ExCAPE: Ex*ascale* C*ompound* A*ctivity* P*rediction* E*ngines*

EXDCI Event

Prague, May 2016

09/05/2016

Author:Tom Ashby and ExCAPE consortium

ExCAPE: in a nutshell

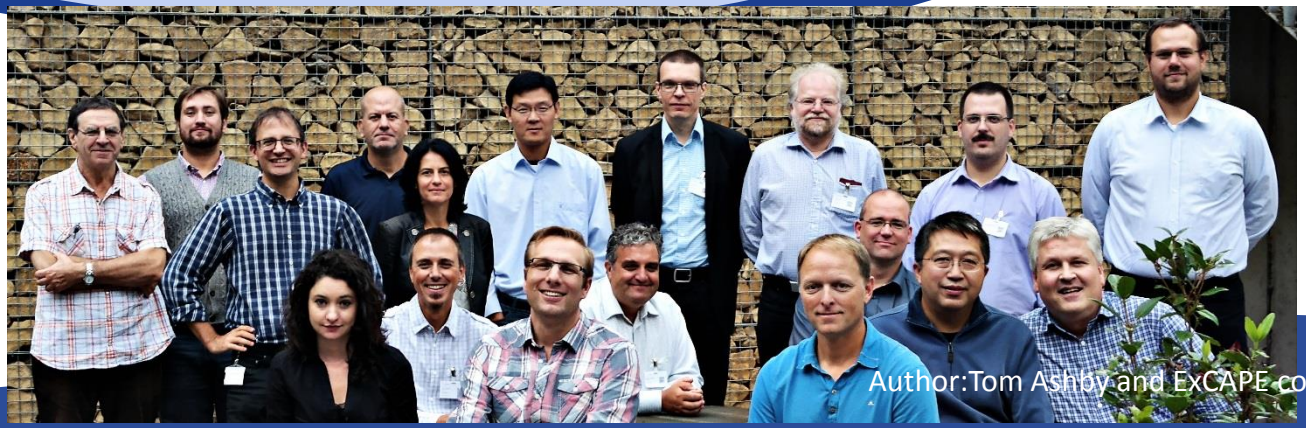
- Exascale machine learning (ML)
  - Machine learning is already a very large consumer of cycles in data centres
  - **New algorithms** need to be developed to turn extra available computation into **more accurate**, **more useful** models
  - Applying analysis and scaling techniques from HPC will help **exploit maximum potential** of these algorithms
    - Enable eventual move to exascale
  - We aim to open up **new areas** for application of HPC expertise with important societal benefits

- **Our proposal**
  - *Apply **Exascale machine learning** to problems in the **pharmaceutical industry***
    - (But techniques applicable to other uses of ML)

ExCAPE: in a nutshell

- The big picture: *the constituent threads*...
  - Exascale supercomputers
    - Large clusters
    - Accelerators
  - Machine learning
    - Supervised and unsupervised learning
    - Confidence estimation, dyadic data etc.
    - Learning performance (accuracy of model)
  - Computation in the Pharma Industry
    - Chemogenomics
    - ADMET

The Partners



Author:Tom Ashby and ExCAPE consortium

ExCAPE - Progress

- **Codes** set up on Salomon@IT4I
  - Deep learning from U.Linz: *binet*
  - Matrix factorization (collaborative filtering) from Imec
  - Scale-out chemogenomics from Janssen Pharmaceutica (on-going)
- **Data sets** on Salomon@IT4I, available for running experiments
  - Tox datasets
  - Chembl + PubChem activity datasets
    - Fingerprint generation etc

ExCAPE - Progress

- **Programming models**
  - First prototype for **scheduling** jobs onto Salomon@IT4I
  - Report done on possibilities to apply **optimisation techniques** to (collections of) algorithms currently considered in the project

- **Algorithms**
  - First discussions have taken place on which algorithms to focus on
    - Deep learning, Group Factor Analysis, Inductive Conformal Prediction

- **Offering to Ecosystem**
  - An *application* and *algorithmic techniques* that are on the edge of the areas covered by traditional HPC
    - Machine learning for life sciences
    - Software and techniques to use
  - A view on how such applications will interact with current and future HPC systems
    - Both hardware and software

- **International collaboration**
  - Progress
    - Inital contact (through Intel) with the ***K supercomputer-based drug discovery (KBDD)*** consortium (Biogrid Pharma Consortium), Japan
      - Presentation and discussion on deep learning infrastructure used by U.Kyoto
  - Help from EXDCI?
    - Possible projects to connect with:
      - RaPyDLI (US: Geoffrey Fox, Jack Dongarra etc)
    - News about other *Machine Learning on HPC* projects...

- **cPPP (ETP4HPC + CoEs)**
  - Current
    - Discussions with *Bioexcel CoE*: looking for application synergy
    - CompBioMed: looks interesting ☺
  - Plans
    - Set up something more concrete with *Bioexcel*
    - Look for hardware and systems partners interested in results on e.g. large scale collaborative filtering

- **Extreme Scale Demonstrators**
  - We are a possible ***application/software*** partner
  - Exciting application: **big data**, machine learning, **pharma industry** etc.
  - **Porting** of application to a demonstrator platform, and **benchmarking**
  - Feedback from potential **industrial end users**

Backup Slides

09/05/2016

Author:Tom Ashby and ExCAPE consortium

ExCAPE: Overview

- The big picture: *combining these!*
  - How should we put the ML algorithms on supercomputers?
    - Programming model, library support, use of accelators...
  - Which ML algorithms will work best for the pharma problems? On which (sets of) datasets?
    - Modelling accuracy
    - Model usefulness (confidence, interpretability)
    - Combining data sets from partners and public data sets
  - Which pharma problems need the most computation?
    - Scalability of the algorithms
    - Complexity of data sets and modeling problems

Drug design and development

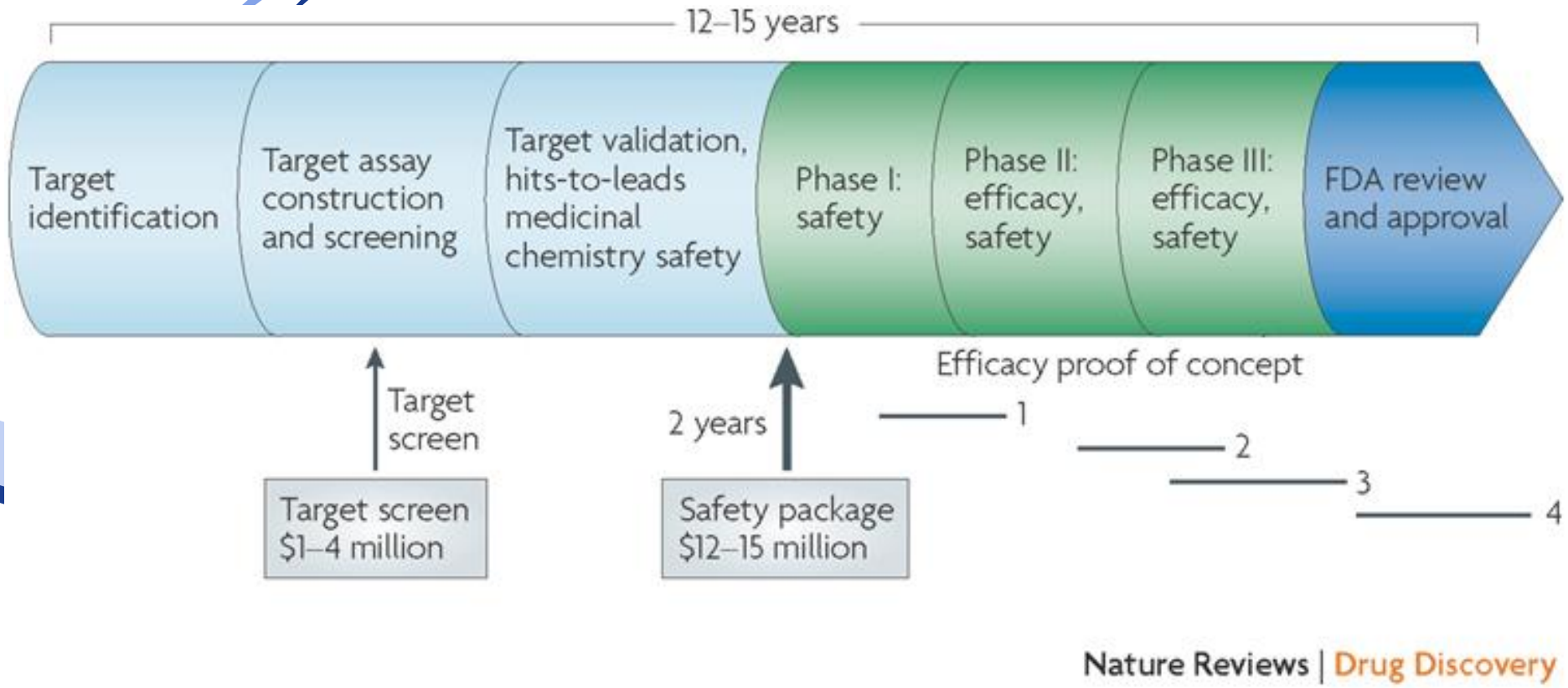
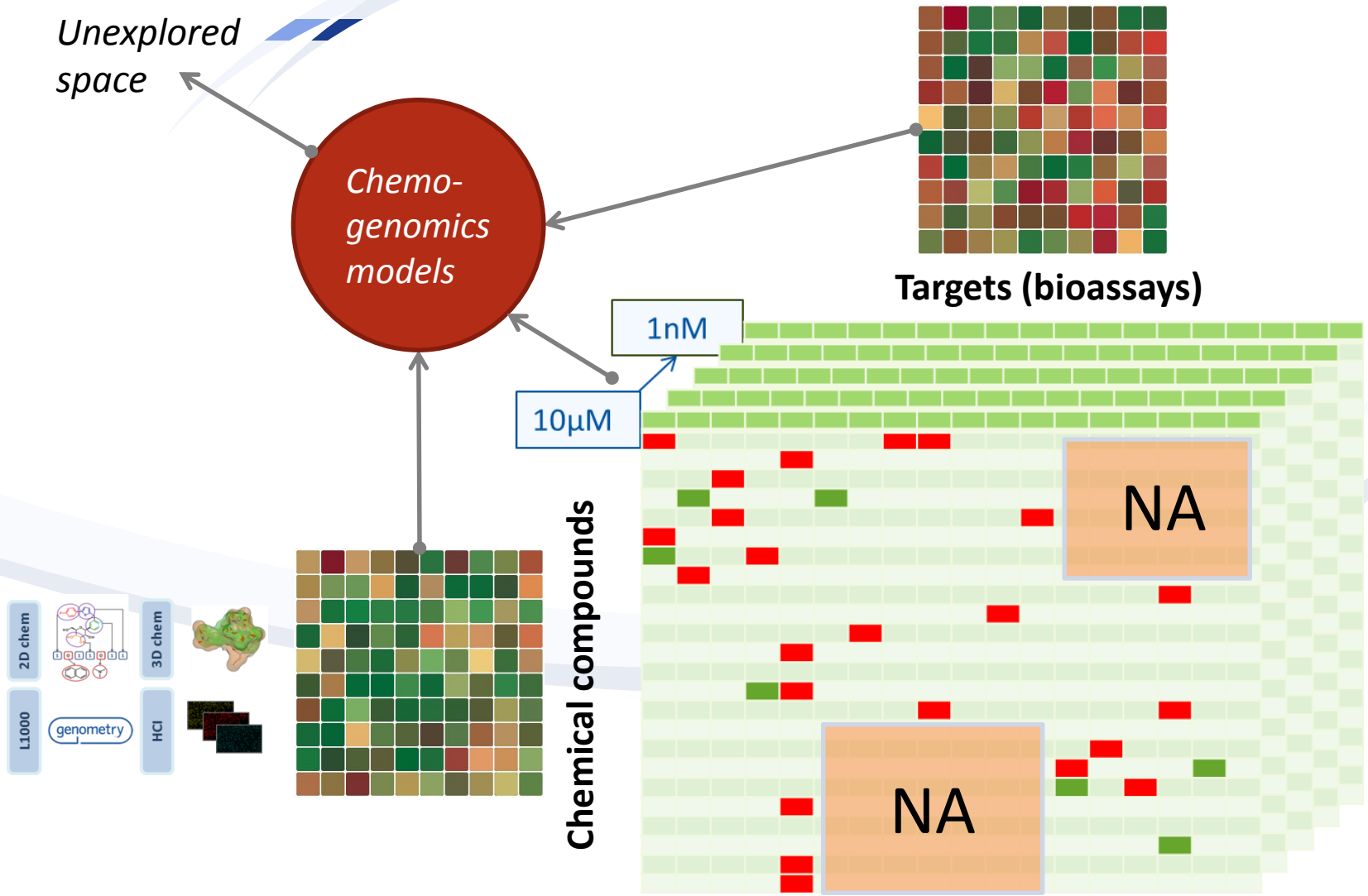


Image from doi:10.1038/nrd2593

Use-case: Chemogenomics



ExCAPE: Workpackages

- Four WPs
  - **WP1**: Exascale Machine Learning Algorithms
  - **WP2**: Programming models and scalable efficiency
  - **WP3**: Benchmarking and validation of machine learning algorithms
  - **WP4**: Management, Exploitation, Dissemination and Communication
- All WPs go from **M1** to **M36**

Opportunities for Collaboration?

- HPC Centres of Excellence
  - **BioExcel**: synergy at the application level
  - **POP**: expertise on parallelism
- FET-HPC projects
  - Hardware:
    - **NEXTgenIO**: Feedback on I/O behaviour
    - **EXTRA**: FPGA tools and expertise
    - (*MANGO*: many core acceleration)
  - Prog models:
    - **READEX**: insight into application phases
  - Algorithms:
    - **NLAFET**: use of NLA libraries to do ML
- PRACE
- ETP4HPC