Joint DDLS advisory committee (DAC) report 2023

Based on the virtual site visit and materials provided to support it by DAC members Søren Brunak, Florian Jug and Jan Ellenberg on June 21, 2023.

The Swedish Data Driven Life Science (DDLS) program is a 12-year research initiative funded by the Knut and Alice Wallenberg Foundation (KAW) with a total budget of about 300 MEUR. SciLifeLab coordinates the program in close collaboration with ten participating universities (Chalmers, GU, KI, KTH, LiU, LU, SLU, SU, UmU and UU) and the Swedish Museum of Natural History. Its focus is to develop and apply data driven approaches in four life science areas, namely cell and molecular biology, precision medicine and diagnostics, evolution and biodiversity as well as epidemiology and biology of infection.

During its 12 year run-time from 2021-2032, the DDLS program aims to pursue six strategic objective (see DDLS strategy, v2), i.e. (i) **attract** scientific excellence in DDLS to Sweden by recruiting a cohort of almost 40 new principal investigators, (ii) **train** the next generation of DDL Scientists by educating over 500 new PhD students and postdocs; (iii) provide DDLS **services** by developing new tools and approaches to empower the Swedish research community in FAIR data management, exploiting the opportunities provided by responsible use of artificial intelligence, and start to create dynamic and four-dimensional models of life; (iv) **bridge** the life and data science communities; (v) form partnerships and create **societal impact**; and (vi) engage in **policy** actions.

The DDLS program is clearly structured and has proposed a roadmap for concrete deliverables in each of the six strategic objectives that are to be achieved in the five phases of its overall run time. Despite starting at the end of the pandemic and the significant coordination challenges with a distributed group of partners, the DDLS program has achieved a lot in its first almost two years of the three-year ramp up phase 1. The main activities have been: to put an organizational structure and team in place, start the faculty recruitment jointly with all participating universities leading to appointment of half the DDLS faculty, prepare the research school and start the general training programs with conferences and symposia, start to provide data science services via SciLifeLab's Data Centre, NBIS/WABI service platforms and establish research area data science nodes at four partner universities, and start to work towards partnerships with complementary Swedish and international partners, including the WASP program, the WCMM and SciLifeLab fellows, industry, as well as EMBL.

Overall, the DAC is very impressed by how much has been achieved in less than two years for such an ambitious, large scale and complex program and congratulates the DDLS management team on their excellent progress. Clearly, this could not have been achieved without the excellent teams, national coordination and collaboration mechanisms already in place in SciLifeLab and the trusted partner for cutting edge life science infrastructure services SciLifeLab has become in its first decade for the national stakeholders. Similarly, DDLS can build on the international excellence of SciLifeLab in data producing life science that develops and uses the latest technologies, which has provided the foundations to make Swedish life science data-driven across the board.

The DDLS program thus rests on a solid foundation and is very timely. When conceived in 2019, it set a leading example of forward-looking life science nationally and internationally. The DAC congratulates the funder KAW for their foresight to engage in this direction and shares their view that this is a flagship program for which they should have the highest expectations and provide generous support for its success. DDLS will be strategically important for the future development of SciLifeLab and Swedish life science, as more and more of it will become data driven and predictive in the future, propelled by the fortuitous cycle between new disruptive technologies that provide increasingly time and space resolved quantitative data on living systems and the ability to computationally and theoretically use this data to make predictive dynamic

models of healthy life, adaptation to environmental change and disease. This will provide critical knowledge so the society can take facts-based decisions to respond to major future challenges.

While the fast and efficient ramp-up and the well-organized 12-year plan is very impressive, moving at this speed to implement a plan made in 2019 also creates some challenges, as there is probably no other area that is changing so quickly in life science as the use of data science and computation in it, as exemplified by the launch of the alpha-fold protein structure prediction tool in the meantime and the recent hot debate regarding "run-away AI". Therefore, this is a critical moment to take stock of what has been achieved so far, where potential gaps are, and what new developments and opportunities have emerged since 2019, where the next phase investments could be targeted to achieve the highest possible impact. In our view, it is key to do that, if the ambition is to maintain an internationally leading, strategic and forward-looking research programme.

During the virtual review meeting, the DAC therefore discussed and probed several, from our point of view, key aspects of how to move DDLS into the future successfully and remain internationally leading. In the following, we would like to make some general strategic, as well as some more specific, recommendations, to support DDLS on its very impressive upward trajectory.

1. Remain leading and avoid early "lock-in"

DDLS is one of the most dynamic areas of life science and changes at a very fast pace. By recruiting all PIs in only two recruiting rounds at the beginning of the program, there is a certain risk to "lock in" the program at a very high level of excellence, yet prioritise already somewhat established research areas and therefore potentially missing out on some of the most impactful areas that are only starting to emerge now or in the near future.

We advise to avoid going only for standard track records of excellence and rapid returns by "picking the low hanging fruits", but rather also remain flexible to incorporate high-risk and high gain emerging areas in DDLS at the PI level. This will be key to achieve the desired societal impact, especially in environment and medicine, where real data driven theoretical work is only starting to get traction. We would therefore advise to:

- Coordinate remaining PI recruitments to fill gaps, slow down recruitment until that is achieved, consider a third round of recruitment for newly emerging high-risk areas, in this very fast-paced and competitive field.
- Engage the already appointed faculty in identifying forward looking topics in a DDLS faculty meeting and encourage the universities to include existing PIs in the recruitment committees of the next round(s).
- Make strategic choices especially for fellows recruited to industry. While Sweden does not have national big pharma companies, it is traditionally strong for startups around new technologies and WALP provides additional measures to boost this area. The use of data science and AI in life science has a lot of potential for new concepts in bioeconomy and appointments should ideally be made in a way that realises this potential on the strong basis of the larger DDLS community.
- Don't fall behind and then run after the AI revolution, but have the ambition to lead its responsible and explanatory use in the life science, where the goal is not only prediction, but mechanistic understanding (see also point 6. below).

2. Form a truly integrated and collaborative DDLS student community

The DDLS predoc community is still to be built and has the potential to become the bottom-up engine to drive the ground-breaking data driven research and new and collaborative ideas between the PIs. Also, it presents a great opportunity for Sweden and the partner universities in terms of attracting extraordinary student quality in a highly competitive sector (computer science/AI) and piloting the new training courses for the next generation of scientists. We would therefore advise to:

- Seize the opportunity and create a truly international and inter-institutional PhD program that recruits in an integrated fashion to attract the best talent and achieve a level of quality and interdisciplinarity than cannot be achieved by any single university.
- Form a collaborative class of each recruitment year, by physically joint introductory and afterwards potentially annual training course that bonds the students as friends and future colleagues. This

can include soft-skill training that students really need (e.g. research project management, scientific writing, presentation skills, postdoc application training workshops on ethics in (data) science and AI) , that can then be adopted by the universities later for the community at large.
- We strongly recommend to underpin the training related suggestions we mention here by an adequate budget allocation.
- Include external DDLS faculty on the thesis advisory committees of predocs, to broaden the scope of the scientific outlook and naturally drive collaborations between DDLS PIs.

3. Use the postdoc programme to link the DDLS groups collaboratively and with the Swedish life science community at large.

Similar to the predoc community, the postdoc community that will now be established is poised to be an invaluable asset for DDLS to create new opportunities for ground-breaking interdisciplinary science. Conceptually, interlinking and co-mentoring them in a similar way to the students, with adjusted measures that fit this more senior career stage has huge potential. To realize it, we recommend to:

- Create postdoc fellowships for collaborative work between two or more DDLS fellows
- Use similar postdoc fellowships to promote joint work with the SciLifeLab and WCMM fellows as well as with WASP, following a similar model as within DDLS and with industry.
- Resource the training schools adequately and provide added value soft skill pilot courses for DDLS postdocs (e.g. lab management & leadership, grant applications), that can then be adopted by the universities later for the community at large.
-

4. Take full advantage of the ability of SciLifeLab to support, integrate and coordinate collaborative work

The rapid ramp-up and success of DDLS would not have been possible without the professional management, data science competence, and integrating and collaborative excellence at SciLifeLab. That the DDLS programme reaches its full potential will depend on continuing to strengthen the coordinating role of SciLifeLab and enter into a long-term synergistic relationship between technology driven data generation and using the data to drive the next generation of life science. We therefore recommend to:

- Establish a national competence and knowledge network by an integrated and staged pre/postdoctoral training programme, whose needs are defined by DDLS PIs and then scale the courses to the national community with the University as desired.
- Provide novel training content for a new generation of interdisciplinary DDL scientists, using the data science tools and services rolled out by the Data Centre and WABI. Topics for courses could, for example, be research software development in the leading software environments and workflow management systems that allow individual tools to be combined in multimodal and integrative data workflows.
- Such courses would empower the national DDLS user community to take full advantage of the new tools and approaches and scale the impact of the programme truly nationally.
- Stimulate and coordinate internationally excellent collaborative science, by promoting interactions between DDLS fellows and with their life science colleagues in SciLifeLab and WCMM through funding for collaborative projects
- Develop services for the new quality of next generation data that is increasingly spatial, temporal, phenotypic, medical and environmental to allow the integration and metadata standardization that will be needed by the new DDLS community to enable cross domain collaborations.

5. Ensure impact by propelling basic and applied data driven life science to the next level

DDLS has enormous potential but, due to its distributed nature, is threatened by fragmentation, dilution of effort and covering too many areas with too little critical mass. To achieve maximum success and impact, we recommend to:

- DDLS is a large player in environmental biology and a comparatively small player in precison medicine and diagnostics. Nevertheless, it should aim to achieve a new quality of interdisciplinary

DDLS research by allowing the PIs to work collaborate by integrating their students, postdocs, tools and methods and dedicating structuring measures and resources towards such joint efforts.

- Engage in policy to enable clinical and health data analysis in Sweden for research, as soon as possible. Since the fellows are starting and need access to this NOW, a pragmatic interim measure would be to use the openly available Danish and Finnish health data sets and use this to accelerate the policy process in Sweden. Everything should be done to avoid that Sweden puts itself in the position to have the leading DDLS health scientists in Europe but does not allow them to work with Swedish health data.
- Lead the discussion and provide the tools for responsible use of AI in life science research, for example by integrating structure prediction and analysis of experimental data.
- Ensure that not only in research but also in support, service and software development the highly competed for data science experts have attractive career tracks in the Swedish research landscape, so the best staff can be recruited and retained.

6. Stay at the cutting edge of AI technology revolution

Due to the disruptive speed of AI technology developments (e.g. large language models such as ChatGPT), there is a danger of falling behind and running after the latest trends in AI, rather than leading the way for its future development and responsible use in research. In our view, the AI methods that will allow to learn directly from the new quality of data that we are increasingly producing on living systems, how they function, with molecular and physical explanations are yet to be developed. To stay leading in this very fast moving field we recommend to:

- ensure effective collaboration with the WASP programme (see also above, point 3.)
- make sure you create a maximally attractive environment for leading AI development applied to the fundamental societal challenges in life science
- use the fact that AI is one of the common interests of virtually all DDLS PIs and their respective teams to engage them in (i) creating and participating in cutting edge training events where new methods or tools are being looked at, (ii) participating in hybrid seminars or round-tables where internationally renowned AI experts are invited to talk on a variety of AI topics from theory to ethics, but also (iii) organizing hackathons and training events where the DDLS community can come together and learn from each other and from invited international guests how to conduct cutting-edge AI research most efficiently.