Topic idea submitted to IHI - Reference Number: TI_001209

Are you submitting the idea:
☐ **in your personal capacity?**
☒ **on behalf of an organisation?**

Please indicate the name of the group organisation: Structural Genomics Consortium and Target 2035 project

Please select from the list below the type of stakeholders your organization represents:
Charity/foundation

1 Title of your idea

Please provide a short title that accurately reflects the objective(s) of your idea:
Harnessing data science and artificial intelligence to scale and democratize small molecule drug discovery

2 Scope

Explain the specific challenges/problems to be addressed by your idea and how these affect relevant stakeholders, taking into account what is already known and/or available in the field:
The open availability of drug starting points (chemical probes) for all human proteins would remove a major roadblock in developing medicines for all diseases, and is encapsulated in the Target 2035 mission (target2035.net). However, at a cost of $2M per chemical probe and at a rate of 6–7 proteins per year, the current technologies are not scalable across all potential drug targets, and are restricted to well-funded institutions and companies. Our aim is to develop computational methods that will deliver drug molecules with high efficiency and at a fraction of the cost of current methods, thus expanding the number of scientists, institutions and jurisdictions that can participate in drug discovery. Our approach is modelled after that which enabled AlphaFold to democratize structural biology and ChatGPT to democratize natural language processing. Specifically, our proposed public-private partnership (comprising large pharma, biotech and tech SME's, contract research organizations, universities, and patient foundations) will create an open public dataset of the interaction of billions of compounds with thousands of human proteins in a machine-learning readable format, and use this resource to train, benchmark and test artificial intelligence (AI) models that can predict new protein/small molecule interactions. We will provide open access to the data, to the drug starting points, and to any AI algorithms developed in the project. The project aims to dramatically increase the number of druggable proteins, lower the price for discovery of clinical candidates, and support a more open, robust and equitable ecosystem for drug discovery.

Please indicate which IHI specific objective(s) (SO), as described in the IHI Strategic Research and Innovation Agenda (SRIA), your idea addresses:

["SO2: integrate fragmented health research and innovation efforts bringing together health industry sectors and other stakeholders, focusing on unmet public health needs, to enable the development of tools, data, platforms, technologies and processes for improved prediction, prevention, interception, diagnosis, treat- ment and management of diseases, meeting the needs of end-users"]

Please select the keywords that are most relevant to your idea:

["Non-communicable diseases"
"Cardiovascular diseases"
"Immune system diseases"
"Metabolic diseases"
"Neurodegenerative diseases"
"Oncology"
"Rare diseases"
"Paediatric"
"Treatment"
"Disease management"
"Digital health"]

In alignment with the IHI specific objective(s) selected above, specify the objectives of your idea:

Many recent advances in AI were made possible by the existence of extensive, freely available datasets to train computational models. AI-driven small molecule drug discovery, by contrast, has access only to chemical bioactivity datasets that are fragmented across databases such as ChEMBL and PubChem, not accessible in a ML readable format, and do not contain high-quality data about inactive compounds. We will bridge this gap by:

1. Fostering a collaborative ecosystem comprising pharma, tech companies, academia and patient foundations
2. Systematically generating and depositing >100 TB of high-quality machine-readable ligand-protein interaction positive and negative data in a publicly available platform enabling data mining and modeling (AIRCHECK: aircheck.ai).
3. Developing predictive AI models from these data
4. Benchmarking the models through multiple open competitions (challenges) linked to experimental testing of predictions at clusters of university and industry labs.
5. Providing open access to drug starting points, data, models, platforms, technologies and processes for improved treatment and management of diseases.

3 Expected impacts to be achieved by your idea

Briefly describe the expected impacts to be achieved by your idea, ensuring that they contribute to IHI general and relevant specific objectives, as described in the IHI SRIA:

Impacts *are wider long-term effects on society (including the environment), the economy and science, enabled by the outcomes of R&I investments. Impacts generally occur sometime after the end of the project, e.g. successful implementation of digital solutions supporting people-centred care.*

IHI general objectives: *1. contribute towards the creation of an EU-wide health research and innovation ecosystem that facilitates translation of scientific knowledge into innovations, notably by launching at least 30 large-scale, cross-sectoral projects, focussing on health innovations; 2. foster the development of safe, effective, people-centred and cost-effective innovations that respond to strategic unmet public health needs, by exhibiting, in at least 5 examples, the feasibility of integrating health care products or services, with demonstrated suitability for uptake by health care systems. The related projects should address the prevention, diagnosis, treatment and/or management of diseases affecting the EU population, including contribution to 'Europe's Beating Cancer Plan'; 3. drive cross-sectoral health innovation for a globally competitive European health industry and contribute to reaching the objectives of the new Industrial Strategy for Europe and the Pharmaceutical Strategy for Europe.*

We will create a large public-private ecosystem comprising clusters that generate analysis-ready protein-ligand data for >1,000 human proteins, organize benchmarking, and test predictions in dedicated experimental laboratories.

By prioritizing protein targets of interest to patient groups and developing innovations that reduce the cost of drug discovery, the project will be people-centred and applicable to all diseases affecting the EU population.

By integrating pharma, tech companies, SME's, and academia, the project will drive cross-sectoral innovation across the ecosystem, and contribute to the objectives of the Industrial and Pharmaceutical Strategies for Europe, particularly advancing digital transformation.

The project will have immediate scientific impact by providing open drug starting points for proteins of relevance to common and rare diseases; immediate economic impact by enabling small biotechnology and AI companies to benchmark and market their algorithms and technologies; and longer-term scientific and societal transformative impact by lowering the price for discovery of hits and ultimately clinical candidates.

4 Why should your idea become an IHI call topic?

Explain why collaboration through a cross-sectoral and multidisciplinary public private partnership is needed in particular:

Why does it require collaboration among several industry sectors (*e.g.* pharma, vaccines, biotech, medical devices, in vitro diagnostics, radiotherapy, medical imaging health ICT)?

Why does it require collaboration between private (industry) and public partners (*e.g.* academia, healthcare practitioners, patients, regulators)?

The project describes a complex ecosystem that involves a range of organizational and technical capabilities distributed in academia, foundations and industry. Activities include:

- creating a Target List and producing purified proteins (patient foundations, academia, biotech and pharma)
- generating high-quality, protein-ligand data using a range of technologies (academia, biotech, tech companies and pharma)
- creating, managing and enabling equitable access to an open dataset, using cloud resources and data science (academia and tech companies)
- organizing benchmarking competitions (academia, pharma and non-profits)
- computational predictions (academia, public, tech companies, biotech and pharma)
- accessing compounds from "make-on-demand" or bespoke libraries, and managing

compound logistics (contract research organizations, pharma and academia)
- carrying out experimental testing of predictions (contract research organizations, pharma and academia)

Why is the contribution of industry needed to achieve the expected impacts?

***Contribution of industry**: Large companies that are members of the IHI industry partners (i.e. COCIR, EFPIA, EuropaBio, MedTech Europe, Vaccines Europe) contribute to the programme, primarily through 'in-kind' contributions (e.g. their researchers' time, laboratories, data, compounds). At least 45% of each project's total costs have to be in-kind contribution.*

The concept for this idea emerged from a meeting hosted by the University of Frankfurt in September 2023. The meeting, which reviewed results from pilot projects, was attended by representatives from >20 pharma and tech companies, and several patient foundations. The conclusions from the meeting that a foundational project of this scale (>100 TB of high-quality data) and complexity (protein science, chemistry, benchmarking, machine learning, data science) was essential, but could not be accomplished without intellectual, monetary, and experimental contributions from academia and a range of private sectors. The assembled group, as well as representatives from other companies and patient foundations, are actively preparing a project plan that will be published as a co-authored white paper.

Industry contributions that were contemplated included funding, in-kind experimental testing of predictions, management time, ML researcher time, screening capacity, compound handling, database support, and protein purification. Foundation partner contributions offered included direct funding, target prioritization, and community engagement.