SHAPE 12th Call - Application Form

At this application stage some SMEs may not be in a position to provide all of the information requested e.g. specific hardware details, quantity of machine time etc. The review process can accommodate this, but please provide as much relevant information as possible to assist in the assessments.

Please overwrite the highlighted passages with your own text.

Project Title	Drug discovery and Development
Company Name	Integrative Biocomputing - IBC
Company Address	Centre d'Affaires Buro Club – BP 97143 Place du Granier 35571 Chantepie Cedex France
Company Website	https://www.ibiocomputing.com
Number of Employees	1
Contact (full name)	Pridi SIREGAR
Contact (job title)	General Manager
Contact (e-mail)	Pridi.siregar@ibiocomputing.com
Contact (phone)	**International code : (+33)** **Phone number:** 04 73 16 39 80

Which of the following describes best the main revenue source of your company?

(X) Private sector

( ) Public funds

Has your company worked with PRACE before?

No

What does your company do?

(max. 300 words)

IBC's core expertise covers Computational biology, serious games and artificial intelligence (AI). In computational biology, IBC has built a prototype Generic Modelling and Simulating Platform (GMSP) designed to generate virtual complex tissues and organs from virtual stem cells. The GMSP is a knowledge-base system (KBS) written in C++ linked to a database that encodes the system's knowledge in cell biology. The dynamics of tissue-generation uses a mixed discrete-continuous formulation where, e.g., discrete objects are cells surrounded by continuous representations of electric and/or chemical fields (Poisson equation, Diffusion equation). Computer-generated virtual tissues and organs produced by IBC's technology have been used in

academia for both research and higher education. The company's 3D dynamic heart simulator was awarded the 1st prize - National award by the French Association for Therapeutic Research in 2000. More recently, multiscale virtual 3D kidney structures from virtual stem-cells were generated by the GMSP and international scientific reports and papers on the subject were published.

Serious games include modelling static and dynamic scenes (e.g. avatars in 3DSMax), programming the game-engine (e.g. Unity3D) and in the case of IBC, incorporating a KBS as a virtual expert that enjoys the role of a tutor. Its serious games in medicine obtained 1st prize at the International Laval virtual Awards 2012. A serious-games prototype in haematology was realized in 2013 in collaboration with the U. Cambridge in the UK.

The company is currently "rebooting" with a strong emphasis in AI.

AI covers machine-learning (ML), deep-learning (DL) and knowledge-base systems (KBS). ML and DL programs are written in python, and KBSs in C++ linked to a database. The company is designing an AI-Core that will integrate these technologies, thus combining bottom-up data-driven inference (ML/DL) with top-down knowledge/hypothesis-driven (KBS) inference. The AI-Core is thus designed to better emulate natural cognition.

Project Abstract

(max. 150 words)

The ultimate (~two years) goal is to create a generic Knowledge-base system (KBS) centered on an AI-Core dedicated to in-silico drug-discovery that includes predicting the pharmacokinetics and pharmacodynamics of new compounds and *de novo* drug design. For the current PRACE-SHAPE project (6 month) we will focus on *de novo* drug design and apply DL methods of the *generative* type such as variational auto-encoders (VAE) and Generative Adversarial Networks (GAN). Indeed, VAEs and GANs can be used to generate novel molecular formulas by compressing high dimensional molecular features into a low dimensional *latent space* from which novel molecular formulas can be generated. In the case of VAE, generation consists in the stochastic sampling of the latent space, while in GANs it is based on reinforcement-learning (RL). As in all DL approaches, training VAEs and GANs can require HPC and in order to do that, expertise from PRACE is needed.

Industrial relevance and potential business impact

(max. 250 words)

Basic research in the life sciences is probably the most multi-disciplinary quest of human inquiry about nature. Despite the tremendous (highly trained) human and financial resources devoted to 'wet-labs' it seems increasingly difficult to find truly novel drugs. When new drug candidates are produced by medicinal chemistry, wet lab studies are carried-out in order to assess their pharmacokinetics and pharmacodynamics properties in four phases (0, I, II, and III) corresponding to the preclinical (phase 0), and clinical phases (I, II, III) that can span over 10 years with failure rates beyond 60%. The total development cost per approved drug that has skyrocketed since 2016 (from ~ $1,5 billion to over $2 billion today). In order to assist experimental drug discovery and mitigate the huge human and financial investments, 'dry lab' technologies have become increasingly popular since, from the resource's standpoint, they "only" require computers and human brains.

The ultimate technical goal is to create a generic KBS dedicated to *in-silico* drug-discovery pipelines including ligand-based and structure-based virtual screening (VS) of novel compounds, bioactivity prediction, lead optimization, drug repurposing, and *de novo* drug design.

The business model is B2B with, as target clients, pharma companies involved in the discovery of new drugs. IBC's offer will be to contribute to the *in-silico* component of drug-discovery pipelines.

Proposed high level Work Plan				
Start date:		January 2021		
Task	**Title**	**Description**	**SME effort (PM)**	**PRACE effort (PM)**
1	Problem-definition	Define the molecular target(s) that will be included in the study and related to cancer of the endometrium.	0.1	0
2	Data-mining	Mine specialized databases such as PubChem, DrugBank, ChemBL,…	1.5	0
3	Database creation	Create a proprietary database that encode molecular features covering physicochemical, biological and (if possible) medical information.	1	
4	Feature importances and selection	Order molecular features based on importance using statistical and ML technics. Define promising feature subsets.	0.5	
5	Molecular representations	Transform compounds' chemical formula into SMILES representation. Transform selected features into Fingerprints. SMILES and Fingerprints will constitute the input data of the DL methods.	0.3	
6	GAN and VEA coding	Write GAN and VAE python programs adapted to the project.	0.5	
7	Preprocessing	Check class balance and other pre-processing tasks.	0.1	0
8	DL	Hand over to the PRACE partner: (1) input data (2) VAE and GAN programs (3) Program to check the chemical validity of the generated molecular formulas (if possible).	0	1
9	Optimization	Based on profiling data from previous task, propose strategies for VAE and GAN hyperparameter tuning and code optimisation.	0.5	0.5
10	Chemical validation and feasibility study	Check the chemical validity and feasibility (e.g. thermodynamics) of the generated molecular formulas.	0.5	0.25
11	Data mining	Mine 3D structures and physical properties of the target molecules (defined in task 1)in data banks such as Protein Data Bank (PDB)	0.1	
12	Molecular docking (if possible)	Automate (or semi-automate) the rather repetitive task of molecular docking using AutoDock MGL, AutoDock Vina and Pymol. Define putative binding sites of the ligands on the target proteins. Ligands are the selected compounds defined in task 10. The target proteins are defined in task 1.	4	4
12'	Alternative to task 12	Check the novelty of the ligands by mining all compounds in open-source DBs (12 and 12' TBD)	4	4
13	Final report	Report on the outcomes of the work.	0.25	0.25
		Total	10.85	6

Technical and business requirements

Please identify technical and non-technical requirements for the project.

Please note that it is understood that not every SME will be able to complete this section. You can mark the section in question with N/A if that is the case. However, **where it is available, provide as much information as possible** in the subsections below.

Technical requirements may include third-party software, preferred hardware architecture, non-standard security levels (like VPNs or data transfers methods not sftp-based), data storage estimates, remote visualization capabilities, etc... If there is a particular PRACE centre/HPC service that you wish to work with also please state this here.

Non-technical requests may include Non-Disclosure Agreements, data locality issues, etc…

Compute Resource

Existing compute resource	A PC
Preferred compute resource	Powerful GPUs: typical DL methods coded in python use TensorFlow/pyTorch that have been optimized for NVIDIA GPUs.
Parallelisation strategy	Whathever they use for GPUs
Storage (Gbyte)	*~50 Gigabytes* - Downloads of 5-10 public ligand DBs such as PubChem, DrugBank, ChemBL … *~500 Gigabytes* – Download protein DBs such as Protein Data Bank
Third party software	Python 3.6.10 - open source MySQL – open source RDKit – open source: http://www.rdkit.org In this project or a future one (see task 12 and 12') AutoDock MGL Tools http://mgltools.scripps.edu AutoDock Vina – open source : http://vina.scripps.edu/ PyMol – open source : https://pymol.org/2/
Typical run	~ 1 – 2 hours on a GPU or 1-2 hours on 700 cores
Core hours	100 – 200 hours on 700 cores
Memory	*Description of requirements of the maximum expected memory usage of the code e.g. 1GB per core, 10GB per node ?*
Other	I/O formats will be excel and/or csv

Non-technical resource

Extract of 'Terms of References': "In the case where particular results may be commercially sensitive to the SMEs business, it is still expected that a white paper could be published with a higher-level discussion of the techniques and approach used, omitting the details of particular sensitivity."

Confidentiality	NDA requested
Other	*Any further relevant non-technical details*