The ELF Honest Data Broker: informatics enabling public–private collaboration in a precompetitive arena

Guillaume Paillard[1], Philip Cochrane[3,4], Philip S. Jones[2], Willem P. van Hoorn[5], Andrei Caracoti[1], Herman van Vlijmen[3] and Andrew D. Pannifer[2]

[1] Dassault Systemes, BIOVIA Ltd., 334 Cambridge Science Park, Cambridge CB4 0WN, UK
[2] University of Dundee, Biocity Scotland, Bo'ness Road, Newhouse ML1 5SH, UK
[3] Janssen EMEA, 2340 Turnhoutsweg 30, 2340 Beerse, Belgium
[4] Finia Consulting Ltd, Unit 8 Dock Offices, Surrey Quays Road, London, SE16 2XU, UK
[5] Ex Scientia Ltd, Dundee Incubator, James Lindsay Place, Dundee, DD1 5JJ, United Kingdom

New precompetitive ways of working in the pharmaceutical industry are driving the development of new informatics systems to enable their execution and management. The European Lead Factory (ELF) is a precompetitive, 30-partner collaboration between academic groups, small–medium enterprises and pharmaceutical companies created to discover small molecule hits against novel biological targets. A unique HTS screening and triage workflow has been developed to balance the intellectual property and scientific requirements of all the partners. Here, we describe the ELF Honest Data Broker, a cloud-based informatics system providing the scientific triage tools, fine-grained permissions and management tools required to implement the workflow.

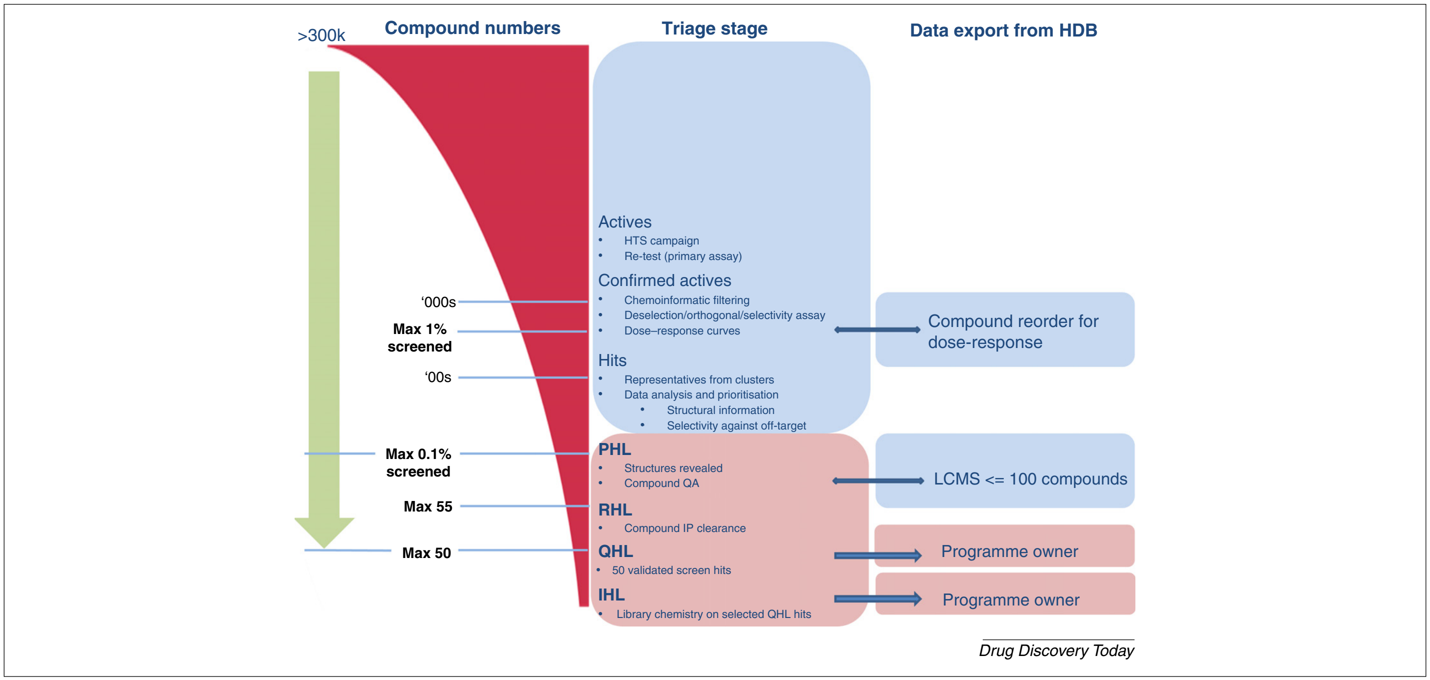
Introduction

The European Lead Factory (ELF) [1] is an Innovative Medicines Initiative (IMI) public–private partnership created to bring pharmaceutical companies, small–medium enterprises (SMEs) and academic laboratories together and stimulate precompetitive drug discovery. The ELF has built an HTS library [the Joint European Compound Library (JECL)] [2] from subsets contributed by the participating pharmaceutical industry partners and through ongoing synthesis within the ELF consortium [3]. This library is available to academic groups and SMEs across Europe via the ELF screening, characterisation and medicinal chemistry infrastructure at the European Screening Centre. The goal of the academic and SME screening campaign is to discover high-quality, well-validated small molecules against novel biological targets that can serve as tool compounds to validate targets or directly form the foundation of a drug discovery project. The entire screening library is also distributed to the seven pharma companies who contributed compounds to the ELF. This enables the companies to screen their own internal programmes against a diverse mid-sized HTS collection far larger than the subsets they contributed and gives access to chemical space unrepresented in their own libraries.

A key consideration in creating the ELF was an intellectual property (IP) model that balances and safeguards the interests of all the parties concerned and enables sustainable partnership. Organisations proposing their biological target for screening (programme owners) require confidentiality for academic or business reasons. Organisations contributing compounds (compound owners) might not want all of the compound structures available for immediate general inspection or analysis because they represent a considerable body of IP. It is also crucial that the hit molecules and SAR emerging from the screening process remain confidential to each programme to enable future development. Conversely, enough information must be made available to scientists carrying out the triage to execute a high-quality compound selection and quality control of screened compounds. To achieve this balance, a workflow has been created that regulates access to compound structures and data using a set of fine-grained permissions and

Corresponding author:. Pannifer, A.D. (a.pannifer@dundee.ac.uk)

http://dx.doi.org/10.1016/j.drudis.2015.11.005



Drug Discovery Today

FIGURE 1

The European Lead Factory (ELF) triage process. The Joint European Compound Library (JECL) is screened against each target and triaged to a final output of up to 50 compounds. During the triage process, a number of formal stage gates are passed [screening result list (SRL), preliminary hit list (PHL), revised hit list (RHL), qualified hit list (QHL) and improved hit list (IHL)] and access to compound structures is controlled depending on a triage team member's permissions [programme team (PT) or programme clearance team (PCT)] and the stage gate. All data stay entirely within the workflow with very little export permitted until release of the final 50 compounds at the end of the process. The only exceptions to this rule are compound ID export for reordering and physicochemical property export to enable LC–MS analysis of up to 100 selected compounds. Monoisotopic mass is required for interpretation of the spectrum, aromatic ring count can assist in cases where there is no UV signal and log *P* can help identify compounds that are not flying in the spectrometer. Even here, no structures are released. Parts of the triage process where structures are blinded are shaded blue and those parts where the structures are visible are shaded pink.

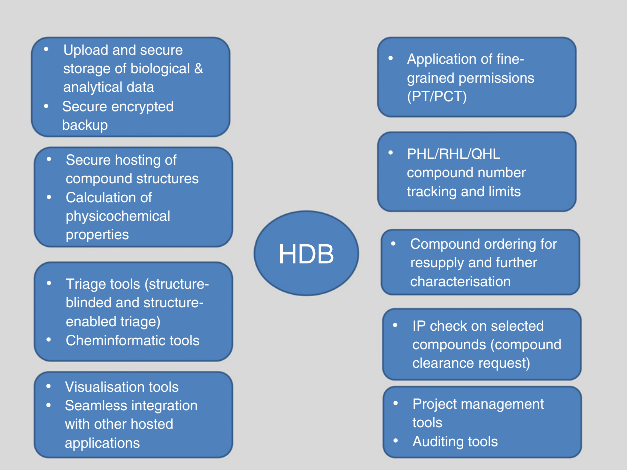
business rules. Fig. 1 gives an overview of the workflow and a glossary of acronyms is provided in Box 1. The entire JECL is screened against each target at a single point to generate the screening result list (SRL). Hits in the SRL are selected by members of the programme team (PT) and confirmed at single point based on biological activity and up to 1% of the SRL can be reordered for further biological characterisation, such as EC_{50} measurement. At this stage, no compound structures are visible; the PT is responsible for triaging the hundreds or thousands of hits to a maximum of 0.1% of the number of compounds on the SRL using an array of cheminformatic tools and biological data. Once the list of up to 0.1% of the SRL is created, these compounds are registered as the preliminary hit list (PHL) and responsibility for the triage passes to the programme clearance team (PCT). For public programmes, the PCT consists of a small team of medicinal chemists within the ELF, whereas for programmes belonging to the seven pharmaceutical companies each has appointed an expert delegate to carry out this role. At this point, structures of the PHL compounds are unblinded and the PCT selects up to 55 compounds to create the revised hit list (RHL). This process enables the JECL to remain largely confidential with only a small number of structures revealed in any programme. These RHL compounds are subject to a final check for IP entanglement by the compound owner through the compound clearance request (CCR) system. The compound owner has no information about the biological target or the programme owner when making this check. Once the IP check is cleared, compounds are then registered on the qualified hit list (QHL) and can be released to the programme owner. For public programmes, there is an opportunity to access further characterisation, medicinal chemistry and modelling capabilities at the ELF to validate and improve the QHL hits further and generate an improved hit list (IHL). The selection of compounds at the PHL, RHL and QHL stages is irreversible in order to eliminate the possibility of PCTs accessing more than the agreed number of structures.

BOX 1

Glossary of acronyms used in the European Lead Factory

ELF: European Lead Factory.
HDB: Honest Data Broker.
IHL: improved hit list – compounds resulting from the medicinal chemistry programme to develop selected QHL compounds.
PCT: programme clearance team – group or person responsible for triage from PHL to QHL.
PHL: preliminary hit list – 0.1% of the SRL selected for visualisation of structures.
PO: programme owner – the group proposing the biological target for screening.
PT: programme team – group responsible for initial stage of triage up to PHL stage.
QHL: qualified hit list ≤50 compounds for release to programme owner.
RHL: revised hit list ≤55 compounds sent for IP check.
SRL: screening result list – HTS output (300 000+ compounds).

The complex system of rules governing access to compound structures, requirements for IP checks, auditing requirements and


Upload and secure storage of biological & analytical data
Secure encrypted backup
Secure hosting of compound structures
Calculation of physicochemical properties
Triage tools (structure-blinded and structure-enabled triage)
Cheminformatic tools
Visualisation tools
Seamless integration with other hosted applications
HDB
Application of fine-grained permissions (PT/PCT)
PHL/RHL/QHL compound number tracking and limits
Compound ordering for resupply and further characterisation
IP check on selected compounds (compound clearance request)
Project management tools
Auditing tools

Drug Discovery Today

FIGURE 2

A summary of the functionality in the Honest Data Broker (HDB). The application supports the full workflow of the European Lead Factory (ELF), providing the triage tools to enable scientists to select the right compounds while applying the rules of the workflow to safeguard intellectual property. Project management tools are also provided together with auditing capability to track the destinations of compounds.

the importance of accurate compound selection within the compound visibility rules created a unique cheminformatics challenge. This has driven the development of a new application to support the entire process: the ELF Honest Data Broker (HDB). The HDB application is hosted on the BIOVIA ScienceCloud platform and provides four key capabilities: (i) fine-grained permissions to support the workflow rules; (ii) security features to maintain confidence in IP protection amongst all the ELF partners; (iii) scientific tools to execute a high-quality HTS triage; and (iv) tools to manage and audit programmes. The functionality of the HDB is summarised in Fig. 2. The ScienceCloud platform was chosen to host the HDB for a number of reasons. A cloud-based solution simplifies deployment and maintenance, ScienceCloud has the high level of security and availability required by the project, it is based on the Pipeline Pilot® platform which provides extensive flexibility in customising the system with the required permissions and additional triage tools, the pharmaceutical industry partners were all users of Pipeline Pilot® and familiar with the chemistry tools and there is access to the growing set of other applications hosted on ScienceCloud.

Permissions

The rules governing access to the different IP-related objects and information are a key aspect of the ELF workflow. Compound structures and ownership information are controlled to restrict access to structural knowledge and to ensure that compound selection is made independently of the origin of the compound, thus protecting the IP of the compound owners. To address all these aspects, a system of fine-grained permissions has been created. All members of a team associated with a programme are assigned either a PT or PCT role. PT members are responsible for initial triaging of the SRL. This is done without access to structures and requires a suite of cheminformatic tools to enable an effective compound selection. PCT members can see the structures of compounds registered on the PHL to provide a ‘chemist’s eye’ check on selected compounds and are able to order compounds for quality assurance (QA) analysis. Compound origin remains hidden throughout the process by a double anonymisation of the compound ID and of the compound owner, this information only being disclosed when compounds on the QHL are released to the programme owners.

In addition to the roles associated with a target programme, further roles are defined to enable execution of the triage process. The CCR permission is held by compound owners and enables a member of this group to see compounds sent to them through the CCR system and clear compounds without IP entanglement. Clearing compounds is performed without knowledge of either the target being screened or the identity of the programme owner. To avoid conflict of interest on targets, compound owners have the ability to prevent their compounds from being screened on a particular target. Other permissions are dedicated to the management of the compound libraries and to maintain the integrity of the overall process. The logistics role is assigned to scientists operating the compound store and allows access to lists of compounds ordered by PT or PCT members. The administrator role is held by a small number of people in the ELF and they are not associated with any programme. This role allows programme setup, supervision, tracking and auditing of information held within the HDB.

Security

The HDB provides all necessary security levels related to data storage and authentication. Data are securely stored and processed in an isolated Oracle® database located in a private cloud. All data in transit are secured with SSL/HTTPS with 2048 bit extended validation certificates; and backups and offline data are encrypted. For data access, the fine-grained data security model allows a role-based access-right definition to the different types of data. High availability of the system is reinforced by the disaster recovery dual instance of ScienceCloud and the HDB.

Triage tools

The requirement that data and compound structure information cannot be exported from the HDB during the triage process requires an effective suite of tools within the system to perform compound selection. Furthermore, the structure-blinded first stages of the process put particular emphasis on cheminformatic tools to allow optimal selection of a few hundred compounds for the PHL from over 300 000 on the SRL. Some key triage tools are listed in Table 1. A principal aim of these tools is to allow PT members to classify compounds by chemical similarity or scaffold to sample the diversity of hit series fully on the SRL and progress attractive series to the PHL. Avoidance of simple potency-based selection in which the PHL might be dominated by a few series and exclude interesting but slightly less-potent series is crucial [4]. All compounds in the HDB are assigned a Murcko scaffold ID [5] and a second scaffold ID based on the ScaffoldTree algorithm [6], adapted at UCB. This groups molecules by the first two-ring scaffold found in the iterative ScaffoldTree trimming process, or by the smallest scaffold if no two-ring scaffold is found. These

TABLE 1

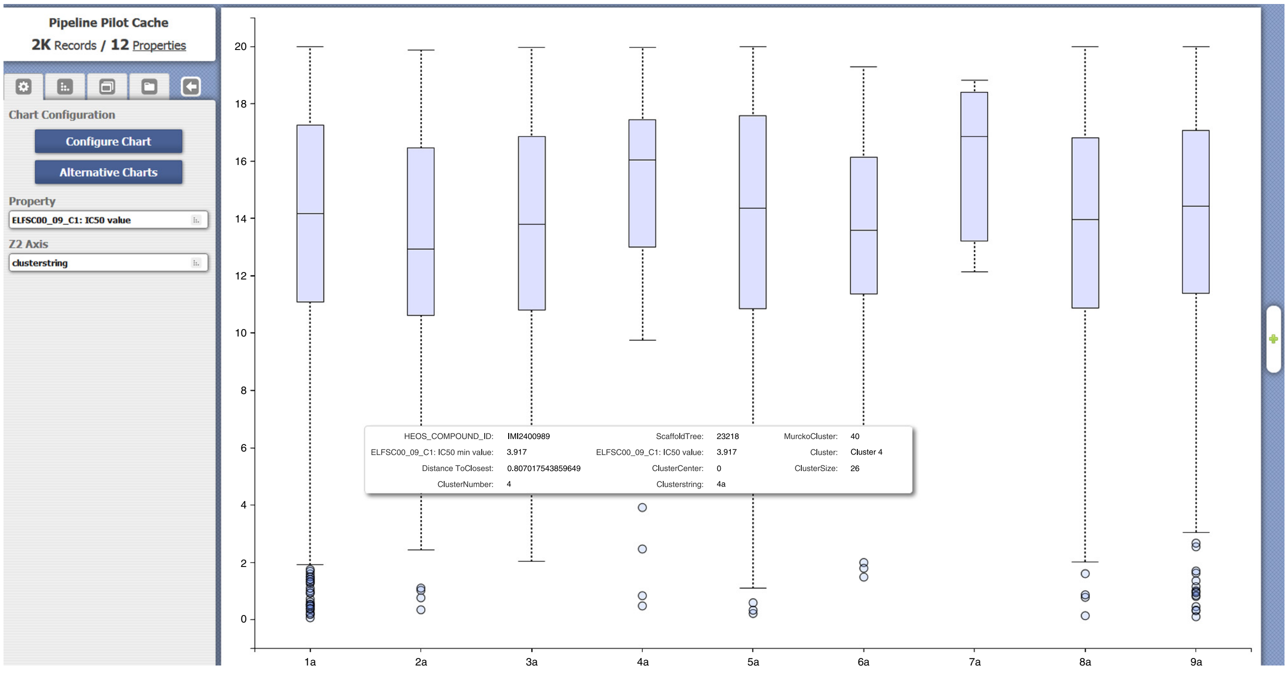
Key triage tools in the Honest Data Broker (HDB). Emphasis is placed on understanding the similarity relationships between compounds when structures are not visible in the early stages of the triage. Deriving knowledge from data deposited in the HDB to inform triage is also made possible and is an area under active development

Triage tool	Function
Property calculators (log *P*, MW, TPSA, etc.)	Generation of key drug-likeness properties
Scaffold clustering	Group molecules by core structure
On-the-fly clustering	FCFP6/Tanimoto clustering of arbitrary compound lists
Filtering	Filter by range, top N, outlier filtering
Activity model calculation	False-negative rescue and/or unusual hit identification
Pre-computed ChEMBL activity models	Off-target biological activity identification
Ligand efficiency, LLE	Size and log *P* normalised activity estimation
List logic	AND/NOT/OR/XOR joining of compound lists
Limited substructure filtering	(De)select compounds with structural feature
Binned similarity search	Identify compounds similar to query molecule
Pareto/Derringer desirability scoring	'Soft' selection of compounds without hard filter limits
Basic math and statistics	User-specified mathematical operations
Cross-programme information	Shows activity of compounds to against enzyme/receptor classes to facilitate selectivity estimation
CNS filter	Desirability model to identify probable CNS-penetrant compounds
SMARTS filtering	Substructures encoding structural alerts to identify potentially undesirable features

features provide the user with a pre-computed classification of the compounds based on the core of the molecule. Clustering of arbitrary user-selected lists of compounds (on-the-fly clustering) is available using FCFP6 fingerprints/Tanimoto similarity in the standard clustering component of Pipeline Pilot® [7,8]. These methods provide different views on the relationships between compounds and facilitate grouping in chemical series. The chemical space around each compound has been analysed and the number of near neighbours for each compound in the SRL is reported enabling the triage scientist to look for SAR and potentially prioritise compounds in clusters over singletons. The HDB can also return compounds similar to a set of query compounds in a single operation. This is useful if, for example, the screen returns a relatively small number of hits and the triage team want to select the near neighbours of each one for visualisation on the PHL and SAR analysis. Rapid deselection of undesirable compounds is also possible. Each company and the European Screening Centre have uploaded a standard set of SMARTS strings [9] identifying undesirable structural features, such as Michael acceptors in molecules. PT and PCT members can see how many queries are matched in their own set and in other sets to get a consensus view if required. In addition, lists of up to 20 substructures, each of up to ten heavy atoms, can be uploaded allowing selection or deselection of compounds on a project-specific basis. This allows, for example, a PT member to deselect chemotypes already well explored and prevent valuable PHL slots being used. Imported structures can also be used in similarity searches for selection of compounds and, in this case, there is no limit on the number of heavy atoms in the query molecules. Measures have been put in place to prevent abuse of such functionality. In principle, a user could import a large number of structures and filter the entire SRL to identify JECL compounds with a similarity of 1. This would reveal compounds identical to, or very similar to, known structures. To prevent this kind of 'fishing' the similarity results are binned in 0.1 Tanimoto similarity ranges. This retains the usefulness as a triage tool while obscuring identical, or very nearly identical, relationships between compounds.

Sophisticated analysis protocols can quickly be created with the available tools. Compounds can be grouped and analysed by cluster to generate an immediate overview of the potential of each cluster to be developed further. Laplacian-modified naive Bayesian classifiers (Bayesian models) [10] can be built using the HTS data. The model predictions can, for example, be used to identify possible false-negatives in the HTS screen and also identify unusual hits in the screen that are likely to be particularly interesting, atypical chemotypes or false-positives. It is also possible to import Bayesian models built outside the HDB and use these to rank compounds. External bioactivity data can also be accessed using a panel of Bayesian models already built using data from ChEMBL [11]. Over 1000 models are available and these can also be used to guide decision making against these targets. The Pareto and Derringer desirability tools [12] allow compounds to be selected without setting hard cutoffs on property values. This has proved useful in selecting central nervous system (CNS)-penetrant compounds where one property being close to an ideal value can compensate for another being outside the conventionally acceptable range.

A feature of performing a screen within the ELF environment is the ability to use knowledge from all the accumulated data within the ELF to inform decisions. Targets assayed in the ELF are given a high level classification (e.g. Ser/Thr kinase, protease) and the number of times a compound has shown activity to each class in all ELF programmes can be displayed. This allows the triage scientist to get an idea of the selectivity of the compound but without compromising confidentiality in these other programmes. The assay technology is also registered in the HDB at



Property
ELFSC00_09_C1: IC50 value

Z2 Axis
clusterstring

HEOS_COMPOUND_ID: IMI2400989
ELFSC00_09_C1: IC50 min value: 3.917
Distance ToClosest: 0.807017543859649
ClusterNumber: 4
ScaffoldTree: 23218
ELFSC00_09_C1: IC50 value: 3.917
ClusterCenter: 0
Clusterstring: 4a
MurckoCluster: 40
Cluster: Cluster 4
ClusterSize: 26

Drug Discovery Today

FIGURE 3

An example of a visualisation from a triage. Compounds are grouped by cluster (*x* axis) and the scatter of EC_{50} values (micromolar values on *y* axis) within each cluster is demonstrated in a boxplot. It is immediately apparent that cluster 7a has few potent molecules and this can be used to inform decision making.

the start of a programme and the number of times a compound is active in a given assay design across all programmes can also be shown. This identifies potential assay-interfering compounds for further investigation.

Combining all these tools can generate complex protocols that can be saved as a template for future use. Analysis and visualisation of the protocol outputs is crucial and a sortable spreadsheet for analysing the output of a triage is available and more sophisticated visualisation and spreadsheeting is available through the Pipette application on ScienceCloud. Pipette has been closely integrated with the HDB to enable seamless transfer of information between the HDB and Pipette, always in compliance with business rules. An example of this visualisation is shown in Fig. 3 where compounds are grouped by cluster and the range of bioactivity is displayed in a box-and-whisker plot.

Programme organisation

Programmes are created within the HDB by consortium members with administrator rights. Mandatory input information such as Uniprot ID followed by automatic comparison with existing programmes prevents the same target being screened twice. Key information such as the programme owner and target class are also entered here. Target class information is required for functionality such as cross-programme information. Once programmes are running, two specific challenges are presented by the ELF. Firstly, the number of programmes running concurrently within the public screening centre is large and PT and PCT members need to deal with a large amount of confidential information. Permissions prevent data and compound selections being visible across programmes, whereas the entirely self-contained triage process within the HDB eliminates the risk of misdirected emails and spreadsheets being sent to the wrong programme owner. The HDB also enables a high level of data organisation within each programme. Biological assays are accessible through a hierarchical tree and arbitrary compound worklists can be saved, time-stamped and annotated for future use. Secondly, PT and PCT members on a project are often located at different sites and the cloud-based nature of the HDB enables cross-site project decision-making by sharing these worklists.

In addition to supporting scientists working directly on the programme, all interactions between the different roles involved in a target programme are supported by the HDB application. For example, the CCR process requires actions from three different roles: PCT members, who create those requests, compound owners, who need to review and approve or reject the requests, and administrators, who are in charge of the consistency of the process. The HDB application provides all necessary features to support the workflow execution. Events such as creation of CCRs by PCT members are automatically generated on registering the RHL and the sending of reminders to compound owners to process requests are automated, preventing unnecessary manual actions. Dashboards and reports are available to compound owners and administrators enabling them to survey which requests are outstanding and require attention as well as allowing audits and review over time.

Concluding remarks

The HDB addresses the growing trend in pharmaceutical discovery for precompetitive partnerships [13] and allows a highly heterogeneous group of organisations comprising big pharma, SMEs and universities to work as a single consortium within a secure environment. A unique set of roles and fine-grained permissions meets the requirements for IP security; and a suite of cheminformatic tools specifically adapted for the unique workflow enables scientists to carry out a high-quality triage within the rules of the ELF. The cloud deployment of the HDB supports rapid introduction of new features in response to user need and defect repair. It also ensures uniformity of tools and application of the consortium's rules rigorously across all partner organisations; an important consideration in such a widely distributed consortium. This approach provides a model for similar large-scale public–private initiatives in the future and is readily adapted for different circumstances.

Conflicts of interest

Guillaume Paillard and Andrei Caracoti are full-time employees of Dassault Systemes BIOVIA.

Acknowledgements

We acknowledge the input and support of the seven EFPIA companies: AstraZeneca, Bayer, Janssen Pharmaceuticals, Lundbeck, Sanofi, Merck-Serono and UCB. This project has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115489, resources of which include financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution.

References

1 Mullard, A. (2013) European Lead Factory opens for business. *Nat. Rev. Drug Discov.* 12, 173–175

2 Besnard, J. *et al.* (2015) The Joint European Compound Library: boosting precompetitive research. *Drug Discov. Today* 20, 181–186

3 Nelson, A. and Roche, D. (2015) Innovative approaches to the design and synthesis of small molecule libraries. *Bioorg. Med. Chem.* 23, 2613

4 Hann, M.M. and Keseru, G.M. (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discov.* 11, 355–365

5 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893

6 Schuffenhauer, A. *et al.* (2007) The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47, 47–58

7 Hassan, M.B. *et al.* (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol. Divers.* 10, 283–299

8 Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754

9 SMARTS – A Language for Describing Molecular Patterns. Available at: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

10 Xia, X. *et al.* (2004) Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* 47, 4463–4470

11 Gaulton, A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107

12 Segall, M.D. (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* 18, 1292–1310

13 Andrews, D.M. *et al.* (2015) Compound Passport Service: supporting corporate collection owners in open innovation. *Drug Discov. Today* 20, 1250–1255