HHS Public Access

Author manuscript

Nat Rev Chem. Author manuscript.

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**[5,17], **Claudia R. Gordijo**[1], **Gerhard Hessler**[18], **Alexander Hillisch**[19,†], **Anders Hogner**[20], **John J. Irwin**[21], **Johanna M. Jansen**[22], **Daniel Kuhn**[23], **Andrew R. Leach**[13,14], **Alpha A. Lee**[24,25], **Uta Lessel**[26], **Maxwell R. Morgan**[1], **John Moult**[27,28], **Ingo Muegge**[29], **Tudor I. Oprea**[10,30], **Benjamin G. Perry**[31], **Patrick Riley**[32], **Sophie A.L. Rousseaux**[7], **Kumar Singh Saikatendu**[33], **Vijayaratnam Santhakumar**[1], **Matthieu Schapira**[1,3], **Cora Scholten**[34], **Matthew H. Todd**[35], **Masoud Vedadi**[1,3], **Andrea Volkamer**[36], **Timothy M. Willson**[37]

[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

[†]Co-corresponding authors: aled.edwards@utoronto.ca and alexander.hillisch@bayer.com.

[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]Sanofi-Aventis Deutschland GmbH, R&D, Integrated Drug Discovery, Frankfurt am Main, Germany

[19]Bayer AG, Pharmaceuticals, Research and Development, Wuppertal, Germany

[20]Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

[21]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA

[22]Novartis Institutes for BioMedical Research, Emeryville, CA, USA

[23]Merck Healthcare KGaA, Computational Chemistry & Biologics, Darmstadt, Germany

[24]PostEra Inc, San Franciso, CA, USA

[25]Department of Physics, University of Cambridge, Cambridge, UK

[26]Boehringer Ingelheim Pharma GmbH & Co. KG, Medicinal Chemistry, Biberach an der Riss, Germany

[27]Institute for Bioscience and Biotechnology Research, Rockville, MD, USA

[28]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA

[29]Alkermes, Inc., Waltham, MA, USA

[30]University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, USA

[31]Drugs for Neglected Diseases initiative, Geneva, Switzerland

[32]Relay Therapeutics, Boston, MA, USA

[33]Global Research Externalization, Takeda California, Inc., San Diego, CA, USA

[34]Bayer AG, Pharmaceuticals, Open Innovation – Public Private Partnerships, Berlin, Germany

[35]School of Pharmacy, University College London, London, UK

[36]In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité Universitätsmedizin Berlin, Berlin, Germany

[37]Structural Genomics Consortium, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

One aspirational goal of computational chemistry is to predict potent and drug-like binders for any protein, such that only those that bind are synthesized. In this Roadmap, we describe the launch of

Critical Assessment of Computational Hit-finding Experiments (CACHE), a public benchmarking project to compare and improve small molecule hit-finding algorithms through cycles of prediction and experimental testing. Participants will predict small molecule binders for new and biologically relevant protein targets representing different prediction scenarios. Predicted compounds will be tested rigorously in an experimental hub, and all predicted binders as well as all experimental screening data, including the chemical structures of experimentally tested compounds, will be made publicly available, and not subject to any intellectual property restrictions. The ability of a range of computational approaches to find novel binders will be evaluated, compared, and openly published. CACHE will launch 3 new benchmarking exercises every year. The outcomes will be better prediction methods, new small molecule binders for target proteins of importance for fundamental biology or drug discovery, and a major technological step towards achieving the goal of Target 2035, a global initiative to identify pharmacological probes for all human proteins.

Introduction

Computational hit-finding is poised to make a major impact in early drug discovery[1–4], enabled by leaps in computational power, increased accessibility to diverse chemical space, improved physics-based methods, and the emerging potential of newer machine learning and artificial intelligence (AI) approaches. However, despite the promise, no algorithm can currently select, design, or rank potent, drug-like small molecule protein binders consistently.

Significant advances in the development of 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[5] (https://predictioncenter.org/) was launched, humans were often better at predicting protein structures than were computational methods. Now machine learning algorithms can predict the structures of many (but not all) globular proteins as accurately as can be determined experimentally[6,7], and progress is being made rapidly to predict the structures of protein complexes[8,9].

In computational chemistry, organizing benchmarking exercises similar to CASP have occurred[10–18], but none are currently operational. In addition, besides the TDT and DREAM benchmarking initiatives[13,14,18] that included a prospective arm to its prediction challenge, there has been no concerted effort to provide experimental testing of predictions, which is in large part because of the associated costs. There is no opportunity to fund the synthesis and quality control of predicted compounds and to test their binding rigorously in one laboratory under standardized conditions that facilitate head-to-head comparison of predictions. One confounding issue has been that commercial sensitivities 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 AI methods. These companies are also likely reluctant to share their methods in any detail, or see them put to the test publicly.

It is now possible to conceptualize a benchmarking exercise that can overcome some of these limitations. From a financial perspective, the creation of ultra-large libraries of chemicals that can be described in silico and procured on-demand[2,19] significantly reduces the cost associated with accessing chemical matter to test predictions. The availability of massive amounts of computational resources facilitates data sharing and democratizes the ability to make predictions.[20]

From an organizational view, there is now community acceptance 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 accepted as an accelerator of biomedical science.[21,22] Critically, this paradigm has provided immense scientific value by normalising the placement of chemical matter, including advanced molecules such as chemical probes, in the public domain without complex and rate-limiting intellectual property agreements[21].

Based on this new landscape, we are creating a public-private partnership called Critical Assessment of Computational Hit-finding Experiments (CACHE) to benchmark computational approaches for the identification of a small molecule that binds a targeted protein with high enough affinity and suitable physiochemical properties to qualify as a credible starting point for a drug discovery project. Modelled after CASP, CACHE will organize hit-finding challenges against selected biologically-relevant 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 will have an experimental arm testing predictions prospectively. Each challenge will typically include two testing iterations to enable refinement and forward application of successful predictive models. Upon completion of a hit-finding challenge, all data generated by CACHE, including all screening data and chemical structures, will be publicly available without intellectual property restrictions.

The genesis of the CACHE concept

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 in November 2020 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 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: target selection and prioritization; virtual library construction; measuring outcomes; and governance. These groups’ ideas for the CACHE project are presented in this Roadmap.

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 generated predictions, and solicit predictions for hit compounds. For evaluation, CACHE will purchase or otherwise procure the compounds that are predicted to bind, experimentally measure their binding to their intended target, calculate other key properties of the active compounds, and share the outcomes openly with the scientific community (FIG. 1). We envision that CACHE, like CASP, will organize multiple rounds of challenges, providing on-going opportunities for computational scientists, molecular modelers, algorithm developers, etc. to improve and test their methods.

CACHE challenges & target selection

CACHE will organize hit-finding challenges that represent the common scenarios encountered in hit-finding (FIG. 2B). The CACHE Target Selection Committee will select targets appropriate for each of these five scenarios. They will define the acceptance criteria for targets in each scenario and use bioinformatics tools to compile a long-list of targets that meet these criteria. Subsequently, they will create a mechanism(s) for the community, including the funders of CACHE, to prioritize from this list of potential targets those that will be included in the benchmarking challenges.

Only targets having two orthogonal cost-effective direct binding assays that can provide rapid, validated, high-quality experimental feedback will be considered. From this list, CACHE and its funders will use a prioritization scheme that maximizes both the structural diversity of the target proteins and the opportunity to discover new biological insights. The aim is for CACHE to benefit both the computational and pharmaceutical communities. We anticipate that a funder (such as a disease-focused charity) might 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. 2A). We also imagine that, in lieu of providing direct financial support, funders, foundations, or companies might also offer in-kind support for CACHE, for example, by offering to evaluate all predictions for a given target, or provide access to computational resources, assay reagents, and/ or laboratory equipment. Over a 5-year period, we aspire to provide CACHE with the resources to pursue 15 targets, representing each of the 5 hit-finding scenarios to enable it to fulfill its goals.

Participation guidance and support

Virtual compound libraries availability

To enable rapid and cost-effective testing of predictions, CACHE will establish a well-defined and robust core make-on-demand virtual library comprising compounds that are readily accessible from commercial vendors, at reasonable cost. A combination of Enamine REAL (now providing 21 billion make-on-demand compounds) and ZINC20[19] (containing over 750 million purchaseable compounds) might comprise the core of this library.

CACHE will annotate compounds in the library with predicted physical properties, such as cLogP, polar surface area (PSA), and the fraction of sp3 carbon atoms (Fsp3), among others, which will be assessed in the challenge's success criteria. Ideally, these annotated properties should enable participants to select individual subsets and/or apply relevant filtering as they see best fit for their challenge, 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 this classification may be required to account for the needs of specific CACHE participants. For example, a 1% diversity set or a 10% diversity set might be preferred when examining computationally intensive approaches, etc. The libraries will evolve, such that more compounds will be added as they become commercially available or accessible, and additional library subsets will be created as a function of their performance.

To accommodate *de novo* design methods, which are not selecting compounds from commercial vendors but designing new molecules, CACHE will test custom-synthesized compounds if the compounds can be procured by participants within 3–6 months of the completion of the *in silico* selection step. In later challenges, CACHE may also incrementally explore mechanisms to provide participants access to a virtual library containing new chemistry, where synthetic chemists within academia or industry would be offered the opportunity to contribute to a virtual library that covers new chemical space. In this initiative, chemists would add compounds that they would be willing to synthesize on-demand in a timely manner, using emerging synthetic chemistry protocols and their own resources.

At regular and defined intervals over the course of the CACHE benchmarking exercises, the CACHE Virtual Libraries Committee 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.

Evaluating predictions experimentally

At the core of the CACHE initiative will be an experimental hub that will provide rapid, high-quality testing of the predicted hits. Predicted compounds will be submitted to the experimental hub, which will procure the compounds and evaluate 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 will be re-tested in dose-response mode, as well as in an orthogonal biophysical assay, which is critical for the robustness of the experimental results. Feedback will be given first to the participant(s), and participants that made successful predictions will have the opportunity to improve on them by submitting a new set of predictions.

Each CACHE challenge round will take ~18 months, with two cycles of predictions per round in order to give participants the opportunity to incorporate learnings from the first round into their next 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.

CACHE benchmarking

Benchmarking computational hit-finding 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 binder among many is the best. The affinity of compounds that are active in a primary screen, typically in a surface plasmon resonance (SPR) assay, will be evaluated with an orthogonal biophysical method. Although binding affinity to the desired protein will be the main benchmarking criterion, selectivity against specific off-targets will be tested if called for in the challenge. The solubility and colloidal aggregation[23] of hit molecules will be determined experimentally by dynamic light scattering. Insoluble and aggregating compounds will be flagged because precipitation and aggregation are confounders in nearly all binding assays. Common pan-assay interference (PAINS) compounds[24], predicted for instance by a strong indication of promiscuity with Badapple[25] will also be flagged. Method specific patterns of binding or inhibition that could be associated with non-specific interaction or aggregation will also be monitored. These include high Hill Slopes of IC_{50} determination plots, linear fitting of SPR data and unreasonable stabilization of proteins measured by differential scanning fluorimetry (DSF). Experimental hits will also be subjected to rigorous analytical quality control to confirm the purity of the samples. CACHE will seek to solve the crystal structure of validated hits in complex with their target when robust crystallization protocols are available.

Before each challenge, CACHE will publish the corresponding success criteria (activity, selectivity, aqueous solubility, lipophilicity, novelty, etc.), and how these will be combined into an overall multi-objective score[26,27] similar to the oralPhysChemScore (oPCS)[28]. Binding affinity, aqueous solubility and logD will be measured. Calculated properties include: corrected molecular weight[28]; PSA[29]; number of rotatable bonds; Fsp^3 [30]; and novelty. This novelty parameter will be defined as the Tanimoto distance relative to most similar structures binding that target as calculated from RDKit. These novelty thresholds were chosen based on previous work with circular fingerprints[18,31]. 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. 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 targets that are more difficult to identify hits against and have no reported 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 the 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 (TABLE 1) 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. This includes human experience on reactivity, synthesizability, chemical stability, potential toxicity, off target activity, etc. Their reflections will not influence the score, but rather will help contextualize the output and provide insight for refinement of the scoring process for future challenge iterations.

CACHE output sharing

CACHE will generate three main outputs for the community: screening data, chemical structures, and algorithm performance (BOX 1). CACHE's mandate is to ensure the screening data and the chemical structures are available to the community without intellectual property or other restrictions on use, and in a digitally readable format according to FAIR principles[32]. These data will also include the composition of the virtual libraries screened, all predicted small molecules (including negative data), all experimental screening results and all screening methods.

CACHE will mandate that participants disclose their computational 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 the 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 software code and, as stated below, intends to provide a range of financial incentives for those participants who release their code, algorithms and workflows under permissive open-source license terms, and ideally who also submit their fully automated workflows. In addition, participants must agree that the identity of those who submit top-performing methods (as determined by prespecified criteria agreed to by CACHE and the participants) will automatically be de-anonymized when the screening data and compound structures are publicly released. Participants that agree to share workflows, code and methodology must do so in a FAIR manner[32].

Participants will be encouraged to seek peer-reviewed open-access publication of 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 open-access publication of overview papers for each challenge, perhaps with dedicated special issues of relevant journals to provide a wider forum for participants.

CACHE organization and management

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) or the Open Group. 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, all under terms that mandate open data sharing. CACHE will create a Secretariat to handle administration, fundraising, project management, and logistics.

CACHE will be funded in part by Members, who will have the opportunity to influence the strategic directions of CACHE through appointments to a Governing Board (FIG. 4). The Governing Board will 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 the strategy for target selection and the metrics for success.

CACHE plans to launch challenges for each of the five hit-finding scenarios shown in FIG. 2, each challenge occurring at least once over 2 years (FIG. 3). 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, paying attention to use resources efficiently.

For each challenge, CACHE will contribute 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. For instance, these documents could be similar to the living reviews found in the Living Journal of Computational Molecular Science or made as contributions to the NCATS Assay Guidance Manual. Challenge Leads, in consultation with the Governing Board, 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 to be run centrally to avoid duplication for participants running many similar baselines. These methods would likely include random local search, simple similarity matching or vanilla docking methods where applicable. Challenge Leads will have the support of the Scientific Advisory Board in making all of these decisions.

CACHE funding strategy

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, could support challenges of particular interest to them. As CACHE matures, participants will be expected to pay a participation fee reflective of a portion of per-compound costs (including synthesis/procurement and assays). To facilitate this, CACHE will develop a transparent cost structure for each challenge. In the interest of encouraging transparency, CACHE aspires to be able to subsidize the cost of participation for participants that agree to share their methods, code, or methodologies.

By centralizing the experimentation, CACHE will not only provide standardized data but also will provide logistical and cost savings over carrying out the activities in individual labs. Within CACHE, we estimate the costs of rigorous experimental testing for 100 compounds is approximately $25,000 USD; this includes purchasing of the compounds, quality control, protein purification, equipment time, primary biophysical assays, and hit confirmation using orthogonal assays. CACHE will procure the compounds on behalf of all participants to facilitate logistics as well as to provide the opportunity to negotiate bulk pricing.

In the first two competitions, CACHE aims to secure sufficient seed funding to purchase and evaluate ~100 compounds for every qualified participant, but in subsequent rounds, these costs will be transferred to participants. If participants wish to test more than 100 compounds, or if the number of participants exceeds the initial available funding, participants may also 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, such as the COVID Moonshot Initiative[20]. Predicted compounds that pose synthetic challenges can be turned into additional community challenges, such as Merck's Compound Synthesis Challenge, 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.

CACHE success criteria

CACHE will be a long-term project that will be assessed against success metrics of organizational capabilities and community engagement in the short-term (1–3 years), and scientific accomplishments in the longer term (year 3 and beyond). Organizational success will be achieved by running the entire workflow of target selection for several rounds. For example, we expect 6 rounds to run over ~2 years; where a round includes hit prediction, chemical synthesis, biochemical/biophysical testing of the compounds and analysis/dissemination of the results (FIG. 3). Community engagement success will be defined as generating a constant flow of targets, hit proposals and experimental results from an increasing number of community members over time. Scientific success can likely be analysed only after 12 rounds (year 4), after which all five types of challenges are performed at least two to three times with different targets. Scientific success metrics will include providing unbiased comparisons of which computational methods deliver suitable hits (chemotypes) as starting points for drug discovery, and the number and quality of novel chemical matter for biologically interesting new targets.

With respect to quantitative metrics, we aspire for CACHE to have deposited experimental screening data for 12 proteins and 30,000 drug-like molecules selected by over 100 participants in the public domain after 4 years. Over this period, we also expect that computational methods will predict unprecedented hits for 25% of the nominated novel targets. We also expect CACHE to provide clearer guidance as to which computational

approaches are most promising for identifying novel small molecules active substances and thus significantly influence computational hit-finding method development on a global scale.

Summary and next steps

A group of ~50 scientists from the public and private sectors 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 experimental and data hub(s), which will support a community of participants in their predictions. 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 enable a community-wide effort to target proteins of interest to them. The intention is to launch the first CACHE challenge in early 2022.

Acknowledgements

We are grateful to Ali Hazrat (Karolinska Institutet, Stockholm, Sweden) for developing the CACHE website: www.cache-challenge.org. Advise on chemoinformatic aspects from Hans Briem (Bayer AG, Berlin, Germany) is gratefully acknowledged. 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). J. D. Chodera acknowledges funding from the National Institute of General Medical Sciences (R01GM124270) and the National Cancer Institute (P30CA008748). These findings are solely of the authors and do not necessarily represent the views of the NIH.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

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).

R. A. Batey acknowledges funding from the Natural Science and Engineering Research Council (NSERC) of Canada.

B.G. Perry acknowledges the following donors for contributing to DNDi's overall mission: UK aid, UK; Médecins Sans Frontières, International; and the Swiss Agency for Development and Cooperation (SDC), Switzerland.

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. Hillisch is on the Board of Directors of the SGC (Structural Genomics Consortium) and the Scientific Advisory Board of Cresset. 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. J. D. Chodera is a current member of the Scientific Advisory Boards for OpenEye Scientific Software, Redesign Science, Interline Therapeutics, and Ventus Therapeutics, and holds equity interests in Redesign Science and Interline Therapeutics. B. G. Perry is on the Board of Directors of Evolia Therapeutics S.A. and the scientific advisory board of Spirochem A.G. All remaining authors declare no competing interests.

Glossary

Hit-finding
identification of a small molecule that binds a target protein, and that has high enough affinity and suitable physiochemical properties to qualify as a credible starting point for a drug discovery project

Chemical probes
chemical compounds used as tools to study the biological function of proteins

cLogP
calculated partition coefficient of a chemical compound between water and 1-octanol

Polar surface area
surface sum over all polar atoms (namely oxygen, nitrogen, phosphor and polar hydrogen) in a chemical compound

Chemical space
ensemble of all possible chemical compounds adhering to a given set of principles and boundary conditions, for drug like small molecules estimated to be 10^60 compounds

Experimental hub
platform where predicted compounds are tested experimentally

Surface plasmon resonance
Label-free method that can be used to measure the binding of a small molecule to a protein immobilized on a chip

Dynamic light scattering
method that can be used to measure the solubility or aggregation of a molecules in solution

Pan-assay interference (PAINS) compounds
chemical compounds often giving false positive results in high-throughput screens as they interact nonspecifically with numerous biological molecules

Differential scanning fluorimetry
experimental method to measure protein unfolding by monitory changes in fluorescence as a function of temperature

oralPhysChemScore
combined score based on certain molecular properties, roughly estimating the suitability of a compound as lead structure for an orally administered drug

Corrected molecular weight

surrogate parameter for molecular volume, correcting the molecular weight of molecules containing halogen atoms

Tanimoto distance
statistic used for gauging the similarity and diversity of sample compound sets

Extended connectivity fingerprints
a certain type of circular fingerprint designed for similarity searches within sets of small molecules

Circular fingerprints
fingerprints representing molecular structures by means of circular atom neighborhoods

References

1. Gorgulla C et al. An open-source drug discovery platform enables ultra-large virtual screens. Nature 580, 663–668, doi:10.1038/s41586-020-2117-z (2020).
2. Lyu J et al. Ultra-large library docking for discovering new chemotypes. Nature 566, 224–229, doi:10.1038/s41586-019-0917-9 (2019).
3. Walters WP & Wang R New Trends in Virtual Screening. J Chem Inf Model 60, 4109–4111, doi:10.1021/acs.jcim.0c01009 (2020).
4. Grebner C et al. Virtual Screening in the Cloud: How Big Is Big Enough? J Chem Inf Model 60, 4274–4282, doi:10.1021/acs.jcim.9b00779 (2020).
5. Moult J, Pedersen JT, Judson R & Fidelis K A large-scale experiment to assess protein structure prediction methods. Proteins 23, ii–v, doi:10.1002/prot.340230303 (1995).
6. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589, doi:10.1038/s41586-021-03819-2 (2021).
7. 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).
8. Evans R et al. Protein complex prediction with AlphaFold-Multimer. Preprint at https://doi.org/10.1101/2021.10.04.463034 (2021).
9. Humphreys IR et al. Computed structures of core eukaryotic protein complexes. Science 374, eabm4805, doi:10.1126/science.abm4805 (2021).
10. 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).
11. Parks CD 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).
12. 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).
13. Jansen JM, Cornell W, Tseng YJ & Amaro RE 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).
14. Jansen JM, Amaro RE, Cornell W, Tseng YJ & Walters WP Computational chemistry and drug discovery: a call to action. Future Med Chem 4, 1893–1896, doi:10.4155/fmc.12.137 (2012).
15. 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).
16. 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).

17. Bannan CC 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).
18. 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).
19. Irwin JJ 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).
20. 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).
21. Edwards AM, Bountra C, Kerr DJ & Willson TM Open access chemical and clinical probes to support drug discovery. Nat Chem Biol 5, 436–440, doi:10.1038/nchembio0709-436 (2009).
22. Müller S et al. Target 2035 – update on the quest for a probe for every protein. RSC Medicinal Chemistry, doi:10.1039/D1MD00228G (2022).
23. McGovern SL, Helfand BT, Feng B & Shoichet BK A specific mechanism of nonspecific inhibition. J Med Chem 46, 4265–4272, doi:10.1021/jm030266r (2003).
24. Baell JB & Nissink JWM Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. ACS Chemical Biology 13, 36–44, doi:10.1021/acschembio.7b00903 (2018).
25. Yang JJ et al. Badapple: promiscuity patterns from noisy evidence. Journal of Cheminformatics 8, 29, doi:10.1186/s13321-016-0137-3 (2016).
26. Wager TT, Hou X, Verhoest PR & Villalobos A Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery. ACS Chem Neurosci 7, 767–775, doi:10.1021/acschemneuro.6b00029 (2016).
27. Cummins DJ & Bell MA 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).
28. 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).
29. Ertl P, Rohde B & Selzer P Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43, 3714–3717, doi:10.1021/jm000942e (2000).
30. Lovering F, Bikker J & Humblet C Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem 52, 6752–6756, doi:10.1021/jm901241e (2009).
31. Muchmore SW et al. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model 48, 941–948, doi:10.1021/ci7004498 (2008).
32. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, doi:10.1038/sdata.2016.18 (2016).

Box 1.

CACHE output

List of methods and strategies

Anonymized list of participants along with a description of their approach.

Predicted structures from each participant

Experimentally determined and calculated properties for all predicted compounds (Table 1), for each of two cycles.

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 using their algorithm.

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 of experimental methods and primary data (yields, purities, etc).

Assay data (screening)

Primary screening data for all predictions and orthogonal confirmation data for active molecules.

Quality control data for compounds

NMR, high-performance liquid chromatography (HPLC), mass spectrometry (MS), solubility.

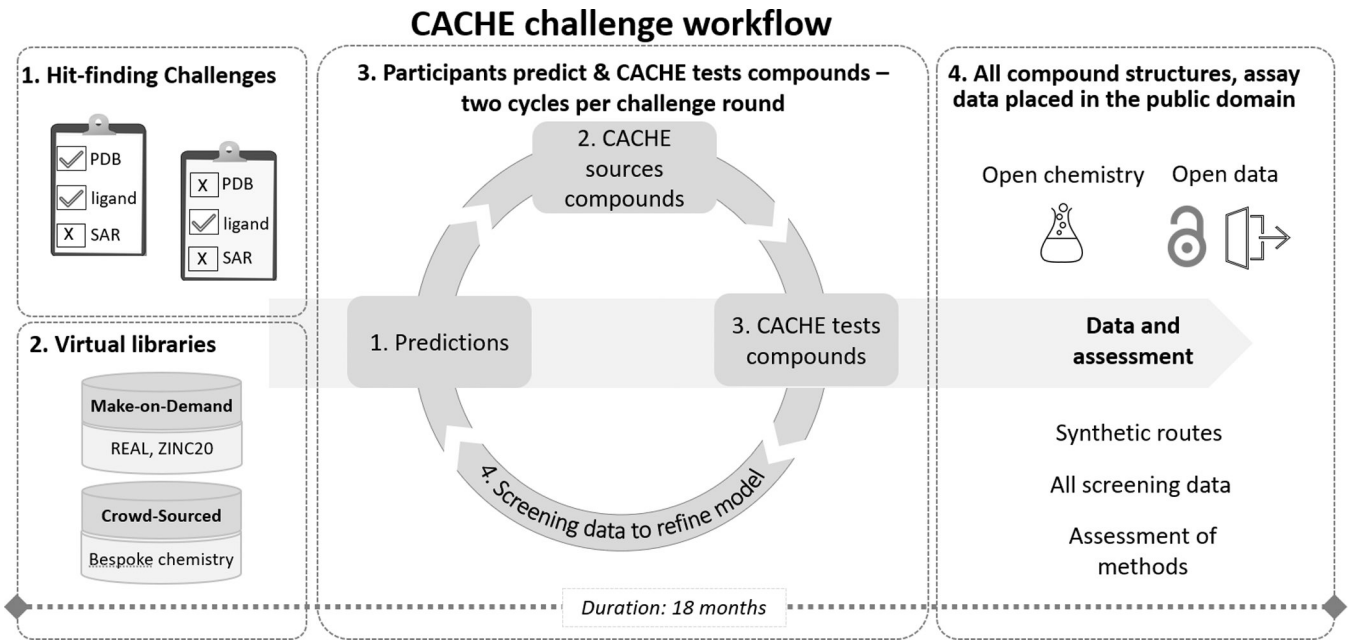

Fig. 1. CACHE challenge workflow.

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.

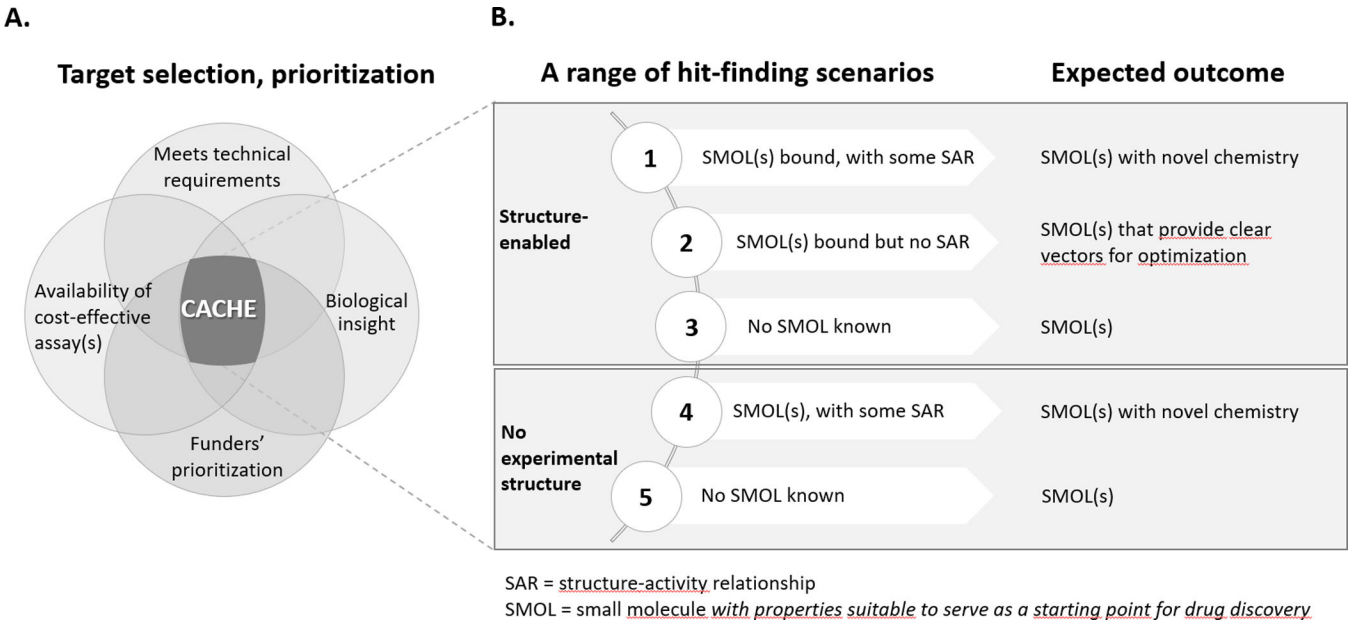

Fig. 2. Target selection consideration and classes of CACHE challenges.

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.

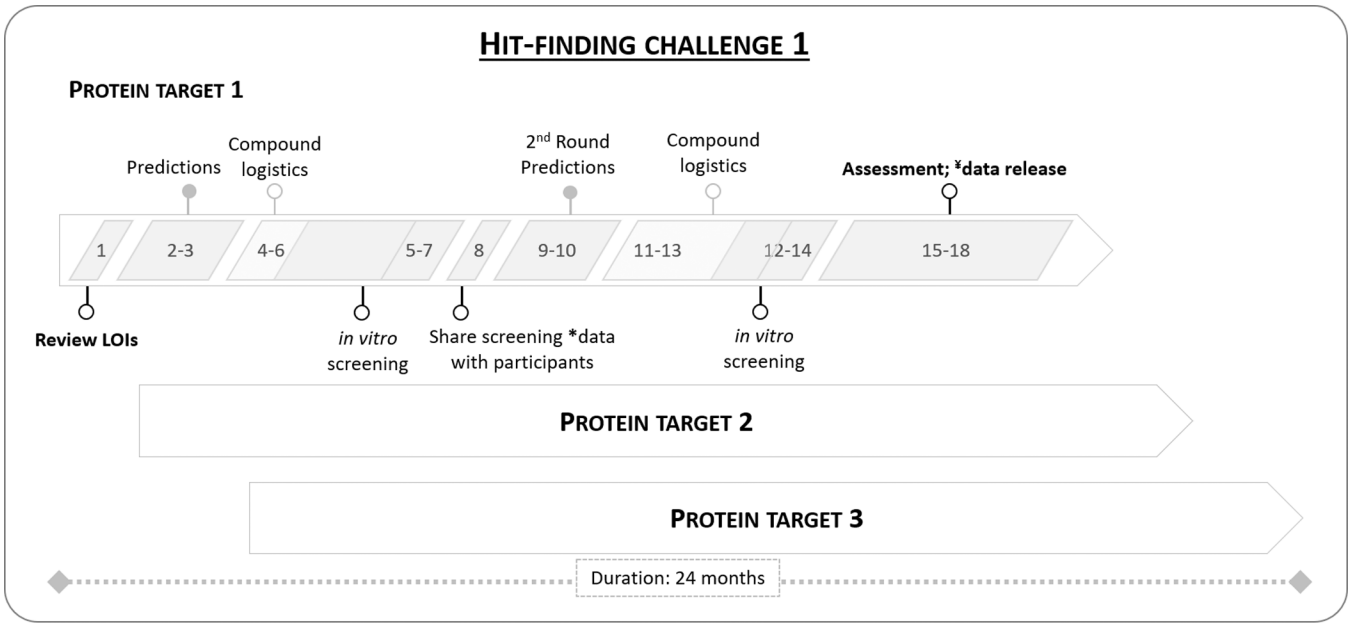

Fig. 3. The timelines of challenge activities.

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

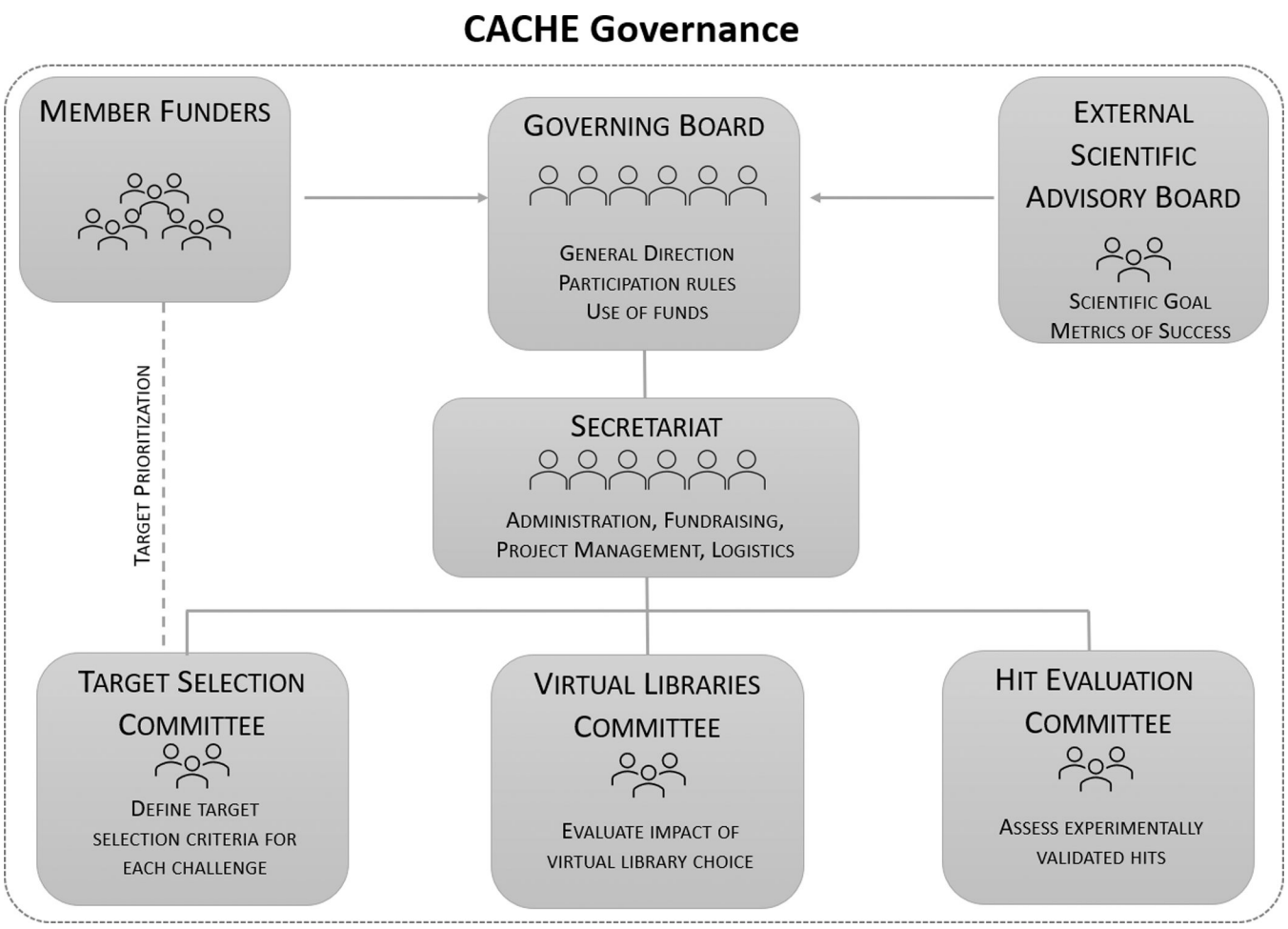

Fig. 4. CACHE governance.

CACHE will be structured as an independent, not-for-profit entity. The CACHE governance will include a Governing Board constituted by funders (Members) and two independent Members selected with input from the scientific community; an External Scientific Advisory Board; and a Secretariat that will oversee day-to-day operations. The Governing Board will create three scientific committees: the Target Selection Committee will select protein targets (with final decision impacted by the Governing Board); the Virtual Libraries Committee will define the virtual chemistry libraries to be screened; and the Hit Evaluation Committee will create the metrics of success and assess performance against the metrics. Funders that do not wish to play an active role in governance can nominate targets for consideration by the Target Selection Committee.

Table 1.

Example CACHE traffic light (TL) scoring scheme for one arbitrary target protein.

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	<0.4
1	10–1 μM	10–50	3–4	400–500	120–140	8–10	0.2– 0.3	0.4 – 0.6
2	>10 μM	< 10	>4	> 500	> 140	≥ 11	<0.2	>0.6

TL = traffic light, Sw = solubility in water, Fsp^3 = fraction of sp^3 hybridized carbon atoms, calculated based on Murcko scaffolds

*Measured experimentally.

†Tanimoto distance relative to most similar structures binding that target as calculated from RDKit.