, calculated based on Murcko scaffolds
*Tanimoto distance relative to most similar structures binding that target in most recent version of ChEMBL using ECFP4 fingerprints as calculated from Rdkit (https://www.rdkit.org/)

SHARING THE CACHE OUTPUT

CACHE will generate three main outputs for the community: screening data, chemical structures, and the performance of algorithms. CACHE's mandate is making available the screening data and the chemical structures to the community without intellectual property or other restrictions on use, and in a digitally readable format. These data will also include the composition of the virtual libraries screened, all predicted small molecules including negative data, and all screening methods.

As a condition of participation, CACHE will mandate that predictors disclose their approaches in sufficient detail to enable an expert in the area to understand the methodology and algorithms. These methodology descriptions will be double-blind peer-reviewed by other participants to ensure they contain sufficient information according to standards of the field. In the interest of encouraging participation from all sectors, participants will not be required to provide access to their code and can remain anonymous. However, CACHE will encourage participants to share their code and, as stated below, intends to provide a range of financial incentives for those participants who open-source their algorithms and workflows, and ideally who also submit their fully automated workflows. In addition, participants must agree that those who submit top-performing methods (as determined by prespecified criteria agreed to by CACHE and the participants – see next section) will automatically be un-blinded. Participants that agree to share workflows, code and methodology must do so in a FAIR manner [20].

Table 2. CACHE output.

List of methods/strategies	Anonymized list of participants along with a description of their approach
Predicted structures from each participant, for each of two cycles	Experimentally determined and calculated properties for all predicted compounds (Table 1)
Performance of algorithms on common set of compounds	Create a virtual library that comprises predictions made by all participants in Cycle 1, and each participant will rank the compounds in that library
Set of top structures	Top ranked structures, including SAR if available
Crystal structures	Coordinates of all complexes of targets and predicted binders
Synthetic routes for top-ranked set of structures	Summary and primary data (yields)
Assay data (screening)	Primary screening data for all predictions and orthogonal confirmation data for active molecules
Quality control data for compounds	NMR, HPLC, MS, solubility

Participants will be encouraged to publish the results of their submissions and detailed analyses of their performance, and to work together to share learnings and identify differentiators of performance. CACHE will organize a workshop following each challenge and coordinate the publication of overview papers for each challenge, perhaps with dedicated special issues of relevant journals to provide a wider forum for participants.

HOW WILL CACHE AND THE CHALLENGES BE ORGANIZED AND MANAGED?

CACHE will be structured as an independent, not-for-profit entity, or fiscally governed by a not-for-profit organization with aligned goals, such as the Structural Genomics Consortium (SGC) [http://thesgc.org] or the Open Group (https://www.opengroup.org). CACHE or its parent organization will receive funding as described below and sub-contract other organizations (academic, government or industry) to carry out CACHE activities, and under terms that mandate data sharing. Industry could provide funding via donation, contract, or fee-for-service mechanisms. CACHE will create a Secretariat to handle administration, fundraising, project management, and logistics.

CACHE will comprise Members, who have the opportunity to influence the strategic directions of CACHE through a General Assembly. Members will include funders of CACHE as well as any participant that makes a meaningful contribution. The Membership will elect a Governing Board to be responsible for making operational decisions, including target selection, participation rules, and use of funds. An external Scientific Advisory Board will be appointed by the Governing Board to provide outside advice on scientific questions such as target choice and metrics for success.

CACHE plans to launch challenges for each of the five hit-finding scenarios shown in figure 2, each challenge comprising 3 different protein targets, and occurring over 2 years. There will be periodic public open calls for participation. For the first rounds, letters of intent will be solicited to better understand the needs and goals of potential participants. All potential participants would be asked to submit brief applications detailing their qualifications to participate and general intended approach. For inclusivity, the initiative should strive to accept every reasonable application, with due attention being paid to use resources efficiently.

For each challenge, CACHE will select a Challenge Lead who will be responsible for the coordination of experiments and logistics. The Challenge Lead will ensure best practices are used in challenge design, execution, and assessment, and codified in iteratively revised documents (such as living reviews or contributions to the NCATS Assay Guidance Manual). Challenge Leads will determine the details of specific challenges, and what compound properties – experimental or computed – beyond affinity for the target will be incorporated into the overall performance scores.

Challenge Leads will also be responsible for determining and executing or delegating the execution of appropriate “baseline methods” (such as random local search, simple similarity matching or vanilla docking methods where applicable) that will be run centrally (to avoid duplication for participants running many similar baselines). Challenge Leads will have the support of the Scientific Advisory Board in making all these decisions.

CACHE FUNDING STRATEGY – SHARING THE COSTS

CACHE intends that its activities, including governance, management, logistics, and data sharing, will be supported by a pool of government, industry, and charitable funders. Ideally, CACHE funding would also be used to provide subsidies for participants from resource-poor environments, providing an overall more inclusive approach.

The funding of the challenges themselves will be shared among interested funders and participants. Funders, such as a disease foundation or a company, could support challenges of particular interest to them. As CACHE matures, participants will be expected to help defray some portion of per-compound costs (including synthesis/purchase and assays) using their own funding. To facilitate this, CACHE will develop a transparent cost structure that can be used in funding applications. CACHE aspires to be able to subsidize the cost of participation for participants that agree to share their methods, code, or methodologies.

In the first round, CACHE aims to secure sufficient funding to purchase and assay ~100 compounds for every qualified participant to distribute among 3 targets. In subsequent rounds, these costs will be transferred to participants. We estimate the costs of experimental testing for 100 compounds is approximately USD$25,000; this includes purchasing of the compounds, quality control, protein purification, equipment time, primary biophysical assays, and hit confirmation using orthogonal assays. To facilitate logistics as well as to provide the opportunity to negotiate bulk pricing, CACHE will procure the compounds on behalf of all participants. If participants wish to test more than 100 compounds in the first rounds, or if the number of participants exceeds the initial available funding, participants may be required to fund some portion of per-compound costs.

CACHE will also be well-positioned to collaborate with other successful community initiatives in order to increase the impact of CACHE. For example, if CACHE includes a viral target among the challenges, then the CACHE predictions might input into community anti-viral development initiatives[†]. Predicted compounds that pose synthetic challenges can be turned into additional community challenges to design and predict the most efficient synthetic pathway for a given small molecule[‡]. Confirmed hits could also be used as starting points to develop new chemical probes[§].

SUMMARY AND NEXT STEPS

A group of scientists from the public and private sector intend to launch a benchmarking initiative to accelerate the development of computational methods to predict small molecules that bind to proteins. The initiative will comprise an experimental and data hub(s), which will support a community of predictors. All data, including chemical structures, will be made available without restriction on use. The initiative intends to attract funding from industry, governments, and foundations to support the infrastructure, and challenge-specific funding, in order to give disease-focused funders the opportunity to target the community-wide effort to proteins of interest to them. An organizational meeting is being arranged, with the intention to launch the first CACHE challenge in early 2022.

ACKNOWLEDGEMENTS

The Structural Genomics Consortium 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 US), Pfizer, Takeda and the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 875510. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Ontario Institute for Cancer Research, Royal Institution for the Advancement of Learning McGill University, Kungliga Tekniska Hoegskolan, Diamond Light Source Limited. This communication reflects the views of the authors and the JU is not liable for any use that may be made of the information contained herein.

M. K. Gilson acknowledges funding from National Institute of General Medical Sciences (GM061300). J. J. Irwin acknowledges funding from National Institute of General Medical Sciences (GM133836). These findings are solely of the authors and do not necessarily represent the views of the NIH.

[†] https://postera.ai/moonshot

[‡] Merck Compound Synthesis Challenge (http://compoundchallenge.merckgroup.com)

[§] https://www.thesgc.org/chemical-probes

The Lee laboratory at the University of Cambridge receives funding from multiple sources, including Pfizer, AstraZeneca, the Engineering and Physical Sciences Research Council and the Winton Programme for the Physics of Sustainability.

T.I. Oprea and C. G. Bologa acknowledge funding from the National Institutes of Health Common Fund program, Illuminating the Druggable Genome (CA224370 and TR002278).

COMPETING INTERESTS

M. K. Gilson has an equity interest in and is a cofounder and scientific advisor of VeraChem LLC. J. J. Irwin is a co-founder of Blue Dolphin LLC, which undertakes fee-for-service ligand discovery. A.A. Lee is the chief scientific officer and a shareholder of PostEra Inc. T. I. Oprea has received honoraria from or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth, and is on the Scientific Advisory Board of ChemDiv and InSilico Medicine.

REFERENCES

1 Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii-v, doi:10.1002/prot.340230303 (1995).

2 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

3 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754 (2021).

4 Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein-ligand poses and affinity rankings. *J Comput Aided Mol Des* **33**, 1-18, doi:10.1007/s10822-018-0180-4 (2019).

5 Parks, C. D. *et al.* D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des* **34**, 99-119, doi:10.1007/s10822-020-00289-y (2020).

6 Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des* **32**, 1-20, doi:10.1007/s10822-017-0088-4 (2018).

7 Jansen, J. M., Cornell, W., Tseng, Y. J. & Amaro, R. E. Teach-Discover-Treat (TDT): collaborative computational drug discovery for neglected diseases. *J Mol Graph Model* **38**, 360-362, doi:10.1016/j.jmgm.2012.07.007 (2012).

8 Jansen, J. M., Amaro, R. E., Cornell, W., Tseng, Y. J. & Walters, W. P. Computational chemistry and drug discovery: a call to action. *Future Med Chem* **4**, 1893-1896, doi:10.4155/fmc.12.137 (2012).

9 Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des* **30**, 651-668, doi:10.1007/s10822-016-9946-8 (2016).

10 Yin, J. *et al.* Overview of the SAMPL5 host-guest challenge: Are we doing better? *J Comput Aided Mol Des* **31**, 1-19, doi:10.1007/s10822-016-9974-4 (2017).

11 Bannan, C. C. *et al.* Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge. *J Comput Aided Mol Des* **30**, 927-944, doi:10.1007/s10822-016-9954-8 (2016).

12 Xiong, Z. *et al.* Crowdsourced identification of multi-target kinase inhibitors for RET- and TAU- based disease: The Multi-Targeting Drug DREAM Challenge. *PLoS Comput Biol* **17**, e1009302, doi:10.1371/journal.pcbi.1009302 (2021).

13 Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224-229, doi:10.1038/s41586-019-0917-9 (2019).

14 Irwin, J. J. *et al.* ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J Chem Inf Model* **60**, 6065-6073, doi:10.1021/acs.jcim.0c00675 (2020).

15 von Delft, F. *et al.* A white-knuckle ride of open COVID drug discovery. *Nature* **594**, 330-332, doi:10.1038/d41586-021-01571-1 (2021).

16 Edwards, A. M., Bountra, C., Kerr, D. J. & Willson, T. M. Open access chemical and clinical probes to support drug discovery. *Nat Chem Biol* **5**, 436-440, doi:10.1038/nchembio0709-436 (2009).

17 Wager, T. T., Hou, X., Verhoest, P. R. & Villalobos, A. Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery. *ACS Chem Neurosci* **7**, 767-775, doi:10.1021/acschemneuro.6b00029 (2016).

18 Cummins, D. J. & Bell, M. A. Integrating Everything: The Molecule Selection Toolkit, a System for Compound Prioritization in Drug Discovery. *J Med Chem* **59**, 6999-7010, doi:10.1021/acs.jmedchem.5b01338 (2016).

19 Lobell, M. *et al.* In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* **1**, 1229-1236, doi:10.1002/cmdc.200600168 (2006).

20 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).