



Evaluation of Mathematics, ICT and Technology (EVALMIT) 2023-2024

Self-assessment for research groups

Date of dispatch: 19 September 2023

Deadline for submission: 31 January 2024

Institution (name and short name): University of Oslo (UiO)

Administrative unit (name and short name): Department of Informatics (IFI)

Research group (name and short name): Scientific Computing and Machine Learning (SCML)

Date: 31.1.2024

Contact person: Ole Christian Lingjærde

Contact details (email): ole@uio.no

Content

1. Organisation and strategy	3
1.1 Research group's organisation	3
1.2 Research group's strategy	4
1.3 Relevance to the institutions	7
1.4 Research group's resources	7
1.5 Research group's infrastructures	8
1.6 Research group's cooperations.....	9
2. Research quality	10
2.1 Research group's scientific quality	10
2.2 Research group's societal contribution	20
3. Challenges and opportunities	22

1. Organisation and strategy

1.1 Research group's organisation

With a staff of more than 400, The UiO Department of Informatics (IFI) is a powerhouse for informatics training and research in Northern Europe. It is the largest such unit in Norway and the one with the strongest international rankings. Around a comprehensive core in computer science, IFI interprets informatics in substantial breadth: from microelectronics and signal processing to participatory design and digital innovation. In 2023, IFI hosts five BSc and eight MSc study programmes, with close to 2500 students. The department is organized in four thematic sections, each of which comprises 2–5 research groups. Our group, Scientific Computing and Machine Learning (SCML) is one out of five research groups in the Section for Machine Learning.

SCML was established in 2001 under the name Biomedical Informatics (BMI) to meet a growing demand for expertise in machine learning and computation in the life sciences. The group was a partner of Center for Bioinformatics at UiO in the period 2018-2023, with Professor Eivind Hovig as the center's director. The group recently dissolved the formal partnership with Center for Bioinformatics and changed its name to SCML to reflect an increased focus on foundational research in scientific computing and machine learning. Since its establishment in 2001, the group has steadily increased in size to meet a growing demand for expertise in machine learning and artificial intelligence in the life sciences and elsewhere in society.

In October 2022, the group comprised the following members:

- **Professors:** Torbjørn Rognes, Geir Kjetil Sandve, Eivind Hovig, Ole Christian Lingjærde and Knut Liestøl (emeritus)
- **Adjunct Professors:** Are Magnus Bruaset, Xing Cai, Rolf Inge Skotheim, Aslak Tveito
- **Adjunct Associate Professor:** Jon Kristen Lærdahl
- **Senior Engineer:** Morten Johansen (50%)
- **Researcher:** Morten Martingsdal (20%)
- **Postdoctoral Fellows:** Mostafa Alwash, Ivar Grytten, Milena Pavlovic, Knut Dagestad Rand, Enrico Riccardi, Christian Bope, Paula Istvan
- **PhD Fellows:** Ghadi al Hajj, Maria Perona Fjeldstad, Sumana Kalyanasundaram, Atle Eskeland Rimehaug, Lonneke Scheffer

Since then, SCML has expanded further to include a new professor (Xing Cai, previously an adjunct professor in the group) and a new associate professor (Pooya Zakeri).

The group's research is to a large extent project based with clearly defined goals, timelines and outcomes, some external funding, and often involving formal collaboration with partners from other groups at the department or interdisciplinary collaboration with partners from other departments or outside the university. However, we also maintain and encourage exploratory and innovative research activities with less strict timelines and primarily or solely supported by core funding. These activities frequently lead over time to the identification of tasks that are suitable for future project applications, the initiation of larger projects, and the establishment of new collaborations with other groups.

Professor Torbjørn Rognes was group manager until September 2023; the current group manager is Professor Ole Christian Lingjærde.

Table 1. List of number of personnel by categories

	Position by category	No. of researcher per category	Share of women per category (%)	No. of researchers who are part of multiple (other) research groups at the admin unit	No. of temporary positions
No. of Personnel by position	Professor (1013)	4	0%	0	0
	Researcher (1109)	1	0%	0	1
	Postdoctoral Fellow (1352)	7	14.3%	0	7
	PhD Fellow (1017)	5	60%	0	5
	Senior Engineer (50%) (1181)	1	0%	0	1
	Adjunct Professor (1013)	4	0%	0	4
	Adjunct Associate Professor (1011)	1	0%	0	1

1.2 Research group's strategy

a) Describe the research group's main goals, objectives and strategies to obtain these (e.g. funding, plans for recruitment, internationalisation etc.) within the period 2012-2022.

The group's main goals for 2012-2022 can be summarised as follows:

- Expand our national and international collaborative network
- Become a leading national actor in development of computational methods for life science
- Contribute to solve important real-world problems in immunology and cancer
- Increase student recruitment

Below is a brief resume of the four goals and our specific and actionable objectives to achieve them:

Expand collaborative network: Expanding our network of national and international research collaborators would provide access to and increase our own expertise in specific areas, and would help identify potential partners for future research projects. Our strategy has been to invite ourselves to other groups of interest and to utilise existing contacts to reach our more effectively to new groups of interest.

Data analysis methodology: We aimed to become a leading national developer in academia of computational methods for data analysis and visualization in the life sciences. We chose to specifically focus on two areas where we already had high competence: (1) compute demanding applications and (2) statistical problems of high dimensionality. To this end, we chose to focus much of our research around developing methods and applications for the analysis of omics data such as DNAseq data, RNAseq data, proteome data and immune repertoires.

Solve real-world problems: To maximise the societal impact of our research, we aimed to deliver significant contributions to the solution of real-world problems in immunology and cancer. Utilising our strong collaborative networks in Norway and abroad to groups working within immunology and cancer biology, we joined several international research projects and gained access to data, resulting in a substantial number of high profile publications where we contributed with expertise in machine learning, sophisticated mathematical modeling and high performance computing.

Student recruitment: We aimed to raise the visibility and interest at all study levels in studies of computational data analysis in the life sciences, thereby contributing to meet the growing demand in the society for such expertise. To this end, we became involved in the period in the development, planning and teaching of several courses at the bachelor, master and PhD level targeting areas of relevance to computational data analysis.

b) Please describe the benchmark of the research group. The benchmark for the research group should be written by the administrative unit in collaboration with the research group. The benchmark can be a reference to an academic level of performance (national or international) or to the group's contributions to other institutional or sectoral purposes.

Benchmark 1: The group should have a substantial scientific production with at least five times as many international publications per year as there are permanent staff members, and should play a central role in a significant proportion of the papers.

Benchmark 2: The group should participate in at least twice as many grant applications every year as there are permanent staff members, and receive funding from at least some of these applications.

Benchmark 3: The group should attract a substantial number of candidates for MSc and PhD projects and each permanent staff member should play an active role in supervision of both MSc students and PhD students.

c) Describe the research group contribution to education (master's degree and/or PhD).

SCML contributes continuously to the department's education at all study levels. The group has for many years played a key role in the planning and teaching of the two introductory courses in programming at the Faculty of Mathematics and Natural Sciences, which combined are taken by more than 1000 students every year from several departments (Informatics, Mathematics, Physics, Chemistry). Courses where SCML has designed and been main responsible for the teaching include:

- | | |
|------------------|---|
| • INF1000/IN1000 | <i>Introduction to object oriented programming</i> |
| • IN1900 | <i>Introduction to programming with scientific applications</i> |
| • INF-MAT2351 | <i>Numerical calculations</i> |
| • IN3200/4200 | <i>High performance computing and numerical projects</i> |

- INF3380 *Parallel programming for scientific problems*
- IN4030/INF4350 *Introduction to bioinformatics*
- GEO4070 *C++ and C programming*
- IN-BIOS5000/9000 *Genome sequence technologies, assembly, variant calling*
- IN5270 *Numerical methods for partial differential equations*
- INF5380/9380 *High performance computing in bioinformatics*
- INF5640 *Advanced numerical methods for partial differential equations*

Courses where SCML has contributed significantly to the teaching include:

- IN3050 *Introduction to artificial intelligence and machine learning*
- IN3130/INF4130 *Algorithms: design and efficiency*
- BIOS4010/MBV4010 *Methods in molecular biology and biochemistry*
- MBV-INF4410/9410 *Bioinformatics for molecular biology*
- INF5620 *Numerical methods for partial differential equations*
- INF-BIO9120 *High Throughput Sequencing technologies and bioinformatics*
- MF9210 *Laboratory course in molecular biology*
- BIOS9905 *Phylogenomics*

SCML recruits master students from the BSc programs *Mathematics with Informatics* (MAMI) and *Informatics: Programming and System Architecture* (PROSA), and is responsible for supervising master students from the study programs *Computational Science* (study option: Bioinformatics), *Data Science* (study option: Data Science and Life Science) and *Informatics: Programming and System Architecture*. The average production of master degrees in the period 2018-2022 was 5 per year, while the total number of PhD degrees in the same period was 14.

d) Describe the support the host institution provides to the research group (i.e., research infrastructure, access to databases, administrative support etc.).

The host institution provides practical technical support for our research activity in the form of acquiring technical equipment; installing software; solving technical issues; and general technical guidance. It also provides administrative support for our research activity in the form of information and guidance about funding opportunities; helping with formalities when setting up research applications; library access to major scientific journals; financial management related to budgeting, accounting, grant management, expense tracking and reporting; ensuring that the research group is regularly and frequently made aware of other activities at the institution that may be of relevance; leadership training; and practical assistance with personell issues of any kind (sick leaves, etc). The host institution also provides support for our teaching activity by taking responsibility for most administrative matters and offering assistance in virtually all matters not directly related to our own lecturing or preparations of such. In addition to the above, the host institution provides access to several computing resources in addition to the desktop computers offered to all employees.

1.3 Relevance to the institutions

Describe the role of the research group within the organisation. Consider the group's contribution towards the institutional strategies and objectives and relate the group's benchmark to these.

The department's strategic plan for 2019-2021 states that there is a major need for the development of competence environments and for the education of candidates with competence and experience in ML/AI. This is an area where the SCML group has great research expertise. We have also participated in the teaching of machine learning and artificial intelligence at the bachelor level at the department. We are participating in the committee currently discussing the establishment of a new bachelor program in machine learning and AI, and will also participate in the planning and teaching of new courses to be offered in ML/AI in the coming years. SCML is also actively involved in the interdisciplinary dScience center where SCML member Sandve is a member of the scientific advisory committee.

UiO's strategic plan for 2021-2030 emphasises the development of the life sciences and academic convergence. Modern life science is highly data driven and generates substantial amounts of data that needs to be analysed and interpreted. Scientific computing, machine learning and artificial intelligence plays a major role in this endeavour. SCML has for many years specialised in developing algorithms and tools for this purpose and has long experience with interdisciplinary collaborations and convergence projects. The strategic plan also mentions further development of the collaboration between UiO and the university hospitals, an area where SCML has been highly active in all the years since the group was established in 2001 with shared position, a wide network, participation in the center of excellence Center for Cancer Biomedicine (CCB) in 2007-2017, and participation in a range of other joint initiatives.

1.4 Research group's resources

Describe the funding portfolio of the research group the last five years (2018-2022).

The group has obtained funding in the period 2018-2022 for research projects coordinated by SCML as well as from projects coordinated by other partners. Funding sources include the Research Council of Norway (**Doctor AI2**: an IKTPLUSS project coordinated by SCML; **A platform for development of peptide-based antimicrobials for treatment of infections with drug-resistant bacteria**: a "Nyskapende bioteknologi" project coordinated by Oslo University Hospital and with SCML as a partner), the Kristian Gerhard Jebsen Foundation (**Codirc**: a KG Jebsen centre with the Faculty of Medicine at UiO as host and with a co-PI from SCML), Health Region South East (**Characterizing in time and space the metastatic process of hormone receptor positive breast cancer**: a research project coordinated by OUS and with a co-PI from SCML), the Norwegian Cancer Society (**Large scale proteomic analysis in cancer-associated chronic fatigue - a search for biomarkers relevant to pathogenesis, diagnosis and treatment**: a research project coordinated by Oslo University Hospital and with SCML as a partner), the Suurph program (**PhD funding for two students**: Lonneke Scheffer and Solveig Næss), and sources at the UiO outside the department (**RealArt**: a UiO:Life Science convergence environment where SCML has a co-PI; **ImmunoLingo**: a UiO Life Science convergence environment where SCML has a co-PI; **Centre for computational inference and evolutionary life science (CELS)**: a strategic research initiative at the Faculty of Mathematics and Natural Sciences).

Table 2. Describe the sources of R&D funding for the research group in the period 2018-2022.

	2018 (kNOK)	2019 (kNOK)	2020 (kNOK)	2021 (kNOK)	2022 (kNOK)
Basic funding	5 330	5 089	6 630	7 254	6 651
Funding from industry and other private sector sources			670	670	670
Commissioned research for public sector					
Research Council of Norway	670	670	1 340	1 774	1 647
Grant funding from other national sources	670	1 340	2 100	2 100	2 100
International funding e.g. NIH, NSF, EU framework programmes				1 340	1 340
Other					

1.5 Research group's infrastructures

Research infrastructures are facilities that provide resources and services for the research communities to conduct research and foster innovation in their fields. [These](#) include major equipment or sets of instruments, knowledge-related facilities such as collections, archives or scientific data infrastructures, computing systems communication networks. Include both internal and external infrastructures.

- 1) Describe which national infrastructures the research group manages or co-manages.

The SCML group has been responsible through its participation in Center for Bioinformatics for the management of the University of Oslo branch of ELIXIR Norway, a national node of the pan-European infrastructure ELIXIR for biological information. This infrastructure supports basic and translational research within the life sciences, and the national node is funded by the Research Council of Norway. Tasks performed by the UiO node include the development of solutions for handling and processing sensitive molecular data, offering services to the scientific community through the bioinformatics core facility at the Oslo University Hospital, and developing diagnostic tools for precision medicine.

- 2) Describe the most important research infrastructures used by the group.

The SCML group is a frequent user of large scale storage facilities, high-performance computing (HPC) resources, and servers. For open and restricted data (green and yellow level), we use high performance and data storage resources offered by Norwegian Research Infrastructure Services (NRIS) in the form of Saga (a computing cluster that can handle sequential and small scale parallel applications up to 128 threads) and the National Infrastructure for Research Data (NIRD) platform which offers storage and cloud services on the same infrastructure and has dedicated resources for data sharing and data intensive computing on a large amount of data. We also utilize the Norwegian Research and Education Cloud (NREC), a cloud service for research and education provided by UiO and UiB where users can provision virtual machines (servers) with network and storage. For restricted and in confidence data (yellow and red level) we also use the Fox high performance

computing cluster and the web portal Educloud Ondemand which allows software to run interactively on the Fox cluster. For strictly in confidence data (black level) such as molecular data from patient biopsies, we use Service for Sensitive Data (TSD) to collect, store, analyse and share the data in compliance with the Norwegian privacy regulations. This service is integrated with Colossus, a high performance computing cluster consisting of 64 compute nodes.

1.6 Research group's cooperations

Table 3. Reflect on the current interactions of the research group with other disciplines, non-academic stakeholders and the potential importance of these for the research (e.g. informing research questions, access to competence, data and infrastructure, broadening the perspectives, short/long-term relations).

Interdisciplinary (within and beyond the group)	SCML has always had strong interactions with other disciplines, and most notably with partners in the life sciences. Examples include participation in Centre for Computational Inference in Evolutionary Life Science (CELS), a multidisciplinary centre addressing fundamental biological questions related to the genotype-phenotype paradigm in an evolutionary context using a computational inference approach; participation in PharmaTox, a strategic research initiative aiming to generate novel insights on effects of pharmaceuticals on human neurotoxicity and neurodevelopment; participation in the K.G. Jebsen Coeliac Disease Research Centre (Codirc); participation in Centre for Bioinformatics (SBI); participation in the Norwegian Centre of Excellence Integreat; collaboration with brain scientists at the Division of Physiology at the Faculty of Medicine; frequent collaboration with statisticians at Department of Mathematics and the Faculty of Medicine, as well as with the Digital Signaling Processing and Image Analysis (DSB) group and other groups inside the department. These interactions are important for our own research in many ways: access to diverse types of data, access to domain knowledge, access to top expertise in various life science fields, and perhaps most importantly the emergence of novel ideas in the intersection between different research fields.
Collaboration with other research sectors e.g. higher education, research institutes, health trusts and industry.	SCML has a strong collaboration with several groups at Oslo University Hospital (OUS), including Department of Microbiology, Department of Medical Biochemistry, Department of Newborn screening, Department of Cancer Genetics, Department of Tumor Biology and Department of Immunology. Three professors in SCML (Rognes, Lingjærde and Hovig) all have adjunct positions at OUS.

<u>Transdisciplinary</u> (including non-academic stakeholders) <i>Transdisciplinary research involves the integration of knowledge from different science disciplines and (non-academic) stakeholder communities with the aim to help address complex societal challenges.</i>	SCML is also involved in transdisciplinary research activities that go beyond the boundaries of traditional interdisciplinary collaborations. Examples include <i>Linguistics-driven machine learning to decipher the molecular language of immunity</i> (ImmunoLingo) and <i>Real world – artificial worlds: improving causal inference in perinatal pharmaco-epidemiology using machine learning approaches on real-world and artificial data</i> (UiO:RealArt).
---	---

2. Research quality

2.1 Research group's scientific quality

Describe the research profile of the group and the activities that contribute to the research group's scientific quality. Consider how the group's work contributes to the wider research within the group's field nationally and internationally.

SCML has a large portfolio of research activities which methodologically can be broadly categorised as follows:

- **Machine learning:** Developing and refining machine learning methodology for analysis of high-throughput data in the life sciences. This includes developing tools for integration of large-scale data from multiple sources (multi-modality data);
- **Statistical modeling:** Developing evolutionary models for cancer initiation and progression. This combines aspects of scientific computing, mathematics, statistics, biology and genetics;
- **Biomarker development:** Developing prognostic and predictive biomarkers for cancer and immunology that helps understand disease mechanisms and predict treatment outcome, recurrence risk and survival;
- **High-performance computing:** Creating high-performance computing tools tailored for biological and medical research, facilitating faster and more accurate analysis;

Concrete examples of applications:

- **Microbiome bioinformatics:** SCML has made significant contributions to the field by developing high-performance tools for microbiome analysis that are widely used. Examples include VSEARCH (>7000 citations) and Swarm (>1300 citations).
- **Adaptive immunity:** SCML is involved in a broad interdisciplinary collaboration concerning the development of machine learning methodology to decipher how adaptive immune cells recognise their target. This has led to several grants from NFR and KG Jebsen and a number of papers providing insights into how immune signals can be used as a basis for diagnostics. Several recent papers provide methodological insights in machine learning and deep learning in particular and have been published in high profile international journals.
- **Cell-based modelling of cardiac electrophysiology:** SCML has been involved together with Simula Research Laboratory in pushing a new paradigm of cardiac modelling which is also applicable to modelling excitable tissues in the brain. Each cardiac cell is fully resolved in this modelling strategy. The unprecedented modelling resolutions is projected to provide new insights into

electrophysiology, but at the same time deliver a new grand-challenge problem in computational science.

- **Prognostic biomarkers in cancer:** we developed and later validated for the first time a whole genome copy number based prognostic biomarker for breast cancer (Complex Armwise Aberration Index; CAAI) in collaboration with Cold Spring Harbor Laboratory, University of Cambridge and Karolinska. We also developed and later validated for the first time a gene expression based predictor of histological transformation of follicular lymphoma to an aggressive phenotype.

SCML is frequently engaged in clinical studies, especially in the realm of cancer and autoimmune disease, where findings have a direct impact on patient care. The group commonly adopts an interdisciplinary approach in its research by integrating knowledge from biology, medicine, statistics, and computing to tackle complex research questions. Transdisciplinary collaborations are encouraged. SCML considers it essential to stay at the cutting edge of technology and methodology in order to innovate machine learning algorithms and computational tools of practical importance to the research society and to the society at large. As a part of this the group has recently started focusing on the convergence of machine learning and high-performance computing (HPC), with the aim to expedite physics-based simulations with the help of physics-informed machine learning and to use HPC techniques to improve the computational efficiency of machine learning libraries on both existing and upcoming hardware platforms. SCML puts large emphasis on dissemination of findings through publications, conference contributions, and to the larger society.

Please add a link to the research group's publication list at the research group's website:

<https://www.mn.uio.no/ifi/english/about/strategy/evalmit/scml.html>

Table 4. List of projects

Project 1: <i>IKTPLUSS project: Doctor AI2 – Artificial Intelligence mining of the Adaptive Immune system to develop an immunodiagnostics platform (2021-2024)</i>	Project owner(s) (project leaders organisation)	University of Oslo
	Total budget and share allocated to research group	Total budget 9.2 MNOK; share allocated to SCML 5.1 MNOK.
	Objectives and outcomes (planned or actual) and link to website	The overall vision of this project is to diagnose disease on the basis of a patient's immune repertoire. The adaptive immune system records each past and ongoing battle with disease by immune receptors - short genetic sequences specific for each disease. Previous attempts to do this have suffered from the lack of machine learning approaches that can exploit the unique biological characteristics of immune receptor repertoires, the lack of ground truth data for machine learning benchmarking, and the lack of a unified and comprehensive computational platform for machine learning analysis of large-scale immune receptor datasets. The objectives of this project were to develop novel AI

		methodology and implement a comprehensive software platform for immune receptor-based diagnostics, and to validate the framework using large experimental and synthetic immune receptor datasets. Website: https://sandvelab.org/research/
Project 2: <i>Centre of Excellence: Centre for Cancer Biomedicine (CCB) (2007-2017)</i>	Project owner(s) (project leaders organisation)	University of Oslo
	Total budget and share allocated to research group	Total budget in the period 2012-2022 approx. 60 MNOK; share allocated to SCML 7.2 MNOK
	Objectives and outcomes (planned or actual) and link to website	The overall vision of the centre was to unite basic and translational cancer research for the benefit of the patient. The objectives of the Statistical Analysis group (led by SCML professors Ole Christian Lingjærde and Knut Liestøl) were to develop methodology and software for analysis of omics data, to develop prognostic and predictive biomarkers in cancer by utilising data generated in the center, and to contribute with statistical and computational expertise in data driven projects initiated by other groups in the center. The many outcomes of the centre participation included (a) developing one of the first statistical models and algorithms (ASCAT) for deriving allele specific copy numbers from SNP data (and later DNA sequence data) - this algorithm is still widely used by the international scientific community and it became a part of the standard toolbox at the Wellcome Sanger Center; developing the first validated gene expression based predictor of transformation to aggressive disease in follicular lymphoma; (b) developing the first validated predictor of poor prognosis in breast cancer that was based on complex DNA arrangements (CAAI); (c) deriving a compact, biologically and clinically relevant representation of the genomic architecture in a tumor (CARMA). The center comprised five groups in addition to the one led by SCML, these focused on respectively cellular membrane dynamics, large scale genomic instability, genetics, lymphoma biology, and intracellular transport. Website: https://www.med.uio.no/ccb/english/index.html
Project 3:	Project owner(s) (project leaders organisation)	University of Oslo

K.G.Jebsen centre: Coeliac Disease Research Centre (Codirc) (2016-2022)	Total budget and share allocated to research group	Total budget 36 MNOK; share allocated to SCML approx. 4 MNOK
	Objectives and outcomes (planned or actual) and link to website	The primary objective of the centre was to translate recent huge advances in the basic understanding of the pathogenesis of coeliac disease to the clinic. Coeliac disease has several autoimmune characteristics, the most notable being highly disease specific autoantibodies. The disease has a strong heritable component with HLA genes as the major genetic factor and shares key features with HLA-associated diseases like primary sclerosing cholangitis, rheumatoid arthritis and multiple sclerosis. The center aimed to improve disease diagnostics, to implement novel treatments and to identify new therapeutic targets for this common disease. Website: https://www.med.uio.no/klinmed/english/research/centres/kgj-coeliac-disease/
Project 4: K.G.Jebsen Centre: Breast Cancer Research (2011-2015)	Project owner(s) (project leaders organisation)	University of Oslo/Oslo University Hospital
	Total budget and share allocated to research group	Total budget approx. 38 MNOK including all funding from K.G. Jebsen, Helse Sør-Øst, Helse Midt, Norwegian Research Council, Norwegian Cancer Society and EU. SCML supervised/co-supervised four PhD students at the centre.
	Objectives and outcomes (planned or actual) and link to website	The overall vision of the center was to open up for stronger collaboration between its clinical and basic scientists, and to motivate exchange of ideas on how to best utilize the huge collection of patient materials and data generated for the benefit of breast cancer patients. Specific objectives were to develop validated stratification criteria based on phenotypic/genotypic profiling of breast tumors, to use validated phenotypic and genotypic stratification criteria for assessing individual response and prognosis in patients with breast cancer, to identify molecular pathways and biomarkers predicting treatment response and/or resistance using cell lines and orthotopic xenograft models representing the various subgroups of breast cancer, and to translate and validate molecular and imaging biomarkers from preclinical models into clinical trials. The Bioinformatics/Biostatistics group (led by SCML

		professor Ole Christian Lingjærde) was involved in achieving all of these objectives and had extensive collaboration with the other five groups in the center (Clinical group, Molecular group, Micro-metastases group, Model systems and functional group, Metabolic profiling and imaging group). The center was a huge success and resulted in > 100 papers of which many were in high profile journals, as well as the establishment of long term international collaborations that remain to this day. Website: https://www.ous-research.no/kgjebesen/
Project 5: <i>Pan-Cancer Analysis of Whole Genomes (PCAWG)</i>	Project owner(s) (project leaders organisation)	International Cancer Genome Consortium (ICGC)
	Total budget and share allocated to research group	This project was largely funded by in-kind contributions from the participating member institutions
	Objectives and outcomes (planned or actual) and link to website	The PCAWG project is an international collaboration with the objective to identify recurrent patterns of somatic alterations in a large set of human cancer genomes (>2,800 cancer genomes from ICGC). The project has produced large amounts of data including somatic mutations, copy number aberrations, somatic structural variations, RNA expression, gene fusions, germline variations, and more. The project involves several working groups focusing on different aspects. SCML has contributed to several of these with methodology and software for copy number segmentation, high resolution analysis of allele specific copy number alterations in cancer, tumor ploidy and tumor cellularity analysis, and subclonality analysis. The project has led to a deep characterization of mutations in cancer and presents a major step forward in the battle to develop tools that predict the optimal treatment for a cancer patient at the time of diagnosis, and in the search for novel molecular targets for precision therapy in cancer. Website: https://dcc.icgc.org/pcawg

Table 5. Research group's contribution to publications

Publication 1: <i>Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction.</i> In Nature Computational Science, 2022.	Authors (Please highlight group members)	Philippe Paul Auguste Robert, Rahmad Akbar, Robert Frank, <u>Milena Pavlovic</u> , Michael Widrich, Igor Snapkov, Andrei Slabodkin, Maria Chernigovskaya, <u>Lonneke Scheffer</u> , Eva Smorodina, Puneet Rawat, Brij Bhushan Mehta, Mai Ha Vu, Ingvild Frøberg Mathisen, Aurél Prósz, Krzysztof Abram, Alex Olar, Enkelejda Miho, Dag Trygve Truslew Haug, Fridtjof Lund-Johansen, Sepp Hochreiter, Ingrid Hobæk Haff, Günter Klambauer, Geir <u>Kjetil Ferkingstad Sandve</u> , Victor Greiff
	Short description	Accurate prediction of antibody–antigen binding with machine learning is desirable, but there has so far been a lack of unified ML formalization of immunological antibody-specificity prediction problems and large-scale synthetic datasets for benchmarking. This paper introduced the Absolut! software suite which enables parameter-based unconstrained generation of synthetic lattice-based three-dimensional antibody–antigen-binding structures with ground-truth access to conformational paratope, epitope and affinity.
	Research group's contribution	Study design, study implementation, writing the paper
Publication 2: <i>CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching.</i> In Bioinformatics, 2022.	Authors (Please highlight group members)	<u>Rognes T</u> , <u>Scheffer L</u> , Greiff V, <u>Sandve GK</u>
	Short description	An important task in adaptive immune receptor repertoire studies is the comparison of different immune receptor repertoires. This paper introduces CompAIRR, a tool that enables ultra-fast computation of adaptive immune receptor repertoire (AIRR) overlap, based on either exact or approximate sequence matching. It improves computational speed thousand fold relative to the state of the art and uses only one-third of the memory. CompAIRR has been integrated with the machine learning ecosystem immuneML to speed up commonly used AIRR-based machine learning applications.
	Research group's contribution	Conceiving the study, study design, study implementation, writing the paper
Publication 3: <i>The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires.</i> In	Authors (Please highlight group members)	<u>Milena Pavlovic</u> , <u>Lonneke Scheffer</u> , Keshav Motwani, Chakravarthi Kanduri, Radmila Kompova, Nikolay Vazov, Knut Waagan, Fabian L. M. Bernal, Alexandre Almeida Costa, Brian Corrie, Rahmad Akbar, Ghadi S. Al Hajj, Gabriel Balaban, Todd M. Brusko, Maria Chernigovskaya, Scott Christley, Lindsay G. Cowell, Robert Frank, Ivar Grytten, Sveinung Gundersen, Ingrid

Nature Machine Intelligence, 2021.		Hobæk Haff, Eivind Hovig, Ping-Han Hsieh, Günter Klambauer, <u>Geir Kjetil Sandve</u>
	Short description	Adaptive immune receptor repertoires are key targets for biomedical research as they record past and ongoing adaptive immune responses. The capacity of machine learning to identify complex discriminative sequence patterns renders it an ideal approach for AIRR-based diagnostic and therapeutic discovery. So far, widespread adoption of machine learning based solutions has been inhibited by a lack of reproducibility, transparency and interoperability. The method proposed in this paper addresses these concerns by implementing each step of the AIRR machine learning process in an extensible, open-source software ecosystem that is based on fully specified and shareable workflows.
	Research group's contribution	Conceiving the study, study design, study implementation, writing the paper
Publication 4: Protein signature predicts response to neoadjuvant treatment with chemotherapy and bevacizumab in HER2-negative breast cancers. In JCO Precision Oncology, 2021.	Authors (Please highlight group members)	Mads Haugland Haugen, <u>Ole Christian Lingjærde</u> , Ingrid Hedenfalk, Øystein Garred, Elin Borgen, Niklas Loman, Thomas Hatschek, Anne-Lise Børresen-Dale, Bjørn Naume, Gordon B. Mills, Gunhild Mari Mælandsmo, Olav Engebråten
	Short description	In cancer the tumor commonly induces the surrounding tissue to grow new blood vessels, a phenomenon called angiogenesis. Drugs that inhibit angiogenesis, thereby reducing the transport of nutrition to the tumor, has proven effective for a number of cancers. However, in breast cancer only a subset of the patients benefit from antiangiogenic drugs. This paper describes how machine learning was used on protein data from patients treated with such drugs to identify a nine-protein signature score (ViRP) that predicts with high accuracy which breast cancer patients benefit from such treatment. The ViRP score was validated both internally and externally in an independent clinical trial. The predictor has led to a patent application (Global Patent Index EP 4100549 A1) and is currently being tested in an interventional clinical study (Neoadjuvant Personalized Therapy in Estrogen Receptor Positive Breast Cancer; NAPEER+) at Oslo University Hospital. It is also tested in retrospective studies on endometrial cancer and colon cancer.
	Research group's contribution	Conceiving the study, study design, data analysis, study implementation, writing the paper

Publication 5: DNA copy number motifs are strong and independent predictors of survival in breast cancer . In Communications Biology, 2020.	Authors (Please highlight group members)	Arne V Pladsen, Gro Nilsen , Oscar M Rueda, Miriam R Aure, Ørnulf Borgan, Knut Liestøl , Valeria Vitelli, Arnoldo Frigessi, Anita Langerød, Anthony Mathelier, Olav Engebråten, Vessela Kristensen, David C Wedge, Peter Van Loo, Carlos Caldas, Anne-Lise Børresen-Dale, Hege G Russnes, Ole Christian Lingjærde
	Short description	Copy number alterations are a common sign of genome instability in cancer. Precise characterisation of the tumor genome architecture may reveal underlying instability mechanisms and allow outcome prediction and treatment guidance. This paper defined six filters quantifying different aspects of the tumor genome architecture. The resulting Copy Aberration Regional Mapping Analysis algorithm (CARMA) was applied to > 2900 tumors, and the derived motifs provided a highly significant independent prognostic predictor for breast cancer. CARMA captures underlying replication and repair defects, and is currently being tested as a tool for patient stratification in a clinical study at Oslo University Hospital (Improved Breast Cancer Treatment; I-BCT).
	Research group's contribution	Conceiving the study, study design, data analysis, study implementation, writing the paper
Publication 6: ICON: a randomized phase IIb study evaluating immunogenic chemotherapy combined with ipilimumab and nivolumab in patients with metastatic hormone receptor positive breast cancer	Authors (Please highlight group members)	Jon A Kyte, Nikolai Kragøe Andresen, Hege Elisabeth Giercksky Russnes, Signe Øien Fretland, Ragnhild Sørum Falk, Ole Christian Lingjærde , Bjørn Naume
	Short description	Immunotherapy with checkpoint inhibitors has emerged as an important treatment modality for several cancer forms. In hormone receptor positive breast cancer, this therapeutic approach is largely unexplored. ICON is among the first clinical trials combining chemotherapy with PD-1 and CTLA-4 blockade, and the first in breast cancer. An important aspect of this trial was that it explored the combination of three different drugs: one drug induces cell death and this evokes an immune response, and the two other drugs ensures that the immune response stays active even in the presence of inhibitory signals from the tumor cells.
	Research group's contribution	Assisting in the planning of the study and the statistical analyses to be performed at different stages of the project. Critically reviewing the paper.

Publication 7: NucBreak: location of structural errors in a genome assembly by using paired-end Illumina reads . In BMC Bioinformatics, 2020.	Authors (Please highlight group members	Ksenia Khelik, Geir Kjetil Ferkingstad Sandve , Alexander Johan Nederbragt, Torbjørn Rognes
	Short description	This paper presented a new tool called NucBreak aimed at localizing structural errors in assemblies, including insertions, deletions, duplications, inversions, and different inter- and intra-chromosomal rearrangements. The approach taken by existing alternative tools is based on analysing reads that do not map properly to the assembly, for instance discordantly mapped reads, soft-clipped reads and singletons. NucBreak uses an entirely different and unique method to localise the errors. It is based on analysing the alignments of reads that are properly mapped to an assembly and exploit information about the alternative read alignments.
	Research group's contribution	Conceiving the study, study design, implementation and analysis, writing the paper
Publication 8: MoDLE: high-performance stochastic modeling of DNA loop extrusion interactions . In Genome Biology, 2022.	Authors (Please highlight group members	Roberto Rossini, Vipin Kumar, Anthony Mathelier, Torbjørn Rognes , Jonas Paulsen
	Short description	DNA loop extrusion emerges as a key process establishing genome structure and function. This paper introduced MoDLE, a computational tool for fast, stochastic modeling of molecular contacts from DNA loop extrusion capable of simulating realistic contact patterns genome wide in a few minutes. MoDLE accurately simulates contact maps in concordance with existing molecular dynamics approaches and with Micro-C data and does so orders of magnitude faster than existing approaches. MoDLE runs efficiently on machines ranging from laptops to high performance computing clusters and opens up for exploratory and predictive modeling of 3D genome structure in a wide range of settings.
	Research group's contribution	Providing critical feedback and assisting in the improvement of the tool and analyses

Publication 9: PathTracer: High-sensitivity detection of differential pathway activity in tumours . In Scientific Reports, 2019.	Authors (Please highlight group members 	Ståle Nygård, Ole Christian Lingjærde , Carlos Caldas, Eivind Hovig , Anne-Lise Børresen-Dale, Åslaug Helland, Vilde Drageset Haakensen
	Short description	Subtle alterations of gene expression is a challenge to detect. This paper proposed a tool for high-sensitivity detection of differential expression in tumours. Samples are projected onto a low-dimensional manifold defined by genes in a given pathway. A score is found by calculating the distance between each projected sample and the projection of a subgroup of reference samples. The reference samples can represent normal tissue or tumour samples with a particular genotype or phenotype. The proposed tool was demonstrated on gene expression data from > 2000 samples and was shown to predict survival, clinical subtypes, cellular proliferation and genomic instability. Predictions outperformed those obtained with other methods.
	Research group's contribution	Conceiving the study, study design, implementation and analysis, writing the paper
Publication 10: Reduced metagenome sequencing for strain-resolution taxonomic profiles . In BMC Microbiome, 2021.	Authors (Please highlight group members 	Lars Snipen, Inga-Leena Angell, Torbjørn Rognes , Knut Rudi
	Short description	Studies of shifts in microbial community composition have many applications. For studies at species or subspecies levels, the 16S amplicon sequencing lacks resolution and is often replaced by full shotgun sequencing. This paper investigated an alternative called Reduced Metagenome Sequencing (RMS) to estimate the composition of a microbial community. The paper proposed a procedure for analysing such data, based on fragment clustering and the use of a constrained ordinary least squares de-convolution for estimating the relative abundance of all community members. Mock community datasets show the potential to clearly separate strains even when the 16S is 100% identical. Improved abundance estimates were obtained when the community had many very closely related genomes.
	Research group's contribution	Study design, analysis, writing the paper

Table 6. Please add a list with the research groups monographs/scientific books. Max 10 listed: title, first three authors et al and link to webpage (if any).

1. *Differential Equations for Studies in Computational Electrophysiology*. Jæger and Tveito (2023). Webpage: <https://link.springer.com/book/10.1007/978-3-031-30852-9>
2. *Rett på Java (3rd Ed.)*. Brunland, Hegna, Lingjærde et al (2021). Webpage: <https://www.universitetsforlaget.no/rett-pa-java-2>
3. *Modeling Excitable Tissue: The EMI Framework*. Tveito, Mardal and Rognes (2021). Webpage: <https://link.springer.com/book/10.1007/978-3-030-61157-6>
4. *Computing Characterizations of Drugs for Ion Channels and Receptors Using Markov Models*. Tveito and Lines (2016). Webpage: <https://link.springer.com/book/10.1007/978-3-319-30030-6>
5. *Conversations About Challenges in Computing*. Bruaset and Tveito (2013). Webpage: <https://link.springer.com/book/10.1007/978-3-319-00209-5>

2.2 Research group's societal contribution

Describe the societal impact of the group's research. Consider contribution to education, economic, societal and cultural development in Norway and internationally.

The group's theoretical research has resulted in novel algorithms that have found widespread use internationally inside and outside academia. For example, the VSEARCH algorithm (Rognes et al., PeerJ 2016) is an open source tool for metagenomics that has attracted much attention due to its speed and accuracy, with more than 6400 citations. One application of this algorithm is to quantify ecosystem diversity, an important element of the monitoring of environmental effects of climate change. Another example is the ASCAT algorithm (Van Loo et al., Proceedings of the National Academy of Sciences, 2010) where the underlying model and estimation algorithm was conceived by a member of SCML. This algorithm allows genome-wide estimation of the precise copy number state of a tumor from microarray data or sequencing data, and became the standard tool for copy number analysis at the Cancer Research UK Cambridge Institute, the Wellcome Trust Sanger Institute, and the Francis Crick Institute. The algorithm played a major role in the pan-cancer analysis projects performed under the umbrella of the Pan Cancer Analysis of Whole Genomes (PCAWG) consortium, projects that led to profound new insights into cancer initiation and progression and to a string of high profile publications that has changed our view of causes and mechanisms in cancer. A more recent example is immuneML (Pavlović et al., Nature Machine Intelligence, 2021), a tool that has already received much attention as a very powerful platform for immune state prediction, for example in the diagnosis of infectious diseases and autoimmune diseases.

The group's applied research has led to novel biological hypotheses that have been widely adopted by the scientific community. One example is the gene expression signature found in Brodtkorb et al. (Blood, 2014) and later validated in a follow-up paper (Steen et al, Haematologica, 2019); this was the first validated predictor of transformation to aggressive disease in follicular lymphoma. Prior to this it was commonly believed that transformation was a random event that could not be predicted in advance; the two papers showed, however, that transformation could often be predicted years in advance, changing the view of the biology behind transformation.

The group's research has also directly impacted patient care at Oslo University Hospital. The two clinical breast cancer studies Improved Breast Cancer Treatment (I-BCT) and NeoAdjuvant Personalized therapy in Estrogen Receptor positive breast cancer (NAPEER+) were initiated as a result of the Complex Armwise Aberration Index (CAAI) and the protein expression based ViRP score, both derived by SCML and partners at Oslo University Hospital

Table 7. The research group's societal contribution, including user-oriented publications, products (including patents, software or process innovations)

No	Name of publication/product	Date of publication /product	Link to the document
1	Efficient algorithms for single- and multi-track copy number segmentation (Copynumber)	2012	Paper: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-591 Software: https://www.bioconductor.org/packages//2.12/bioc/html/copynumber.html
2	Robust and fast clustering method for amplicon-based studies (Swarm)	2014	Paper: https://peerj.com/articles/593/ Software: https://github.com/torognes/swarm
3	A versatile open source tool for metagenomics (VSEARCH)	2016	Paper: https://peerj.com/articles/2584/ Software: https://github.com/torognes/vsearch
4	Copy aberration regional mapping analysis (CARMA)	2020	Paper: https://www.nature.com/articles/s42003-020-0884-6 Software: https://github.com/arnevpladsen/CARMA/
5	A pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations (PANORAMA)	2020	Paper: https://www.biorxiv.org/content/10.1101/2020.05.04.076901v1.full Software: https://shinyibv02.uio.no/panorama/
6	A predictor for breast cancer patients of response to neoadjuvant treatment with a VEGF inhibitor (ViRP)	2021	Paper: https://ascopubs.org/doi/full/10.1200/PO.20.00086 Patent filing: https://worldwide.espacenet.com/publicationDetails/biblio?CC=WO&NR=2021156137A1&KC=A1&FT=D#
7	Computing adaptive immune receptor repertoires (Compairr)	2022	Paper: https://academic.oup.com/bioinformatics/article/38/17/4230/6646451 Software: https://github.com/uio-bmi/compairr
8	A predictor of survival for older DLBCL patients treated with immunochemotherapy: the geriatric prognostic index (GPI)	2023	Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10483351/ Software: https://wide.shinyapps.io/GPIcalculator/
9	Database of scientific publications by authors affiliated with Oslo University Hospital (Publika)		Software: https://www.publika.no/publika

3. Challenges and opportunities

Information about the strengths and weaknesses of the research group is obtained through the questions above. In this chapter, please reflect on what might be the challenges and opportunities for developing and strengthening the research and the position of the group.

Challenges:

The SCML group was until recently and partly for historical reasons branded solely as a bioinformatics unit, and our expertise in statistics and machine learning has therefore been lesser known to students and scientific partners. This has impeded our ability to attract top students and science partners seeking machine learning expertise.

Scientific computing has had a low profile at the department for years, due to previous reorganisations moving most of these activities out of the department and to either the Department of Mathematics or to the Simular Research Center. Our recent strengthening of this field requires profiling to attract students, partners and funding.

The group's current office location in the NEMKO building is suboptimal as all other research groups in the Machine Learning Section are located in the Ole Johan Dahl building. While the buildings are only a short walking distance apart, the different locations reduces the day-to-day interaction and the potential for unplanned exchanges of ideas.

The lack of a dedicated BSc program in machine learning and/or scientific computing reduces the group's recruitment potential.

The SCML group needs a software engineer - expertise in software develop, ...

Opportunities:

The SCML group has currently rebranded its scientific profile and is in the process of doing the same for its educational profile. We predict that this will attract new student groups and scientific partners, offering fresh opportunities for increased recruitment of MSc and PhD students in machine learning and scientific computing as well as opportunities for new scientific alliances.

The group's recent strengthening of the core areas of scientific computing and machine learning opens for exploration of synergies, including computationally demanding learning problems and learning problems subject to physical or other constraints most easily formulated as differential equations.

More focus on foundational machine learning issues combined with existing strong links to life science research groups presents an important arena for developing, testing and deploying novel methodologies and algorithms.