The Advanced Machine Learning for Innovative Drug Discovery (AIDD) project

Newsletter 3, April 2022

The **Advanced Machine Learning for Innovative Drug Discovery** (AIDD) project is a Marie-Skłodowska-Curie Innovative Training Network (ITN) for Early Stage Researchers (ESRs) funded by the European Commission under the Horizon 2020 Programme, Marie Skłodowska-Curie grant agreement No 956832. The project brings together fifteen academic and industry partners from ten European countries and the University of British Columbia (Canada) to train sixteen PhD students in close collaboration with associated partners from the USA, Australia, China, Israel, and other countries.

Project development

The **AIDD project started on 01.01.2021** and on 18.10.2021, the first day of the First Summer School, all 16 fellows including ESR16 joined the School.

The Project Management Team has been preparing the first Summer School during the period from July 2021 to October 2021. The Consortium invited experts to give talks related to chemoinformatics, machine learning, and applied computer science. Among speakers who accepted the invitation to give a lecture were several high-profile scientists from academic institutions, MIT, Google Research, and NVIDIA.

The school took place in Munich on HMGU Campus located 85764, Neuherberg, Building 57, Room 052. The School lasted two weeks 18th October - 29th October. Not all the speakers and the students managed to come to Munich due to COVID-19 restrictions, and the event was partially organized by Zoom.

Figure 1. AIDD participants at the HMGU Campus

First Summer School

The School started on the 18th of October with a “Welcome Party” where the fellows could meet each other in a formal setting. Bavaria is well known in the World for its Biergarten culture. The Project Management Team decided to make the first meeting in a famous touristic place — SeeHaus in English Garden. AIDD fellows came and enjoyed traditional Bavarian food while getting to know each other. Every day all fellows, as well as guests of the school, were invited to the local HMGU canteen for lunch.

The lectures started on the 19th of October according to the following schedule, see. Fig. 1.

			Building 57 / Room 052		Building 57 / Room 052		Building 57 / Room 052		Building 57 / Room 052	
First week	**18.10.21**		**19.10.21**		**20.10.21**		**21.10.21**		**22.10.21**	
Time	**Monday**		**Tuesday**		**Wednesday**		**Thursday**		**Friday**	
10:00-11:00			RDkit: basics	Gregory Landram, Zoom	Workshop on PyTorch	Thomas Viehmann			Reinforcement Learning	Jose Arjona, 45 mins
11:00-12:00			History of AI	Mark Embrechts, live			Transformers for Computer Vision	Thomas Unterthiner, Zoom	Reinforcement Learning	Philipp Renz, 45 mins
12:00-13:00			Lunch		Lunch		Lunch		Lunch	
13:00-14:00			How did we do deep learning in the BIGCHEM: experiences, challenges and opportunities	Josep	Neural networks: tricks of the trade	Pieter-Jan Hoedt, 45 mins, Zoom	Python: advanced (numpy, decorators, network)	Andreas Poehlmann, Zoom	SMILES based modelling, tools and models	Esben Jannik Bjerrum, Zoom
14:00-15:00			Introdcution to chemoinformatics	Tropsha	Workshop on PyTorch	Thomas Viehmann	QSAR: substructure analysis	Peter Ertl, Zoom	Introduction to best practices	Samuel Genheden, Zoom
15:00-16:00			Personal presentations of ESRs (5 people)	ESR1-ESR6			Introduction to modeling chem reactions with ML	Marvin Segler, Zoom	Recurrent networks, LSTM, and transformer	Michael Widrich, Zoom
16:00-17:00	Welcome party, SeeHaus in Englischen Garten									

	Building 57 / Room 052		Building 57 / Room 052		Building 57 / Room 052		Building 57 / Room 052			
	25.10.21		**26.10.21**		**27.10.21**		**28.10.21**		**29.10.21**	
Second week	**Monday**		**Tuesday**		**Wednesday**		**Thursday**		**Friday**	
10:00-11:00			Python: testing	Samuel Genheden, Zoom						
11:00-12:00	QMAR: quantitative cell morphology-activity relationships	Adam Arany, live	Python: questions and answers	Andreas Poehlmann, Zoom	Optical chemical structure recognition	Clevert, live	Computational Cancer Pharmacogenomics	Michael Menden	Die Josefsthaler Wasserfälle	
12:00-13:00	Lunch		Lunch		Lunch		Lunch			
13:00-14:00	Atomic Simulation Environment (ASE)	Leonardo Medrano Sandonas (live)	Molecular property prediction	Andreas Mayr	Convolutional Neural Networks	Mark Embrechts, live	Similarity-based data mining with applications to drug discovery	Petra Perner	https://docs.google.com/document/d/1FNgqjcD-x1oiizutOlXzIjJG3sdNmlxdK1tZDVQrD1g/edit?usp=sharing	
14:00-15:00	Introduction to Reaction Informatics	Igor Baskin, Zoom	High Performance Computing / Python (cluster, muli-GPU, etc)	Martijn Oldenhof, live	Placebo		Personal presentations of ESRs (5 people)	ESR7-ESR11		
	Coffee break		Coffee break							
15:00-16:00	Molecular representations	Rafael Gomez-Bombarelli, Zoom	Structure Query Language: basics	Pavel Karpov, live	Recurrent Neural Networks	Hans Georg Zimmerman	Personal presentations of ESRs (5 people)	ESR11-ESR16		
16:00-17:00	Accelerated data science with applications in natural language processing	Anthony (Nvidia)	Generative models and optimization for molecules	Rocio Mercado, zoom, US			Organic reactions modeling: start-of-the-art and perspectives	Connor W. Coley, Zoom		
17:00-18:00	Introduction to CUDA with Python	Christian Hundt (Nvidia)								

Fig. 1: Schedule of the First Summer School

Tuesday, 19.10

RDKit: Basics with Gregory Landrum (10:00-11:00, Zoom)

The founder and creator of RDKit, Greg Landrum, gave a talk on the open-source cheminformatics software including an overview of the toolkit and a tutorial of the basics in Jupyter Notebook. The tutorial covered topics such as molecule manipulations, various representations, calculation of descriptors, and chemical reactions.

History of AI with Mark Embrechts (11:00-12:00, Live)

A talk on the overall history of Artificial Intelligence (AI) including early ideas and developments, as well as critical milestones in the various fields that we today consider AI. Early developments were the McCulloch-Pitts neurons, Hebb's theory of unsupervised

learning of neurons, the Turing test defining artificial intelligence, human computing by John von Neumann, Frank Rosenblatt setup of the perceptron. Also mentioned was the AI winter, including its trigger and what should be learned from it.

Take-aways from BigChem with Josep Arús-Pous (13:00-14:00, Zoom)

One of the students from the previous MSCA ITN project BigChem, in the same format as AIDD, talked about his experience in the project and his research. He recently finished his PhD. Some tips related to the research in general were, to focus on the why and to fail often and early. Regarding deep learning, he suggested that the first part of the deep learning workflow, understanding the data, is the most important part since it means understanding the problem.

Introduction to Chemoinformatics with Alexander Tropsha (14:00-15:00, Zoom)

Lecture on general chemoinformatics by Alexander Tropsha. Chemoinformatics can be described as a way of transforming data into information and information into knowledge, with the goal of making faster and better decisions in drug discovery.

The traditional approach to chemoinformatics consists of QSAR studies: correlating structural properties of chemical compounds with their biological activities. However, the term chemoinformatics is becoming broader, as there are many diverse sources of chemical data. One of them is literature: text mining has become prevalent in chemoinformatics. Correlations can be established between physiological, bio/chemical, and disease data present in scientific literature.

Molecules can be represented as vectors of chemical descriptors. The distribution of all possible molecules in this multidimensional space is called the chemical space.

Wednesday, 20.10

The Project Management Team wanted to dedicate the entire day to a workshop on developing a complex neural network architecture from scratch using PyTorch, the machine learning framework that is a standard tool in the AIDD Project. We were able to identify a prominent PyTorch developer, Dr. Thomas Viehmann, who contributed a full day seminar on this topic.

PyTorch Workshop with Thomas Viehmann (10:00-12:00 and 14:00-17:00, Live)

Most of the day was spent on a workshop in PyTorch. Concepts in PyTorch such as high-dimensional tensors, autograd, models, and predefined models in torchvision, optimization, etc. were covered including some hands-on experience. A vision transformer was also constructed step-by-step on a dataset of images with traffic signs.

Dr. Viehmann also shared a great deal of useful tips on PyTorch style and efficiency. Dimension reductions can usually be done without *.unsqueeze()*, simply by setting the desired dimension to *None*. Model state *dicts* should always be stored instead of the models themselves. The *.item()* call should never be called inside the inner-loop of PyTorch training since it creates unnecessary synchronization points of the parallelization. Instead, it should

only be called once, at the end of an epoch to store statistics running over the epochs. The main advantage of using Adam optimizer over SGD is that it decouples the learning rates for weights and biases. This can be useful/crucial when the magnitudes of their respective gradients vary significantly, since in this case, it could be so that no learning rate works in SGD. The reason for this is that the learning rate, in this case, could be too big for the weights whilst too small for the biases or the other way around.

Tricks of the Trade in Neural Networks with Pieter-Jan Hoedt (13:00-14:00, Zoom)

A lecture summarising takeaways from the book Neural Networks: Trick of the Trade, split into the topics: data, model, and learning. Regarding the data, areas such as preprocessing, data leakage, etc. were covered. The main takeaways were to: understand your data, keep your model simple, and remember that learning benefits from tuning.

Thursday, 21.10

Transformers for Computer Vision with Thomas Untherthiner (11:00-12:00, Zoom)

Thomas Untherthiner from Google Research gave a talk explaining the background to transformers in computer vision and how they are being used at Google for various cutting edge tasks. The background included some of the milestone models during the development of deep neural networks, such as AlexNet, ResNet, self-attention, that eventually lead to the creation of transformers. BERT is the transformer that made the initial significant improvements in Natural Language Processing (NLP).

At Google, they have their own internal image dataset called JFT-300, containing 300M images compared to the 14M images contained in the well-known public dataset ImageNet. Untherthiner made a comparison between Vision Transformer (ViT), Google's model, to Big Transfer (BiT) which is based on ResNet. For ViT the huge amount of data in the JFT-300 dataset is crucial whereas BiT works better on smaller datasets such as ImageNet.

Advanced Python with Andreas Poehlmann (13:00-14:00, Zoom)

This talk gave a detailed tutorial on Numpy, decorators, network, etc. in Python, using this GitHub repository. Poehlmann explained the underlying mechanics of how numpy stores data, demonstrated the importance of using decorators and introduced the request library for HTTP getting and posting amongst other things.

QSAR: Substructure Analysis with Peter Ertl (14:00-15:00)

A lecture on substituents and functional groups, also referred to as substructure analysis, and their role in QSAR. Functional groups (FGs) are sets of atoms and bonds that are responsible for a set of chemical reactions, and which provide interesting information about properties of the molecules in which they are present. This talk covered from descriptors that can be used to characterize FGs (size, hydrophobicity, acceptance or donation of electrons), to methods that can be employed in the drug discovery process with regards to FGs. One example of such a method is bioisosteric design, which consists in replacing part of the molecule with another substituent that is similar but has slightly different properties, with the aim of keeping biological activity but reducing toxicity or increasing bioavailability. An

algorithm to identify functional groups in organic molecules, developed by Peter Ertl, was also presented.

Chemical Reaction Modelling using ML with Marwin Segler (15:00-16:00, Zoom)

Marwin Segler held a talk on modelling of chemical reactions. This can be done on different levels, e.g. on individual reactions as molecular graphs or on a global network of reactions as a knowledge graph with molecule and reaction types. Depending on the problem one can also be interested in the reaction mechanism, i.e. the sequence of elementary steps from reactants to products. Individual reactions can also be modelled in the forward sense, called synthesis, or in the backward sense, called retrosynthesis. There is a large number of questions that can be asked regarding reaction modelling, such as predicting products, possibility of reaction happening, reaction class.

Another aspect of reaction modelling is the representation of the reactions. Many representations exist but for machine learning certain ones are more suitable, such as fingerprints as they are bit vectors. A general advice given for deep learning research in reaction modelling is that the field moves forward quickly. Hyperparameter tuning matters a lot, but there are usually too many hyperparameters to explore.

Friday, 22.10

Introduction to Reinforcement Learning with Jose Arjona (10:00-11:00, Zoom)

Jose Arjona from Dynatrace gave some background to reinforcement learning including terminology, basic concepts, Markov Decision Processes, Bellman's Equation, policy iteration, and value functions. Also covered was a comparison of differences between Dynamic Programming, Monte-Carlo, and Temporal-Difference used to search the action space in a reinforcement learning setting.

Generative Models for Molecules with Philip Renz (11:00-12:00, Zoom)

An introductory overview on two methods for generating molecules, distribution learning and goal-directed generation, including some of the work done in both these areas by Philip Renz and colleagues at JKU. Distribution learning in this context means to find a new molecule similar to existing ones, i.e. according to the distribution of a given group of known molecules. The so-called Guacamol metrics can be used to evaluate different aspects of generated molecules, such as novelty, validity, uniqueness etc. AddCarbon is a generative model of molecules built on distribution learning developed by Philip Renz et al. It's performance was compared to other generative models during the talk. Goal-directed generation on the other hand, aims to generate molecules that satisfy some property profile.

SMILES-based Modelling with Esben Jannik Bjerrum (13:00-14:00, Zoom)

Esben Jannik Bjerrum from AstraZeneca talked about how the SMILES-representation of molecules can be used for modelling chemistry as well as useful tools and the different models suitable and available. SMILES-strings are foremost useful in NLP, where they should first be tokenized into a vocabulary. PySMILES utils is a package created for this

purpose by Bjerrum, available in GitHub. As with other types of data, SMILES can be augmented to increase the size of the training set. Step-by-step de novo compound generation was demonstrated using a Recurrent Neural Network (RNN). Another method for molecule generation is the beam search algorithm, which can be used to generate multiple SMILES-strings deterministically.

Good Software Development Practices with Samuel Genheden (14:00-15:00, Zoom)

Also from AstraZeneca, Samuel Genheden spoke more on the side of software development. In modern software development, the workflow is typically an iterative process in an agile framework, such as Scrum. Good coding practices are important for repeated use of code, handing over code to others, fixing bugs, and extending its function.

Recurrent Networks including LSTMs with Michael Vidrich (15:00-16:00, Zoom)

An overview and background of Recurrent Neural Networks (RNNs), including standard pitfalls and drawbacks such as the vanishing gradient problem. Extending RNNs with Long Short-Term Memory units was also discussed briefly as a means to solve the vanishing gradient problem. The talk also touched on transformers, attention and modern hopfield networks.

Monday, 25.10

QMAR with Adam Arany (11:00-12:00)

A talk about machine learning approaches to Quantitative cell Morphology-Activity Relationships (QMAR) by Adam Arany from KU Leuven. QMAR can be modelled as a multi-task learning setting. Such machine learning tasks can be solved in a number of different ways, such as with matrix factorization in probabilistic models on one hand and using deep learning methods on the other hand. Netflix held a machine learning competition to improve their recommender system, where they released a dataset of user information including ratings of categorised movies. Since a recommender system is a multi-task setting, this led to developments in this field especially in deep learning.

Adam Arany also presented research that he worked on together with other academic and industry partners, where high-throughput data from a single imaging assay were repurposed to predict biological activity of compounds in other assays. In this work a Bayesian matrix factorization method Macau was compared to a deep neural network as well as random forest and k-nearest neighbour models.

Atomic Simulation Environment with Leonardo Medrano, (13:00-14:00, Live)

Atomic Simulation Environment (ASE) was introduced in this talk, as a set of tools and Python modules for setting up, manipulating, running, visualising, and analysing atomistic simulations. As an example the deep tensor neural network, SchNetPack, can be used as a machine learning calculator in the ASE calculations.

Reaction Informatics with Igor Baskin, (14:00-15:00, Zoom)

Igor Baskin held a talk on the history of reaction informatics together with some general background in the field. The first part covered how the classification of chemical reactions has evolved over time. Weygand, Telheimer (addition, elimination, rearrangement and substitution), Balaban, Arens and Vladutz (superimposed reaction skeleton graph) classification methods were mentioned. Then, the history of chemical reaction descriptions was also described. Hendrikson unified this description by assigning bonds between atoms to different types of lines depending on whether there was a bond breakage or formation. Ugi-Dugundji's model encodes information about chemical bonds in three different matrices: reactants, products and reaction itself.

Molecular Representations with Rafael Gomez-Bombarelli, (15:00-16:00, Zoom)

An overview of the ways that molecules can be represented, their advantages and disadvantages as well as when certain ones are most suitable. Fingerprints are suitable for ordinary neural networks whereas graph neural networks make it possible to work directly with molecular graphs. The basics of how message passing is done in Graph Convolutional Networks (GCN) was covered followed by how GCN can be applied to 2D molecular graphs, with varying features, and to 3D point clouds of molecules. In the 3D case, the data can still be considered a graph by connecting points at a given cut-off distance to each other.

Accelerated Data Science with Anthony Costa, (16:00-17:00, Zoom)

Anthony Costa from Nvidia introduced RAPIDS for end-to-end GPU accelerated data science with libraries for pre-processing (cuIO, cuDF instead of pandas), machine learning (cuML insteal of scikit-learn), deep learning (TensorFlow, PyTorch), graph analytics (cuda-X instead of NetworkX), and visualisation on GPU. The transfer from CPU to GPU results in lightning fast speed-up performance, from hours to seconds. The talk also included a tutorial on some of the mentioned libraries, specifically performing a TF-IDF vectorization on a dataset of Covid-19 tweets, in order to demonstrate the speed-up achieved with the GPU acceleration.

Accelerated Data Science using CUDA with Christian Hundt, (17:00-18:00, Zoom)

Also from Nvidia, Christian Hundt talked more specifically about how CUDA can be used to accelerate data science in Python. GPU’s should be used because of its ability for massive parallelism, which in terms of hardware comes down to the large number of cores. The acceleration was mostly demonstrated through another tutorial.

Tuesday, 26.10

Python testing with Samuel Genheden, (10:00-11:00, Zoom)

Samuel Genheden gave a second seminar, this time on how to test Python code, with a focus on the pytest library. Topics covered included, levels in which testing can be carried out (unit, integration, end-to-end, acceptance), mocking and pytest fixtures. Also mentioned was that testing should be divided into three steps: arrange (importing libraries and setting up objects, files, etc), act (actually calling the function) and assert (check that the obtained

result is equal to the expected result). The second part of the seminar was dedicated to going over some example unit tests.

Python: Questions and Answers with Andreas Poehlmann, (11:00-12:00, Zoom)

A seminar reserved for asking python related questions. Concurrent programming was mainly covered, including a demonstration of the basics of forking and joining threads using a number of different python libraries. Besides the usual threading python library, a more modern one that allows interacting with threads is “concurrent.futures”, from which the class ThreadPoolExecutor can be imported.

Molecular Property Prediction with Andreas Mayr, (13:00-14:00, Zoom)

An overview of the task to predict molecular properties, covering the context of the task, public data sources and machine learning approaches. Assay measurements were explained together with relevant molecular representations such as graphs, SMILES, InChI. Suitable machine learning approaches for each representation were discussed as well as research done in the different fields by Andreas Mayr and colleagues at JKU.

High Performance Computing with Martijn Oldenhof, (14:00-15:00, Live)

In this talk high performance computing was discussed. It starts with hardware improvements, but eventually parallel computing is needed in order to achieve high performance computing. The concepts of strong and weak scalability were covered. Parallel speed-up should be weighed against the parallel efficiency, i.e. how much of the resources are spent on useful computation versus overhead for the parallelism. Even when multi-processing is available the code should always be optimised as much as possible, it is not enough to have a scalable solution. Oldenhof also mentioned that Python in itself is slow, partly due to the Global Interpreter Lock (GIL) and the dynamic types. However, there are many useful libraries that have been written to improve the speed of Python scripts.

Structured Query Language with Pavel Karpov, (15:00-16:00, Live)

AIDD Project Manager Pavel Karpov gave an introductory lecture about Structured Query Language (SQL), the language for writing and handling large databases. Databases are usually row-based, meaning that examples or entries are listed in rows. The talk included some SQL providers and functions typically provided, such as adding and removing items, joining databases in various ways, handling of NULL objects, etc.

Molecular Generation and Optimization with Rocio Mercado, (16:00-17:00, Zoom)

Rocio Mercado, from MIT, gave a talk on deep generative models for molecular generation, optimization, and evaluation of such models, as well as an outlook on the future of the field. Molecular generation is interesting in the context of molecular design where certain desired molecular properties are the goal. This includes both traditional design and de novo design. Generation schemes can be done in a single-shot approach or iteratively in an autoregressive way. For molecular optimization, we can use transfer learning, evolutionary algorithms, Bayesian optimization or reinforcement learning. GraphINVENT was also presented. This is a platform for graph-based molecular generation using Graph Neural Networks (GNNs) that Rocio Mercado was a part of developing.

In the evening, all the students, as well as the guest speakers, spent unforgettable hours during the official AIDD Dinner at Augustiner Keller, the traditional Bayerische restaurant.

Wednesday, 27.10

Optical Chemical Structure Recognition with Djork-Arné Clevert, (11:00-12:00, Live)

Djork-Arné Clevert from Bayer gave a talk about his research on Convolutional Neural Networks (CNNs) to learn representations of images of molecular graphs and decoding them to SMILES.

Convolutional Neural Networks with Mark Embretch, (13:00-14:00, Live)

A talk diving into the idea of convolutions in neural networks, rather than the mathematical theory behind the computations. Simple examples were used to demonstrate how a small convolutional filter examines the spatial data to identify a single pattern, where present. These filters are especially powerful when followed by a pooling layer resulting in a summarized form of where the pattern is present. Multiple parallel filters are used to identify multiple patterns, often referred to as channels, and convolution followed by pooling is typically done iteratively in layers on sequentially lower-resolution data. This concept can also be done in reverse to upsample data.

Recurrent Neural Networks with Hans Georg Zimmerman, (14:00-16:00, Zoom)

Initially Hans Georg Zimmerman introduced the basics of neural networks, including the mathematical formulations and the error backpropagation. Further, the background on RNNs was given and their effectiveness on time-series data was demonstrated for the application of forecasting dynamical systems. The talk also touched on topics such as exploding gradients.

Thursday, 28.10

Computational Cancer Pharmacogenomics with Michael Menden, (11:00-12:00, Zoom)

This talk by Michael Menden introduced the concepts of cancer genomics and pharmacogenomics. Regarding the former, the lecture covered from the general dogma of molecular biology (how information is passed from DNA to RNA to protein) to the effect that different types of mutations can have in the resulting phenotype. I.e: indel mutations in general are more aggressive than in frame coding mutations. The concepts of driver and passenger mutations were also explained. Finally, it was described how natural selection occurs when targeting cancer, as only drug resistant cells survive, complicating further treatments. A suitable approach for this is selecting smart combinations of drugs.

Similarity-based Data-mining with Petra Perner, (13:00-14:00, Zoom)

Petra Perner gave a talk about similarity-based data-mining applications in drug discovery. Similarity between objects can be measured in different ways, often some sort of distance in the feature space is used. Topics covered were concepts of case-based reasoning and

applications of CBR to image processing and information management tasks that have been applied to drug discovery.

Organic Reactions Modelling with Connor W. Coley (16:00-17:00, Zoom)

Associate Professor Connor W. Coley from MIT, held a talk about current state-of-the-art methods and his own perspective on organic reaction modelling. The goal of this kind of modelling is a property profile and the process typically iterates between predictions using e.g. QSAR and validation using e.g. assaying. This is done in the hope of leading to the discovery of a new candidate. Data sources in this field vary greatly in accessibility, however a few large datasets are publicly available, such as USPTO, Pistachio, Reaxys, SciFinder and Open Reaction Database (ORD). Synthesis planning can be done in a single step or in multiple steps.

Currently, graph-based models are ahead in the field, but closely followed by sequence-to-sequence modelling predicting product SMILES from reaction SMILES. A problem with the current models is that they can sometimes come up with products which are not synthesizable. To mitigate this, Coley's team are considering generative models with the constraint of synthesizability.

Hiking in the Schliersee region

Because of the changeable weather forecast, it was not known by Friday, 29.10 morning, whether the hiking would take place at all. Pavel Karpov organised a hike to "Die Josefsthaler Wasserfälle", which is an easy and doable route. However, the weather was fantastic, and the group decided to reach one of the tops, namely Brecherspitz (1685 m), and enjoy the panoramic view on Schliersee, Fig. 2.

Fig. 2: View on Schliersee. The photo was taken by Pavel Karpov.

The AIDD ESRs received the AIDD school well, as indicated by their feedback report. During the School, there were 33 lectures (8 in the live format, 25 Zoom), the social events (Welcome Party, Dinner, and Hiking).

Additional information

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement No 956832.