GRANATUM
THINK TANK FOR BIOMEDICAL PROGRESS
http://www.granatum.org/

LiSIs
Life Sciences Informatics System
http://lisis.cs.ucy.ac.cy/

GRANATUM-LiSIs: Making complex *in silico* predictive models accessible to wet-lab biologists

Ioannis Kirmitzoglou[1], Ioanna Kalvari[1], Christos C. Kannas[2], Kleo G. Achilleos[2], Zinonas Antoniou[2], Christos A. Nicolaou[2], Christiana M. Neophytou[3], Christiana Savva[3], David Scherf[4], Clarissa Gerhauser[4], Andreas I. Constantinou[3,*], Constantinos S. Pattichis[2,*], Vasilis J. Promponas[1,*]

[1] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[2] Department of Computer Science, University of Cyprus, Nicosia, Cyprus
[3] Cancer Biology and Chemoprevention Laboratory, Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[4] Cancer Chemoprevention and Epigenomics Workgroup, German Cancer Research Center, Heidelberg, Germany.

∗ Correspondence to: pattichi@ucy.ac.cy; andreasc@ucy.ac.cy; vprobon@ucy.ac.cy

Introduction

Chemoprevention research aims to finding drugs/natural substances to prevent the occurrence of a particular disease and elucidating their mechanism of action. The discovery of novel chemopreventive agents is severely hampered by the lack of high throughput assays to screen quickly and reliably promising chemical compounds.

We present LiSIs (http://lisis.cs.ucy.ac.cy), a platform in the context of an ongoing cross-disciplinary project (GRANATUM; http://www.granatum.org) aiming to bridge the gap between biomedical researchers by ensuring their seamless access to the globally available information needed to perform complex experiments and to conduct studies on large-scale datasets.

Scientific Workflows

Scientific workflows (SWs) are used to describe in abstract form the actions that need to be taken in order to complete a complex task. A SW is represented as a directed graph, where each node represents a step implemented by a software component (e.g. a local program or a remote web service). The graph edges represent either data flow or execution dependencies between nodes, coordinating the inputs and outputs of the individual steps, forming the data flow.

SWs provide a simple, yet powerful, environment and facilitate interdisciplinary collaborations by (i) sharing workflows and computational components, and (ii) jointly undertaking research initiatives requiring end-to-end scientific data management and computational analysis. Advances in grid technologies allow workflows to exploit parallel executions enabling large-scale data processing, where workflows are used as a parallel programming model for data-parallel applications. Moreover, web services allow ease of access to local and distributed data sources as well as data aggregation from highly heterogeneous environments.

The LiSIs Platform Aims

The Life Sciences Informatics System (LiSIs) aims to provide cancer chemoprevention experts with a set of online tools to create, update, store and share virtual screening SWs for the discovery of new chemopreventive agents. LiSIs is available via a web interface (http://lisis.cs.ucy.ac.cy) through a password protected, tiered login process, providing different level access to platform functionalities based on the user profile. Regular users are able to assemble SWs utilizing available *in silico* models and tools. “Power users” may build new models and tools through the development of custom SWs. Workflows execute on the system server and results are stored on the user's GRANATUM workspace, enabling accessing, manipulating or sharing SWs, datasets and results with other users.

The LiSIs platform is built on top of Galaxy ([1-3]; http://galaxy.psu.edu/). Galaxy is an open, web-based workflow system that comes preloaded with a big selection of tools designed for data intensive biomedical research. LiSIs utilizes the core elements of Galaxy to offer a series of essential capabilities, such as workflow and history sharing as well as shared data libraries (**Figure 1**). At the same time LiSIs takes advantage of the expandable nature of Galaxy to provide easy to use tools commonly implemented in drug discovery pipelines, for example chemical compound filters and property generators, 3D docking tools and predictive model generation and utilization.

Results & Discussion

LiSIs is a modular platform comprised of five major modules: input, pre-processing, processing, post-processing and results/outputs. Each module hosts a collection of component categories essentially implementing a variety of functionalities. A component category may offer different variations of the same functionality, for example enabling data input from either user-defined chemical and biological data files or from the GRANATUM Linked Biomedical Data Space.

In particular, the Predictive Models component enables building novel (Power users - **Figure 2**) or using existing (Regular users - **Figure 3**) data-driven models to predict biochemical properties of interest to the user for the selection of compounds with acceptable predicted properties. Such models fall into the category of Quantitative Structure – Activity/Property Relationship (QSAR/QSPR) models used by the drug discovery community to predict relevant properties of molecules.

The model building process is assisted by a custom developed “Hierarchy of Cancer Chemoprevention Properties/Activities”. In brief, this light ontology-like effort aims to make available to modeling experts possible (independent) ways by which a substance may act as a cancer chemopreventive agent.

During model construction (**Figures 2 & 4**), a list of training set compounds with their respective property values is provided as input, along with the settings for the selected predictive modeling application. The output includes the trained predictive model and a log file containing measures of the quality of the model estimated by cross-validation or other appropriate techniques. In the model usage phase (**Figure 3)** a list of test compounds is provided as input to a specific predictive model. The output consists of a list of compounds with their associated predictions and a log file documenting the results.

The Predictive Models component currently makes use of four popular machine learning algorithms widely used by the chemoinformatics community for predictive modeling, namely: Decision Trees, Random Forests, Support Vector Machines and k-Nearest Neighbours. Implementations of the aforementioned algorithms are available by interfacing the CARET package [4] for the open source R Environment for Statistical Computing (R Development Core Team, 2012; http://www.r-project.org/). The modular architecture of the system facilitates future extensions, with the possibility of adding other appropriate predictive modeling algorithms.

As part of the platform, LiSIs offers a series of built-in predictive models of some biological/chemical properties commonly employed as filters in drug discovery. **Table 1** displays the key characteristics of these models as well as their performance compared to selected studies. It is worth noting that the models described are the result of a thorough process of model generation using a variety of descriptors/fingerprint and learning algorithm combinations and a detailed analysis of the performance of the resulting models.

Acknowledgments

This work has been partially supported through the EU-FP7 GRANATUM project, contract number ICT-2009.5.3. We gratefully thank our partners from the members of the GRANATUM consortium for their crucial role in this work.

Πανεπιστήμιο Κύπρου University of Cyprus
dkfz.
GRANATUM THINK TANK FOR BIOMEDICAL PROGRESS
SEVENTH FRAMEWORK PROGRAMME

Read me later!
http://sdrv.ms/14WcZos

References

1. Blankenberg D *et al.*, *Current Protocols in Molecular Biology*. 2010; Chapter 19:Unit 19.10.1-21.
2. Giardine B *et al.*, *Genome Research*. 2005; 15(10):1451-5.
3. Goecks, J *et al.*, *Genome Biology*. 2010;11(8):R86.
4. Kuhn M, *Journal of Statistical Software*. 2008; 28(5).
5. Fjodorova *et al.*, *Chemistry Central Journal.* 2009; 4(Suppl 1):S3
6. Cassano *et al.*, *Chemistry Central Journal.* 2009; 4(Suppl 1):S4
7. Hansen *et al.*, *Journal of Chemical Information and Modelling*. 2009; 49(9):2077-2081
8. Roncaglioni A *et al.*, *SAR and QSAR in Environmental Research*. 2008; 19(7-8):697-733.
9. Delaney J, *Journal of Chemical Information and Modelling*. 2004; 44(3):1000-1005

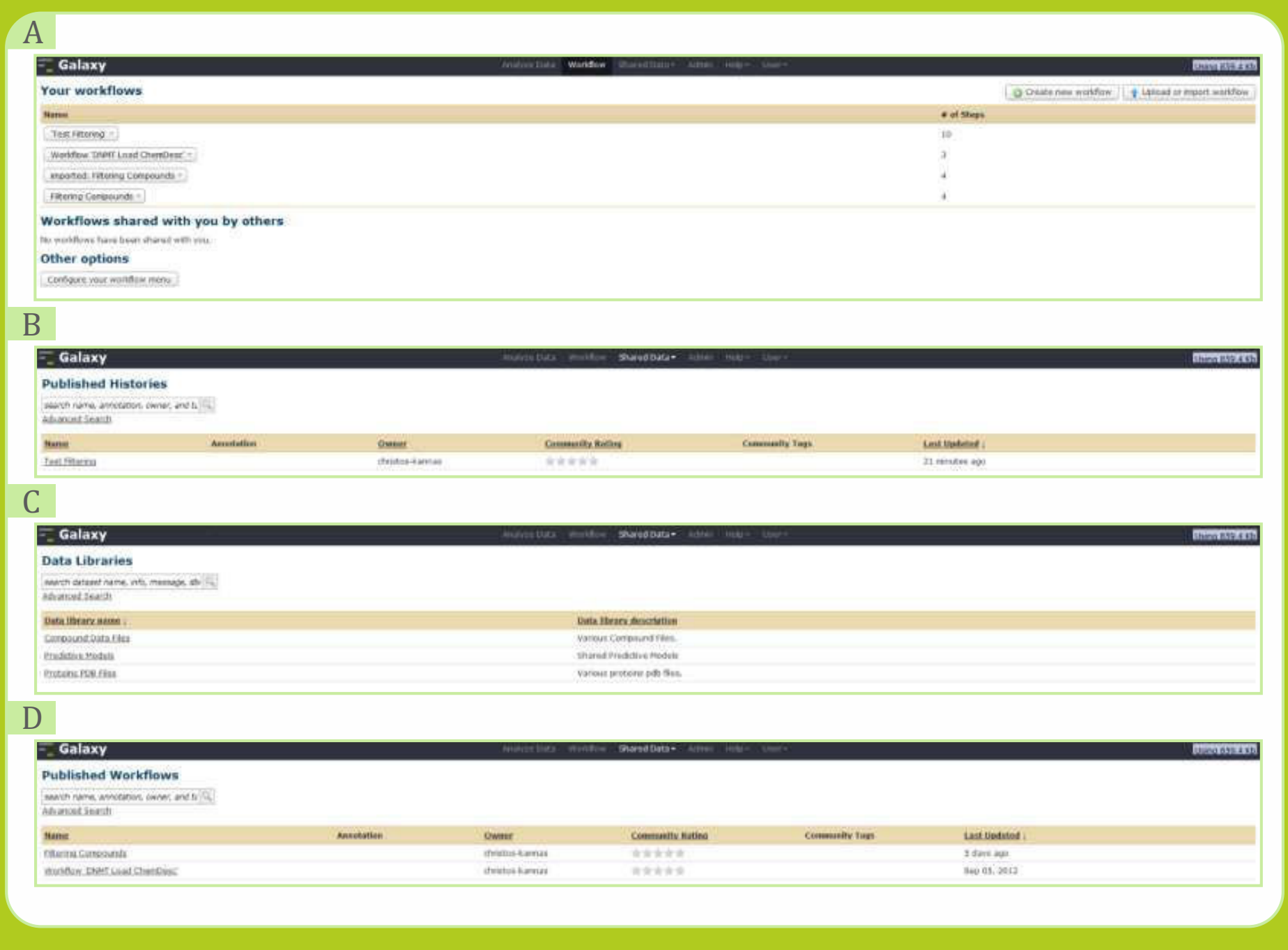


Figure 1

Core Galaxy functionalities exploited by the LiSIs platform:
A. Users may access available workflows, i.e. workflows they had created or published by other users. These workflows are, strictly speaking, the abstract workflows which specify the computational steps of an *in silico* experiment.
B. Users also have access to available histories, which (in the Galaxy jargon) refer to workflows accompanied with the actual data of a specific run. Histories enable reproducibility of earlier *in silico* experiments, error checking/debugging or inspection of intermediate computational results.
C. The data libraries panel provides access to data already uploaded on the Galaxy/LiSIs platform. The system is fully customizable (at the administrator or power user level) both in terms of content (e.g. supported file types and their associated content) and access levels.
D. Users of all levels are enabled to publish workflows giving free access to all authenticated users on the system. A specific panel facilitates access and search for published workflows meeting specific criteria.

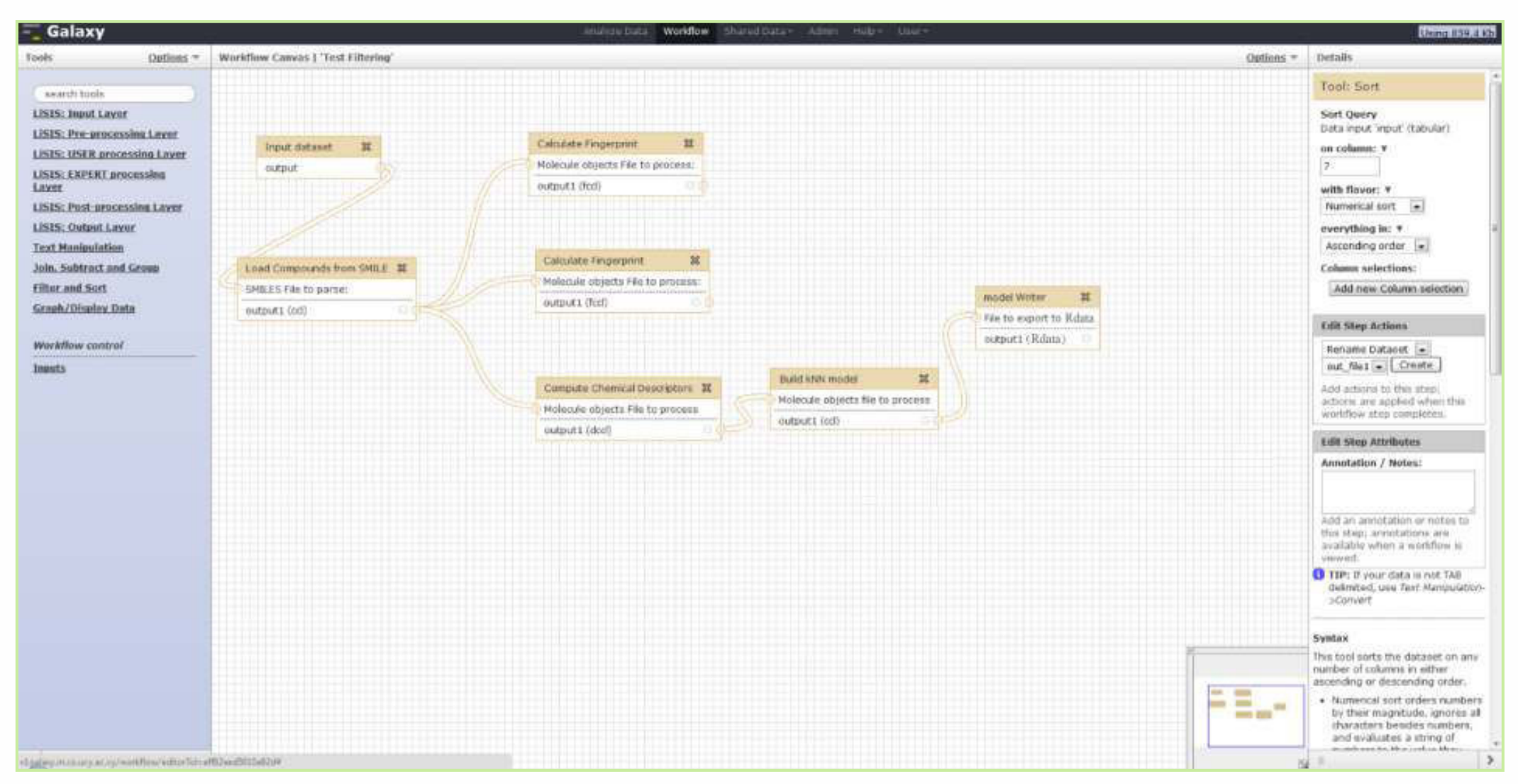


Figure 2

An example workflow of a predictive model creation:
The expert user (EU) has to define a series of parameters such as the input file, the descriptors to be calculated and the classifier. In this specific example the EU has chosen the k-Nearest Neighbours classifier with its default parameters (*build kNN* node). The *build kNN* tool automates the procedure of selecting the best *k* value based on an EU selected metric such as accuracy, sensitivity or specificity. The EU can review the final model by examining the output of the tool which also contains cross-validation data. At the final step of the workflow the model is saved and exported to GALAXY's public libraries, ready to be used by the normal users through the *Predict* tool.

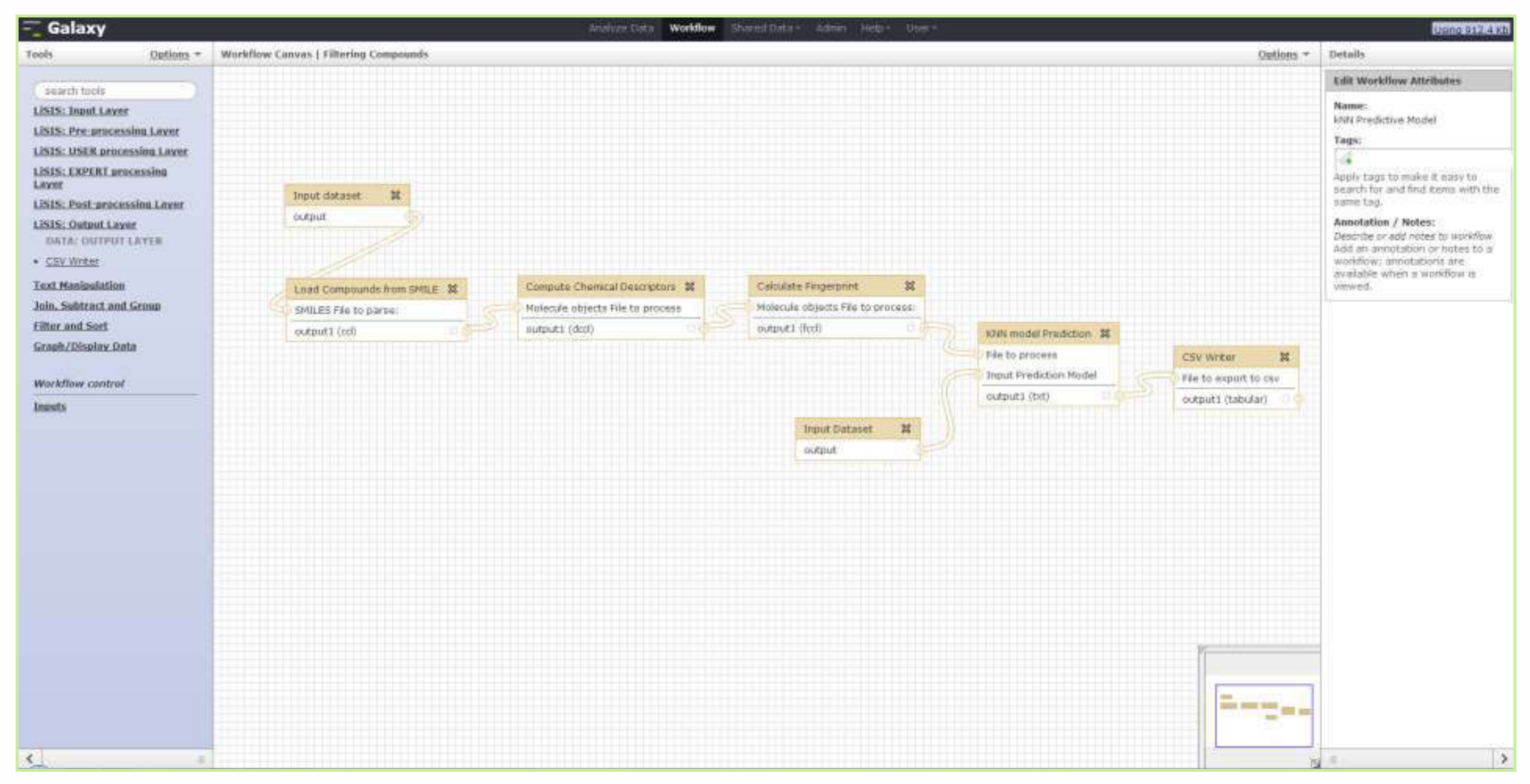
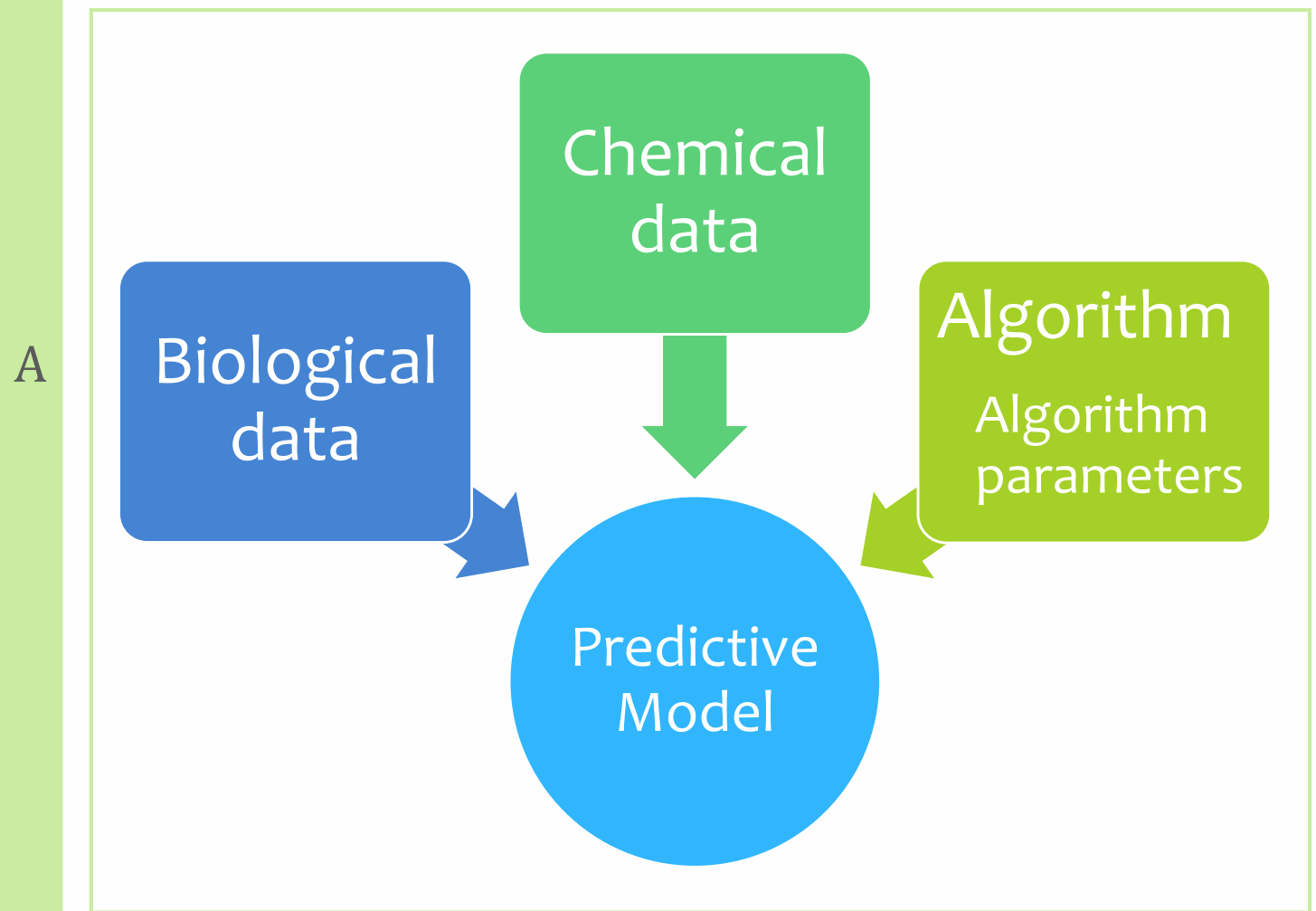
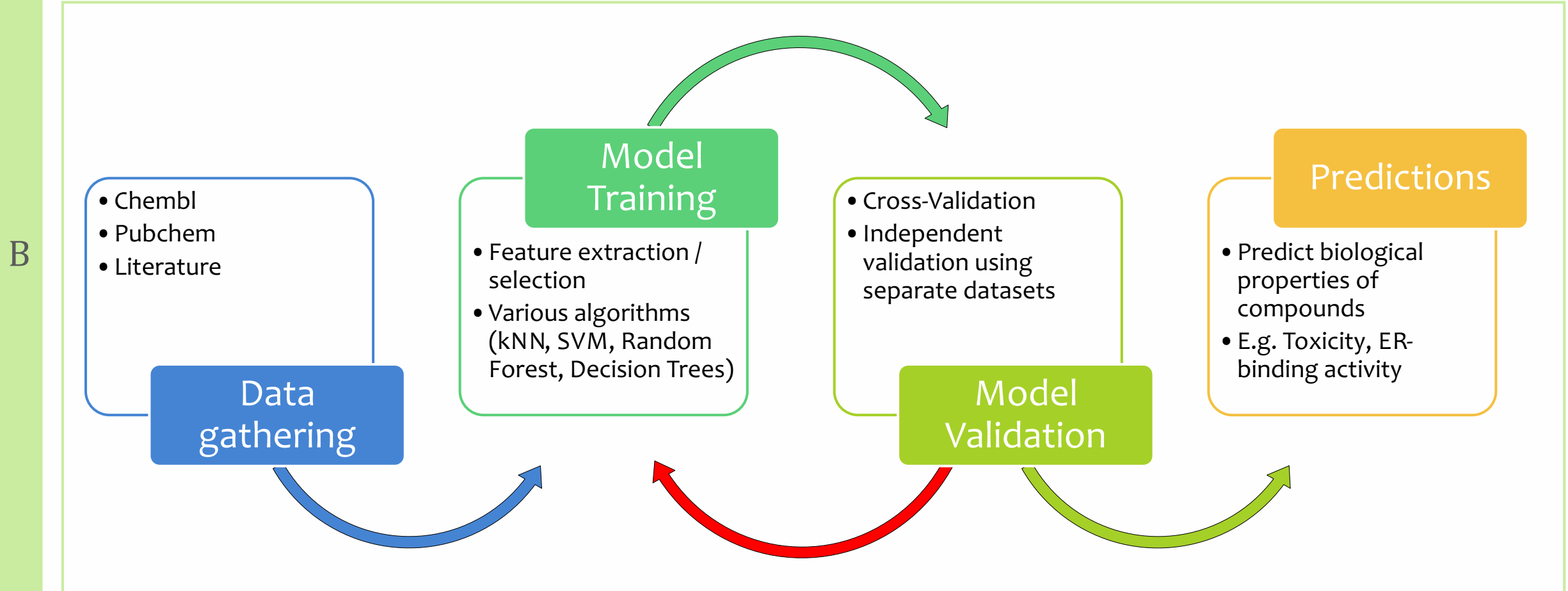


Figure 3

Utilizing an existing predictive model:
The initial steps required to use a predictive model (compound loading and chemical descriptors generation) are identical to those used in the model building phase. Importantly, the LiSIs platform intelligently provides automatically parameter selections compatible to the predictive model to be used in the downstream computation. This feature not only ensures the correct execution of the models but also makes the prediction process transparent to the regular user. The output of the *Predict* tool is a list of predicted biological properties for the input molecules which can be post-processed by special tools provided within LiSIs. Functionalities such as sorting or selecting compounds with specific predicted properties enable further analyses or report generation.

Figure 4




B
Chembl
Pubchem
Literature
Data gathering
Model Training
Feature extraction / selection
Various algorithms (kNN, SVM, Random Forest, Decision Trees)
Cross-Validation
Independent validation using separate datasets
Model Validation
Predictions
Predict biological properties of compounds
E.g. Toxicity, ER-binding activity

General procedure for the creation of a predictive model:
A. For the creation of a predictive model through the LiSIs platform an expert user (EU) must define 3 key components:
i) Data containing biological properties of the compounds that will be used to train the model. These properties are usually obtained through wet-lab experiments and can either be a true/false statement (e.g. Toxic/Non-toxic) or a real value (e.g. IC50 values of binding affinity to a specific target).
ii) Data describing the chemical properties of the compounds. These descriptors can either define simple chemical properties of a compound (e.g. molecular weight, number of rings and molecular complexity) or can be derived by the actual structure of the compound in ways that also take into account pharmacophore features (e.g., hydrophobic centroids, aromatic rings and hydrogen bond acceptors or donors) or a suitable combination. The LiSIs platform has the ability to automatically calculate a variety of descriptors for a list of chemical compounds.
iii) The configuration parameters for the model's classifier. Currently, the LiSIs platform supports choosing among different classification algorithms (e.g., k-Nearest Neighbours or Support Vector Machine) as well as a series of parameters related with model training and cross-validation.

B. For the creation of a good model every part of the process is crucial:
Biological properties data must be of high quality. Chemical descriptors should be selected and tuned for each specific model through a semi-automated feature selection process. Selecting the right classifier (along with its parameters) is also a crucial and laborious process. The whole process of feature selection and parameter tuning is guided by the cross-validation results of each intermediate model. The best intermediate model is also independently validated against an independent data-set (i.e., data not available to the model during the training phase). The LiSIs platform contains a series of custom tools built to partially automate, and thus accelerate, the creation and validation of predictive models.

Table 1

Model Name	Description	Method	Status	Sens / Spec LiSIs	Sens / Spec Literature
Carcinogenicity	Predict carcinogenic potency	Random Forests	Operational, published	0.70 / 0.67	0.66 / 0.62 [5]
Developmental Toxicity	Predict developmental toxicity	Random Forests	In development, published	0.600 / 0.82	0.95 / 0.59 [6]
Mutagenicity	Predict mutagenicity potential	Random Forests	Operational, published	0.82 / 0.80	0.83 / 0.75 [7]
ERα binding	Predict ERa binding activity potential	Random Forests	Operational, published	0.72 / 0.91	0.81 / 0.92 [8]
DNMT binding	Predict DNMT binding activity potential	Linear SVM	In development	0.94 / 0.34	—
Solubility	Predict compound solubility	Regression equation	In development	—	—

Predictive models built-in the LiSIs platform:

LiSIs currently offers six built-in predictive models while more are constantly developed. All the models where validated using 10 independent experiments of leave 25% out cross-validation.

Published models are the result of a thorough process of model generation using a variety of descriptors/fingerprints and learning algorithm combinations and a detailed analysis of the performance of the resulting models.

Carcinogenicity
Dataset source: Fjodorova *et al.* [5].
Number of compounds: 805 (421 carcinogenic, 384 non-carcinogenic.)

Developmental Toxicity
Dataset source: Cassano *et al.* [6].
Number of compounds: 292 (116 toxicant, 176 non-toxicant)

Mutagenicity
Dataset source: Hansen *et al.* [7].
Number of compounds: 6444 (3503 mutagenic, 3009 non-mutagenic)

ERα binding activity
Dataset source: Roncaglioni *et al.* [8]
Number of compounds: 802 (286 binders, 516 non-binders)

DNMT binding activity
Dataset source: A collection of compounds gathered by several Chembl arrays and publications, compiled by the Cancer Chemoprevention and Epigenomics Workgroup, German Cancer Research Center, Heidelberg, Germany.
Number of compounds: 258 (221 binders, 37 non-binders)

Solubility
Implementation of a regression equation described by Delaney [9]

We also compare the results of the LiSIs predictive models to those from the respective publications, observing that our results are usually in par or better than the current state-of-the-art. Apparently our results illustrate that in most of the cases Random Forests models perform better that the rest of the tested algorithms.