PrECISE

# D7.4

# Integrate methods, including ACSN and Watson

| | |
|---|---|
| **Project number:** | 668858 |
| **Project acronym:** | PrECISE |
| **Project title:** | PrECISE: Personalized Engine for Cancer Integrative Study and Evaluation |
| **Start date of the project:** | 1st January 2016 |
| **Duration:** | 36 months |
| **Programme:** | H2020-PHC-02-2015 |

| | |
|---|---|
| **Deliverable type:** | Other |
| **Deliverable reference number:** | PHC-668858 / D7.4 / V1.0 |
| **Work package contributing to the deliverable:** | WP7 |
| **Due date:** | December 2018 – M36 |
| **Actual submission date:** | 21st December, 2018 |

| | |
|---|---|
| **Responsible organisation:** | IBM |
| **Editor:** | Róbert Alfoldi, Matteo Manica |
| **Dissemination level:** | PU |
| **Revision:** | V 1.0 |

| | |
|---|---|
| **Abstract:** | Methods developed by IBM as cloud services are integrated in the SmarBiobank. |
| **Keywords:** | Method integration, molecular data analysis, web services |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.

This work was supported (in part) by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0324-2. The opinions expressed and arguments employed therein do not necessarily reflect the official views of the Swiss Government.

## Editor

Róbert Alfoldi (ABT)

Matteo Manica (IBM)

## Contributors

Laszlo Puskas (ABT)

María Martìnez Rodríguez (IBM)

Jelena Čuklina (ETH)

## Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The users thereof use the information at their sole risk and liability.

# Executive Summary

We report on the work performed to integrate IBM Research services developed during the project in the SmartBiobank (SBB).

IBM and ABT collaborated to link the different APIs allowing the display of the IBM Cloud services as frames inside a page hosted on the SBB Cellanalyzer.

Analogously to the work performed in D7.3 to integrate ACSN we show how the integrations has been performed and how SBB users can have direct access to the IBM Cloud services implemented in the context of PrECISE: COSIFER, INtERAcT, Chimaera and PIMKL.

The integration is compatible with Chrome and Safari and can be freely used for academic/research purposes.

The impact in clinical studies of the services developed has been exhibited and validated by applying the algorithms on the data produced by partners UZH and ETH. The analysis and results obtained can be found in WP1, WP4 and WP6 deliverables.

# Contents

**Chapter 1 Introduction ........................................................................................ 1**

**Chapter 2 IBM Cloud services ................................................................................ 2**

2.1 COSIFER........................................................................................................ 2

2.2 INtERAcT........................................................................................................ 3

2.3 PIMKL............................................................................................................. 3

2.4 Chimaera ........................................................................................................ 4

**Chapter 3 SmartBiobank integration ..................................................................... 6**

**Chapter 4 Conclusion.............................................................................................. 8**

**List of Abbreviations ............................................................................................. 9**

# List of Figures

Figure 1: Schematic description of the consensus network inference implemented in COSIFER. ........................................................................................................ 2

Figure 2: Schematic description of the algorithm implemented in the backend of INtERAcT web service. ........................................................................................................ 3

Figure 3: Schematic description of the algorithm implemented in the backend of PIMKL web service.......................................................................................................... 4

Figure 4: Depiction of the clonality inference problem from multiple biopsies solved in Chimaera......................................................................................................... 5

Figure 5: SmartBiobank Dashboard with four integrated systems. Clinical Database, Data Storage System, ACSN/NaviCell Maps and IBM Cloud services. ......................................... 6

Figure 6: The developed  REST API is able to communicate with PIMKL, COSIFER, INtERAcT and CHIMAERA web based servers. ................................................................. 7

# Chapter 1 Introduction

During the work performed by different partners in PrECISE a variety of software tools have been developed.

IBM deployed on IBM Cloud a collection of web services that implements all the algorithms designed for WP1, WP3 and WP4 deliverables. This set of services can be used in orchestration or as single modules to analyse different types of omics data and make available to a broad user community methodologies to infer molecular networks, stratify patients, predict samples phenotype and analyse tumor clonal composition. We strongly believe that these services will make available to cancer researchers (e.g., biologists and clinicians) with limited computational expertise an easy access to state-of-the-art methodologies useful for their research.

To enable a seamless integration with other tools developed by other partners the IBM Cloud services have been integrated in the SmartBiobank as described in the following Chapters.

In this deliverable we describe the work carried out on IBM and ABT side to integrate the services and we provide a description on how to use the system.

The integration of ACSN, from CI, has been already described in D7.3 and it was excluded from the content of the current document.

# Chapter 2 IBM Cloud services

IBM during the work in PrECISE has developed four algorithms completed with an open access web serviced implementation to allow the research community an easy access to the tools adopted in the project. The services implemented tackle different challenges in systems biology: consensus network inference from molecular data (COSIFER), network inference from literature (INtERAcT), network-based patient stratification (PIMKL), tumor clonal composition inference from multi-area biopsies sequencing (Chimaera) (see WP4 and WP1 deliverables for details about the specific algorithms).

## 2.1 COSIFER

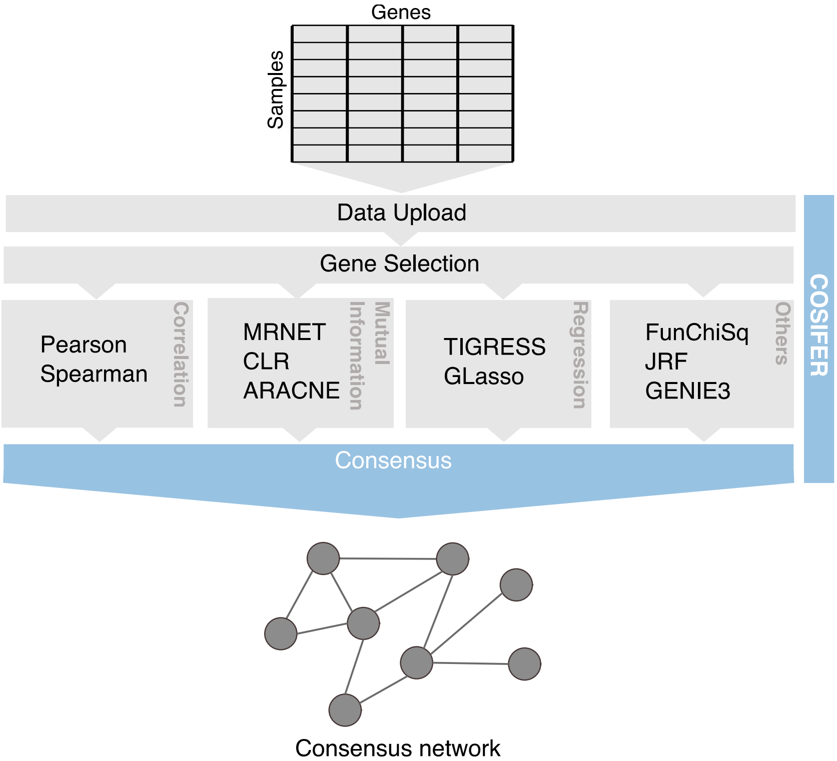

Figure 1: Schematic description of the consensus network inference implemented in COSIFER.

Intracellular networks regulate every kind of cellular decision such as differentiation, proliferation and apoptosis. When these control mechanisms fail, cancer and other diseases may occur. The complexity of these networks originates from the large number and various interactions of the molecules involved. High-throughput technologies such as microarrays and RNA sequencing provide snapshots of the transcriptome and enable insights into the internal regulatory apparatus of a cell. However, inferring the topology of these networks and identifying their key regulators is a challenging task, and international consortia have intensively worked on the development of computational methods tackling this problem. Despite the efforts to compare and develop gene regulatory network inference methods, the research community still lacks easy to access inference tools available to everyone. COSIFER is a web-based platform providing a service for the inference of molecular networks using a consensus between state-of-the-art methodologies given molecular measurements and a list of molecular entities of interest.

The user provides gene/protein expression measurements and the molecular entities of interest and can select different inference approaches that will be combined in a consensus network.

## 2.2 INtERAcT

In recent years, the number of biomedical publications freely available in the literature has grown enormously, resulting in a rich source of untapped new knowledge. However, most biomedical data is buried in the form of unstructured text, and their exploitation requires expert knowledge and time-consuming manual curation of published articles. Hence the development of novel methodologies that can automatically analyze textual sources, extract facts and knowledge, and produce summarized representations that capture the most relevant information in a timely fashion.

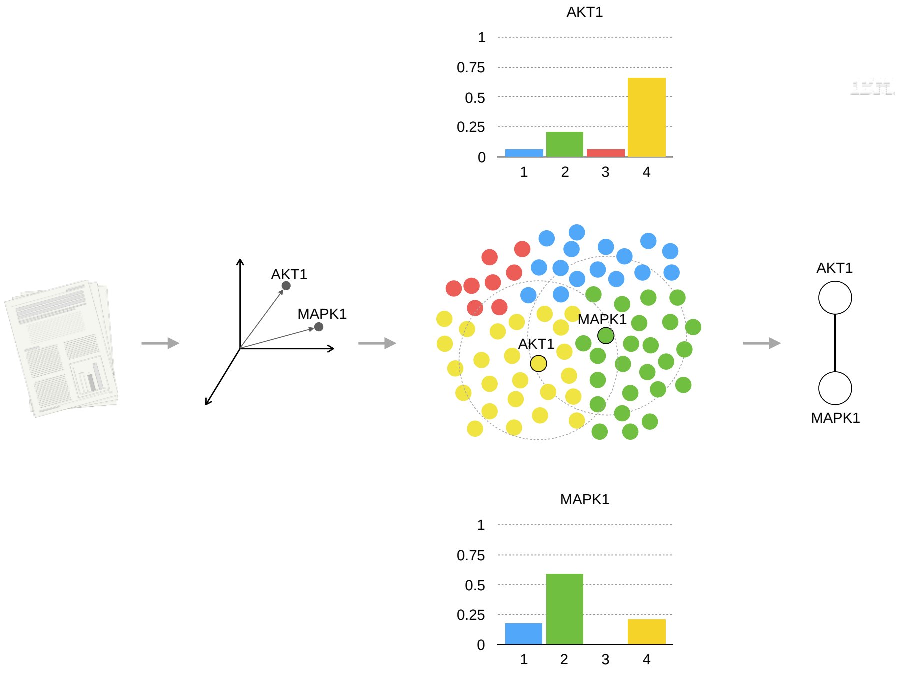

Figure 2: Schematic description of the algorithm implemented in the backend of INtERAcT web service.

INtERAcT represents a novel approach to infer interactions between molecular entities extracted from the literature using an unsupervised procedure that leverages recent developments in automatic text mining and analysis. INtERAcT implements a new metric that acts on the vector space of word representations to estimate an interaction score between two molecules.

For detailed instructions on how to use the web service see the user guide included in the deliverable: *interact_user_guide.pdf*.

## 2.3 PIMKL

Reliable identification of molecular biomarkers is essential for accurate patient stratification. Although state-of-the-art machine learning approaches for classification continue to push boundaries in terms of performance, most of these methods are not able to integrate different data types and lack generalization power, limiting their application in clinical settings.

Furthermore, they behave as black boxes, and provide limited insights about the mechanisms that lead to the prediction.

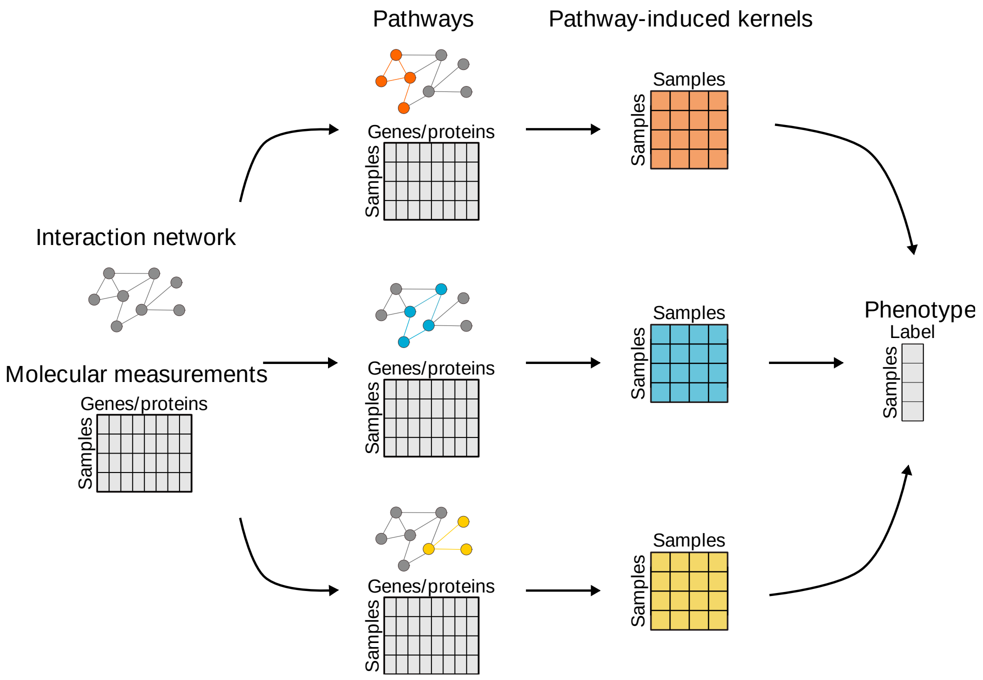

Figure 3: Schematic description of the algorithm implemented in the backend of PIMKL web service.

Pathway-induced multiple kernel learning (PIMKL) is a novel, interpretable methodology to reliably classify samples highlighting molecular mechanisms responsible for the prediction of a phenotype of interest. PIMKL exploits prior knowledge in the form of a molecular interaction network and annotated gene sets, by optimizing a mixture of pathway-induced kernels using a multiple kernel learning (MKL) algorithm. The model provides a stable molecular signature, interpretable in the light of the ingested prior knowledge, that can be further used in transfer learning tasks.

For detailed instructions on how to use the web service see the user guide included in the deliverable: *pimkl_user_guide.pdf*.

## 2.4 Chimaera

Knowledge about the clonal evolution of each tumor can inform driver-alteration discovery by pointing out initiating genetic events as well as events that contribute to the selective advantage of proliferative and potentially drug-resistant tumor subclones.

A necessary building block for the reconstruction of clonal evolution from tumor profiles is the estimation of the cellular composition of each tumor subclone cellularity. These, in turn, are based on estimates of the relative abundance frequency of subclone-specific genetic alterations in tumor biopsies.

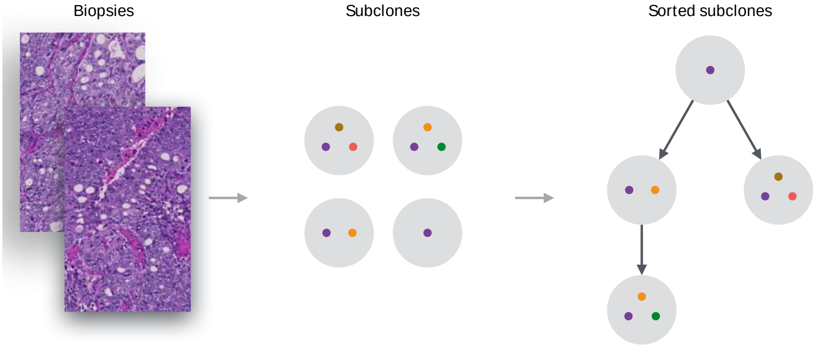

Figure 4: Depiction of the clonality inference problem from multiple biopsies solved in Chimaera.

Estimating the frequency of genetic alterations is complicated by the high genomic instability that characterizes many tumor types. Chimaera is an algorithm that allows an accurate detection and estimation of subclone frequencies using whole exome sequencing (WES) and whole genome sequencing (WGS) data from multi-area biopsies that enables the subclones to be sorted in an evolutionary tree structure.

The user provides mutation frequencies in terms of variant allele frequencies (VAF) for the different profiled tumor areas and obtains as an outcome the clonal composition of the studied tumor: mutation frequencies corrected by chimaera grouped in the different subclones.

# Chapter 3 SmartBiobank integration

In the frame of task 7.4 ABT finished the interface development of a fast and stable connection between SmartBiobank Dashboard (https://smartbiobank.astridbio.com/) and IBM Cloud services developed in PrECISE (Figure 5).

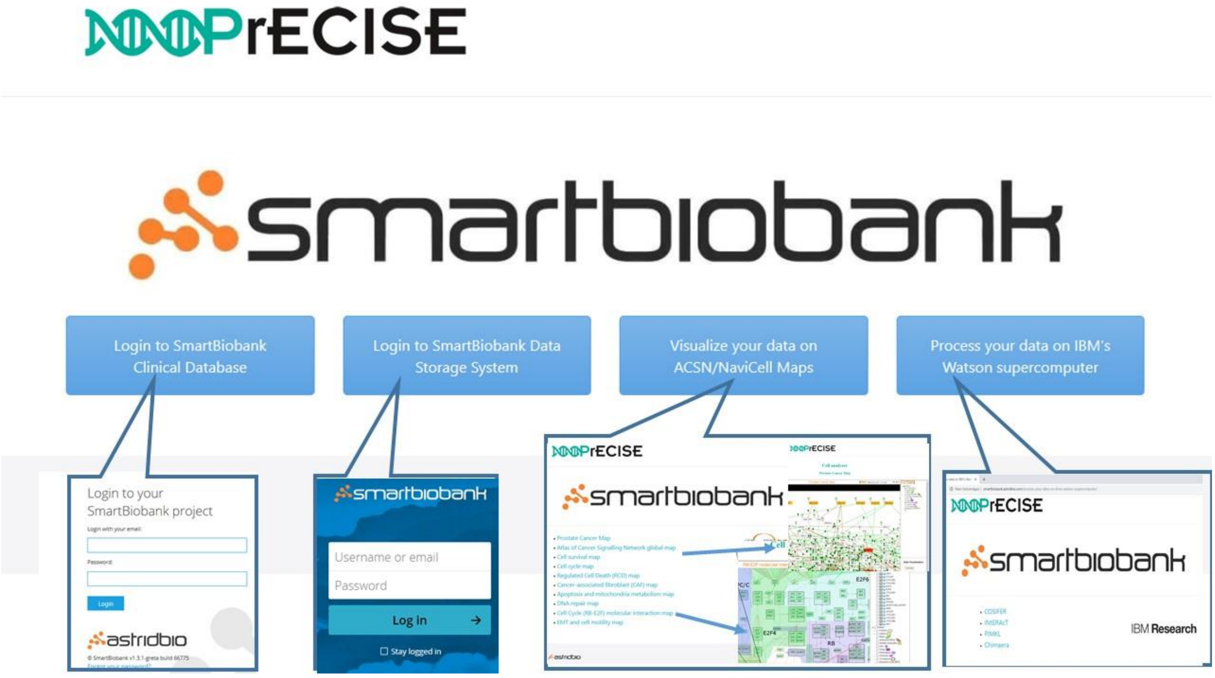
PrECISE
smartbiobank
Login to SmartBiobank Clinical Database
Login to SmartBiobank Data Storage System
Visualize your data on ACSN/NaviCell Maps
Process your data on IBM's Watson supercomputer
Login to your SmartBiobank project
Login
astridbio
smartbiobank
Username or email
Password
Log in
Stay logged in
PrECISE
smartbiobank
Prostate Cancer Map
Atlas of Cancer Signalling Network global map
Cell survival map
Cell cycle map
Regulated Cell Death (RCD) map
Cancer-associated fibroblast (CAF) map
Apoptosis and mitochondria metabolism map
DNA repair map
Cell Cycle (RB-E2F) molecular interaction map
EMT and cell motility map
E2F6
E2F4
RB
astridbio
PrECISE
smartbiobank
COSIFER
INtERAcT
PIMKL
Chimaera
IBM Research

Figure 5: SmartBiobank Dashboard with four integrated systems. Clinical Database, Data Storage System, ACSN/NaviCell Maps and IBM Cloud services.

During the workflow, ABT first created a new landing page for T7.4, but finally needed to extensively redesign the dashboard in strong collaboration with IBM's specialists in order to make it compliant with IBM's strict security policies. These security policies have been put in place especially to allow a safe way to manipulate and process sensitive data. Therefore, ABT finetuned the previously developed REST API named Cellanalyzer for use in Task 7.4. The API now allows communication with IBM's accessible services via HTTP requests. The developed REST API is now ready to support stable communication between the SmartBiobank's dashboard and the different IBM Cloud services (PIMKL, COSIFER, INtERAcT, Chimaera) (Figure 6).

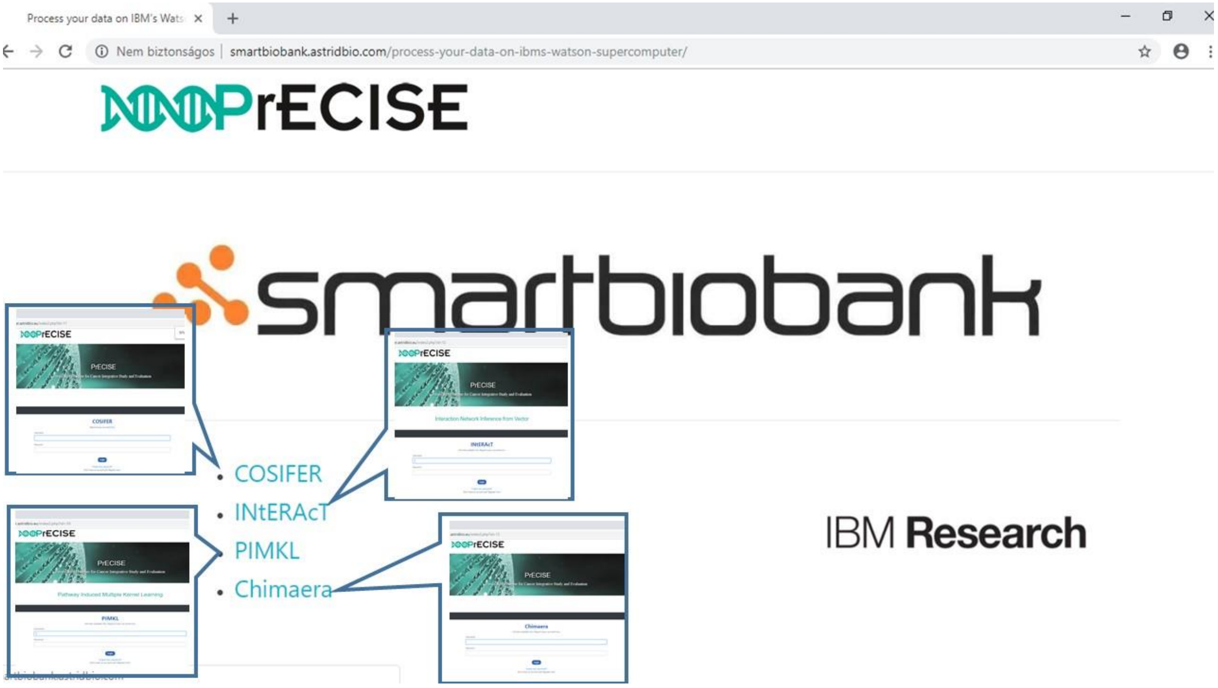
Process your data on IBM's Wats
Nem biztonságos | smartbiobank.astridbio.com/process-your-data-on-ibms-watson-supercomputer/
PrECISE
smartbiobank
COSIFER
INtERAcT
PIMKL
Chimaera
IBM Research
PrECISE
Interaction Network Inference from Vector
Pathway Induced Multiple Kernel Learning

Figure 6: The developed REST API is able to communicate with PIMKL, COSIFER, INtERAcT and CHIMAERA web based servers.

IBM’s platform is built on top of a Python based web framework which can be remotely controlled through a REST API. The platform's web-page has a special setting which is required by the IBM policy. It ensures that only the accepted web-pages can embed the original page. It is set to accept the "astridbio.eu" domain for embedding the original platform. This setting requires advanced browser security which currently can be provided by Google Chrome browser and Safari. Thus, Chrome is recommended for using the system. To manage the embedding ABT uses a PHP http client. On PHP code level we describe requests to the platform and also validate the currently used browser if it is acceptable.

# Chapter 4 Conclusion

In this deliverable we have presented the integration of the cloud services implemented by partner IBM into the SmartBiobank by partner ABT.

The ability to access directly from the SBB platform the services as embedded frame will represent an easy way for researcher to analyze the data stored in the system with the algorithms developed by the IBM researchers.

Moreover, our efforts pave the way to enable the integration in the SmartBiobank of other services in an analogous way, increasing enormously the potential of the platform implemented by ABT.

In conclusion, we are convinced that this integration represents a tangible outcome of PrECISE and will provide the community with an extremely valuable resource. These integrated services represent a fast and effective way, for non-computational researcher, to access state-of-the-art methodologies that can accelerate and improve cancer research in generating novel hypothesis and answer specific biological question leveraging multiple data types and sources of information.

# List of Abbreviations

| | |
|---|---|
| ACSN | Atlas of Cancer Signaling Network |
| API | Application programming interface |
| CI | Institute Curie |
| REST | Representational State Transfer |
| SBB | SmartBiobank |
| PIMKL | Pathway induced multiple kernel learning |
| COSIFER | Consensus network inference service |
| INtERAcT | Interaction networks from vector representation of words |
| MKL | Multiple kernel learning |