doi:10.2533/chimia.2017.667 Chimia 71 (2017) 667–677

The Screening Compound Collection: A Key Asset for Drug Discovery

Christoph Boss*, Julien Hazemann, Thierry Kimmerlin, Modest von Korff, Urs Lüthi, Oliver Peter, Thomas Sander, and Romain Siegrist

Abstract: In this case study on an essential instrument of modern drug discovery, we summarize our successful efforts in the last four years toward enhancing the Actelion screening compound collection. A key organizational step was the establishment of the *Compound Library Committee (CLC)* in September 2013. This cross-functional team consisting of computational scientists, medicinal chemists and a biologist was endowed with a significant annual budget for regular new compound purchases. Based on an initial library analysis performed in 2013, the *CLC* developed a *New Library Strategy*. The established continuous library turn-over mode, and the screening library size of 300'000 compounds were maintained, while the structural library quality was increased. This was achieved by shifting the selection criteria from 'druglike' to 'leadlike' structures, enriching for non-flat structures, aiming for compound novelty, and increasing the ratio of higher cost 'Premium Compounds'. Novel chemical space was gained by adding natural compounds, macrocycles, designed and focused libraries to the collection, and through mutual exchanges of proprietary compounds with agrochemical companies. A comparative analysis in 2016 provided evidence for the positive impact of these measures. Screening the improved library has provided several highly promising hits, including a macrocyclic compound, that are currently followed up in different *Hit-to-Lead* and *Lead Optimization* programs. It is important to state that the goal of the *CLC* was not to achieve higher HTS hit rates, but to increase the chances of identified hits to serve as the basis of successful early drug discovery programs. The experience gathered so far legitimates the *New Library Strategy.*

Keywords: Compound Library Committee (CLC) · Screening compound collection · Screening compound selection process · Virtual TNT-library

1. Introduction

Today, pharmaceutical companies rely heavily on High-Throughput Screening (HTS) campaigns to identify novel compounds (so-called hits) to initiate their medicinal chemistry programs. Two key elements determine the success of an HTS: A biologically significant and technically robust biological assay, and the structural and physical quality of the screening compound collection. In this article, we focus on the structural aspects of the screening compound collection, and the parameters we considered important for the design and assembly of our collection.

*Correspondence: Dr. C. Boss
Drug Discovery & Preclinical Development
Idorsia Pharmaceuticals Ltd
Hegenheimermattweg 91, CH-4123 Allschwil
E-mail: christoph.boss@idorsia.com

1.1 *Structural Quality*

First, the structural ***quality*** of the screening compounds needs to be carefully assessed. Since the seminal paper of Lipinski,[1] the importance to control the physicochemical properties of medicinal chemistry as well as screening compounds is broadly understood and accepted. During the hit to lead (H2L) optimization process, lipophilicity, molecular weight, and complexity of molecules tend to increase.[2] To compensate for this, the rules to select druglike compounds for the screening library, were changed to favor more leadlike compounds.[3] In addition, rather than applying hard cut-offs for parameters like MW, cLogP or PSA, modern approaches employ scoring functions or multi-parametric optimization (MPO) procedures. An interesting example of a scoring function was developed by Pfizer in their CNS MPO workflow.[4]

A good library needs to be continuously maintained. Problematic compounds, such as frequent hitters or reactive chemicals, have to be identified and should either be physically eliminated, or flagged. To exclude such structures from the beginning, alerts such as the REOS and PAINS lists of undesired structural elements have been developed, which are now broadly applied.[5] Medicinal chemists depend on high quality lead structures, based on high quality hits, to start a lead optimization program. They will, and should, disregard problematic chemical starting points exhibiting undesired functionalities or properties.

The ***three-dimensional (3D) aspect*** of a compound heavily impacts its physicochemical properties, pharmaceutical effects, and potential side-effects. As shown by Lovering and colleagues, spatial complexity (as measured by the number of Fsp3 carbons) as well as the presence of chiral centers, correlate with success as compounds transition from discovery through clinical testing to drugs.[6]

1.2 *Structural Diversity*

Second, the structural ***diversity*** of the collection is essential. The development of combinatorial chemistry in the 1990s permitted the rapid synthesis of large numbers of structurally very similar compounds. As a consequence, all major pharmaceutical companies built very large screening libraries with relatively low structural complexity. This enabled ultra-high-throughput screens (uHTS) covering millions of compounds using relatively simple (preferably homogeneous and miniaturized) biological activity assays. However, this incurred experimental constraints and high financial costs. Therefore, even large pharmaceutical companies now tend to screen

smaller library subsets initially, and then iteratively expand the screen based on the initial results. For smaller companies such as Actelion, smaller screening compound collections assembled from high quality compounds covering a large area of the chemical space were always the realistic option.

A generalized screening compound collection has to deliver hits on different targets, such as GPCRs, ion channels, kinases, proteases; it may serve different therapeutic areas with specific requirements, such as blood-brain barrier penetration or bacterial accumulation; and it should be compatible with different assay technologies, such as enzymatic or cell based assays, high content screenings, or phenotypic assays. Therefore, it is tactically useful to divide the compound collection into specific subsets. For example, tool compounds annotated with known biological activities help to classify other hits according to their mode of action, or allow to scrutinize biological pathways. A library subset spanning the chemical space of the library with approximately 10% of the compounds enables small proof-of-concept screens. A set of small molecules enables alternative hit discovery approaches; such as fragment-based screens. In summary, the screening compound collection needs to cover a large and diverse chemical space with a rather limited number of compounds.

1.3 *Novelty*

Third, the ***novelty*** of the compounds is important to commercial drug discovery, as it is the basis for a competitive advantage at a later stage, and to generate intellectual property. The novelty assessment can be easily performed using cheminformatics tools. Several sources can be explored to identify inventive structures. Natural products or natural product-derived compounds can offer valuable and original starting points. Compound exchanges with other pharmaceutical or agrochemical companies is an efficient way to access additional diversity and uncovered chemical space. The screening compound collections of pharmaceutical companies are often dissimilar, as demonstrated, for example, by Bayer and AstraZeneca.[7] This is why Sanofi and AstraZeneca recently realized an exchange of 200'000 compounds.[8] Also, contract research organizations (CROs) offer screening compounds, designed or focused libraries, or scaffolds suitable for extensive decoration. Although a large number of compounds is commercially available, their originality, price, or exclusivity varies greatly.

In the following, we detail our strategy and the tactical measures implemented to adapt Actelion's corporate library and the screening compound collection to the newest scientific findings in hit and lead discovery, and propose new concepts toward managing and maintaining screening compound collections.

2. The Actelion Screening Compound Collection: Strategy and Tactics

The Actelion screening compound collection comprises a dynamic library of approximately 300'000 compounds with an average lifetime of five years per 'catalogue compound', and ten years per 'premium compound' (definitions see below). Every year, the oldest 60'000 compounds are removed, and the lost chemical space is analyzed. The result of this analysis is, together with other factors, the basis for the selection of 60'000 new compounds to bring the collection to 300'000 compounds again. Hence, 120'000 compounds are physically moved every year, 60'000 in and 60'000 out (Fig. 1).

The collection includes selected internal project compounds as well as external compounds. The compounds are divided into two distinct categories. The first category is called 'premium compound' and has a lifetime of ten years, encompassing our own project compounds, and rare, novel, expensive or custom-made commercial compounds. The second category is called 'catalogue compound' and has a lifetime of five years, encompassing the inexpensive or easily commercially available compounds.

In 2013, the Actelion screening compound collection of 300'000 compounds was evaluated internally as well as by an external company. The following convergent conclusions were drawn from these two independent analyses:

- The library was structurally highly diverse.
- The physicochemical property distribution was well balanced, but rather in the classical druglike range (Lipinski's rules of 5) than the desired leadlike range.
- The proportion of flat versus non-flat compounds was high (see 'Flatland'[6]).
- The proportion of commercial compounds versus proprietary compounds was very high.

As a reaction to this analysis, in early 2013 it was decided to create the 'Compound Library Committee' (CLC), a group of scientists covering the diverse disciplines involved in early drug discovery, to be responsible for all strategic and tactical aspects related to the Actelion screening compound collection. Today, the CLC consists of a medicinal chemist, two computational chemists, one representative from HTS biology and a hit-to-lead chemist. The Chemistry Group Leader responsible for the budget allocation attends the committee meetings as a permanent guest. The CLC tightly collaborates with compound management in order to implement technology changes and the significant compound logistics efforts triggered by the CLC's decisions. The major task of the CLC was and still is to develop and implement a screening compound collection strategy, ensuring that the screening compound collection is fully in line with the strategic and tactical needs of Actelion's early drug discovery projects. The first measure that the CLC took was to correct the shortcomings identified in the analysis of the Actelion screening compound collection.

Within a year, the CLC formulated ideas to be implemented into the *new library strategy* to enhance the probability to discover hits with the potential of high-quality leads. To gain acceptance for these changes among medicinal chemists, the commitment to full transparency was key. All CLC proceedings are made available to all Actelion drug discovery scientists. The CLC also presents its vision, plans, activities and efforts twice a year to all interested stakeholders within the research departments. This openness not only led to a broad understanding and acceptance of the CLC, but also to an increased interest in HTS campaigns and their results.

The main elements of the *new library strategy* were (i) to keep the *annually updated compound turn-over (rolling mode)* of the Actelion screening compound collection; (ii) to explore new, not yet or only marginally covered chemical space such as macrocycles, natural-product derived compounds, or focused- and designed libraries; (iii) to implement compound exchange programs with agrochemical companies and, potentially, with other pharmaceutical companies; (iv) to move from a druglike

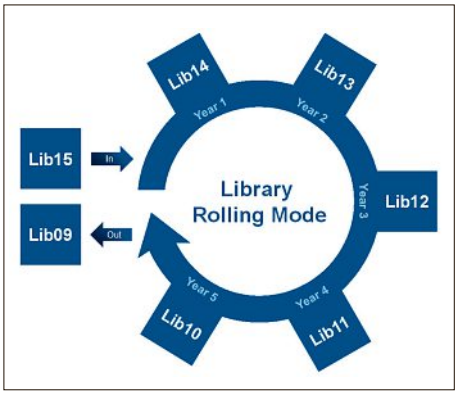

Fig. 1. The 'Rolling Mode' of the Actelion Screening Compound Collection.

to a leadlike structural bias, and from flat to non-flat structures; (v) to apply multiparameter optimization (MPO) scores for compound selection; (vi) to enhance the REOS and PAINS filters for compound de-selection; (vii) to assess new technologies, for example DNA-Encoded Library (DEL) screening, or increase fragment-based lead discovery; (viii) to strengthen the in-house resources dedicated to experimental hit follow-up of both structure-based and ligand-based virtual screenings; (ix) to invest a significant part of the budget assigned to the screening collection to acquire exclusive 'premium compounds'.

At a practical level, it was also decided to keep the library formatted on 384 well plates for screening. To maximize the benefit from the more expensive 'premium compounds', it was decided to double their storage lifespan to ten years, as compared to the five years for 'catalogue compounds'. At the same time, the amount of substance ordered was increased, and the concentration of the compound stock solutions was raised from 4 mM to 10 mM in 100% DMSO. This allows to satisfy the increasing demand of material for hit follow-up investigations, and to increase the screening concentration if needed. Previous experience had shown that insufficient solubility at 10 mM is no longer a problem with compounds of high structural quality.

As a practical consequence, larger automated compound stores were acquired to accommodate the single tube compound aliquots.

The strategic measures mentioned above align the size of the Actelion screening compound collection (300'000 compounds) with the available plate storage capacity, compound management processes, and to the 384-well format used in the High-Throughput Screening (HTS) workflow. An analysis performed by others suggests that a library of this size can representatively cover the leadlike space.[9]

To efficiently execute the compound selection, library management, and compound logistics processes, proprietary software has been developed.[10]

The continuous library turn-over is laborious and costly. However, in our experience, this approach holds important advantages. The library always reflects the newest developments in medicinal chemistry. Limiting the compound age increases the likelihood that a compound can still be re-ordered, and compounds remain in a good physical condition. LC-MS analysis of the latest screening library revealed that 88% of the compounds are of good quality at the beginning of their useful life (Fig. 2).

One of the CLC's most important responsibilities is the annual replacement of compound sets according to the 'rolling mode'. The means by which this is done are explained in the following sections. It is important to mention that only screening plates are discarded. Compounds which are still available as powders and/or individual solutions are kept in the stores. Hence the total number available from Compound Management at Actelion, including duplicates and salt forms, is 850'000 compounds representing 600'000 unique structures. Although approximately 300'000 compounds are not part of the screening set at a given time, they can still be accessed within days for targeted screening and rapid hit expansion. In contrast, hit expansion through external providers is often a time-consuming process.

2.1 *Compound Selection and Library Assembly*

The compound selection and library design is performed with the Knime Analytics Platform and additional internally developed software. This process was adapted and improved based on the outcome of the analysis of the Actelion screening compound collection in early 2013 as described above.

The optimized Actelion screening compound selection workflow consists of four main steps:

1. MultiParameter Optimization (MPO) scores calculation
2. Substructure filtering and removal of unwanted chemical functionalities
3. Library comparison and gap filling
4. Clustering and dissimilarity selection

2.1.1 *Step 1*

The very first step of the selection process, performed with Knime, is the calculation of leadlike MPO scores on a pool of candidate compounds. Our leadlike MPO scores calculations are adapted from a similar, published approach of CNS MPO scores calculations.[4b] As depicted in Fig. 3, a set of eight physico-chemical properties (MW, clogP, HBA, HBD, Molecular Flexibility, aromatic ring count, Fsp3 and stereo center count) were used to build our own leadlike and non-flat compound MPO scoring algorithm. The following considerations underpin the choice of these parameters.

As previously mentioned, the size of compounds tends to increase during optimization cycles, hence compounds with a MW < 400 were considered as preferred starting points.

Highly lipophilic compounds have an increased risk to interact promiscuously with proteins, leading to target selectivity

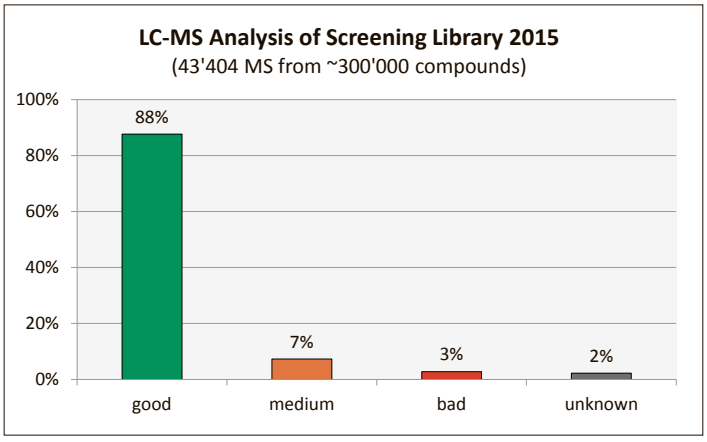

Fig. 2. Quality Control (QC) results of the Actelion screening compound collection. According to their integrity and purity compounds were flagged as *good* (>85% pure), *medium* (50–85% pure) or *bad* (<50% pure, or wrong compound).

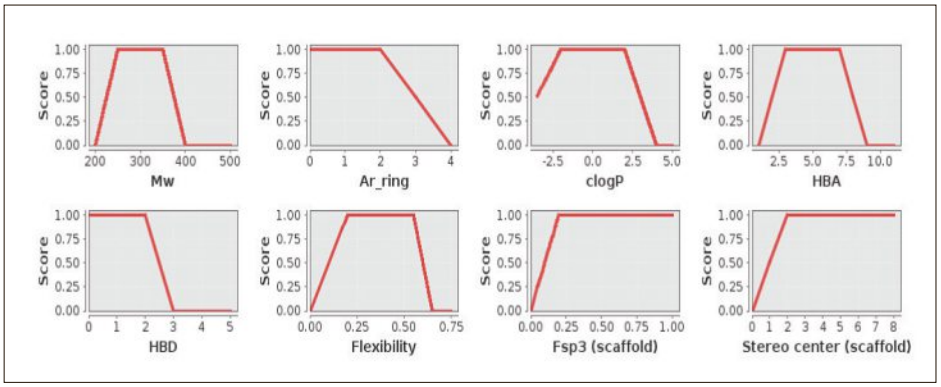

Fig. 3. Each plot represents one of the eight physicochemical property desirability functions used to generate our leadlike MPO. Multiparameter optimization methods are commonly used to assess and balance the effects of several variables, weighted based on their importance to the overall objective.

issues and off-target toxicity. They also often display suboptimal metabolic stability. On the other hand, highly hydrophilic molecules may be less membrane permeable. Therefore, the optimal clogP range was set between –2 and 2.

The number of aromatic rings in a molecule has been shown to influence the chance of drug development success and should be maintained below four.[11]

To enable the 'Escape from Flatland'[6], three-dimensional complexity (conferred through Fsp3 carbons) and the presence of chiral centers at scaffold level are taken into account for the scoring.

Since it is connected to membrane permeability,[1] the number of Hydrogen Bond Acceptors (HBA) and the number of Hydrogen Bond Donors (HBD) were also considered for the MPO score calculation.

Finally, Molecular Flexibility (detailed in section 4) was found to be a predictor for good oral bioavailability. The optimal range was set between 0.20 and 0.55.[12]

The use of MPO scoring for the selection of compounds allows for greater flexibility in drug design beyond the use of single parameters or hard cutoffs. Ultimately, MPO scoring should allow to identify compounds with higher probability of success.

A trapezoidal membership function is used for the calculation of each individual desirability score of each physico-chemical property as defined in Fig. 4.

Then an overall desirability score D is calculated as defined below, where S_p is the individual score per property, and W_p is the weight associated with each property:

$$D = \sum S_p W_p / \sum W_p$$

The highest possible D score is 1. Compounds displaying an overall score D below 0.6 are discarded at this stage of the selection process.

2.1.2 *Step 2*

The second step consists of eliminating unwanted chemical features, for example reactive functionalities which often result in promiscuity of compounds toward many targets as well as anti-targets, and therefore enhance the risk of unwanted or toxic effects. This step is performed based on the substructure filtering functionality within the Knime workflow. Originally, PAINS[5e] and REOS[13] SMARTS strings were used. With time, our collection of SMARTS was continuously enriched through the input and feedback of medicinal chemists. Currently, our list of SMARTS strings contains 670 entries. Each SMARTS string defines how often a given substructure is allowed to appear in a given compound. Fig. 5 represents a snapshot of some SMARTS strings with their associated rule used to filter compounds. From a technical point of view, each candidate compound is parsed against 670 SMARTS strings before being allowed to enter the next selection round, if compliant with all rules.

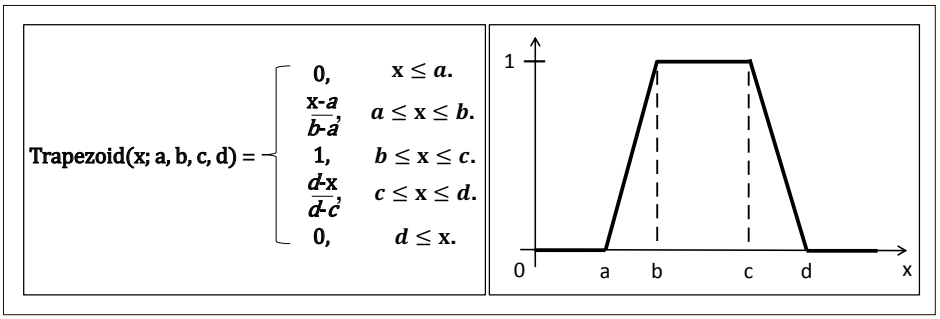

Fig. 4. A trapezoidal membership function (MF) is specified by four inflexion points.

2.1.3 *Step 3*

The third step involves an internally developed tool called VSCmd which allows to perform a library comparison between the pool of candidate compounds and the Actelion screening compound collection. The process begins by virtually discarding the 60'000 oldest compounds from the screening compound collection as a consequence of the 'rolling mode'. Then, SkeletonSpheres (see section 4) descriptors are calculated for both sets of compounds, the pool of candidate compounds and the remaining Actelion screening compound collection. The molecular descriptors are used to compute similarity coefficients. Each candidate compound is compared against each Actelion screening compound, in order to find the nearest neighbor. Then, to ensure broad chemical space coverage, all candidate compounds having a similarity coefficient above 0.8 are discarded. Remaining candidate compounds are now all dissimilar to the compounds of the Actelion screening compound collection and the gap lost by removing 60'000 compounds is covered again with new, similar or identical, compounds.

2.1.4 *Step 4*

The last step of the selection workflow is the clustering and dissimilarity selection using an internally developed tool based on OptiSim technology (see section 4). This allows the selection of a diverse set of candidate compounds which efficiently cover the chemical space of interest. Based on SkeletonSpheres descriptors, all compounds with a dissimilarity coefficient below 0.2 are discarded. This approach allows the selection of a diverse set of compounds eligible to enter the Actelion screening compound collection.

2.2 *Screening Compound Collection: Tactical Measures to Optimize the Content*

As mentioned, our original analysis revealed that the screening compound collection initially contained too many flat molecules, and they were more drug- than leadlike. Increasing the number of chiral centers, of sp3-carbons, and of saturated carbo- and heterocycles has been proposed to obtain more globular molecules, thereby enhancing the chance to discover compounds for clinical success.[14] In a screening compound collection such as ours, containing only 300'000 compounds, the pharmaceutically relevant chemical space can only be covered by optimal use of molecular diversity descriptors and by avoiding redundancy by all means.

SMARTS	Comment	Rule
[CX3H1](=O)[#6]	aldehydes	Max: 0
FCS(=O)(=O)[F,Cl,Br,O]	triflates	Max: 0
[SX3](=O)	sulfoxides	Max: 1
[OH]cccc[OH]	para phenols	Max: 1
C(=O)[O-,OH]	carboxylic acids	Max: 1
c12ccccc1cccc2	Naphthalenes	Max: 1
N-[OH]	Hydroxylamine	Max: 1
[#6][#6](=O)-&!@O-&!@[#6]	esters	Max: 2
c[OH]	aromatic hydroxyls	Max: 2
cBr	aromatic bromide	Max: 2

Fig. 5. Examples of SMARTS strings.

Therefore, we enriched the Actelion screening compound collection with compounds from several diverse sources described in the following paragraphs.

2.2.1 *Scaffold Selection and Non-exclusive Screening Compound Design*

In order to enlarge our collection of novel, innovative and non-flat compounds, we requested two CROs to propose novel scaffolds. The following requirements were stated: The central core should not appear in public databases, and it must have two exit vectors, a weak one and a strong one. The weak exit vector was used to prepare five sub-scaffolds from each initial scaffold. The strong exit vector served to achieve structural diversity. Thus, for each sub-scaffold, 40 final compounds were synthesized, resulting in 200 compounds around one initial scaffold. To maximize diversity, 200 different building blocks were selected for the final decoration of the strong exit vector. This clearly challenged the synthetic capacity of the CROs. Each CRO worked on ten scaffolds, resulting in a total of 2'000 compounds per CRO.

Although large numbers of innovative compounds were obtained in this way, the CROs did not succeed to prepare all selected central cores. Indeed, several scaffolds selected based on 'paper chemistry' turned out to be synthetically not accessible in a flask. Also, to keep the costs of this approach reasonable, the compounds had to be obtained on a non-exclusive basis.

In a second approach, off-the-shelf designed libraries were acquired from several suppliers on a non-exclusive basis. This approach was considered attractive since such libraries often represent the leftovers of non-exclusive designed libraries ordered by potential competitors. As for exclusive and non-exclusive library compounds, the off-the-shelf designed library compounds were selected for their novel and sp3 rich scaffolds. Additionally, it was possible to reach a higher chemical diversity with these compounds, as this cherry picking approach allowed to select only a few analogs around one central core.

2.2.2 *Scaffold Selection and Exclusive Screening Compound Design*

We decided to also design and synthesize proprietary and exclusive libraries. Therefore, suitable scaffolds had to be identified, either in our in-house or in external collections. Again, the selection started with the identification of scaffolds exhibiting a high number of sp3-carbons and a low molecular weight. For example, spirocyclic scaffolds represent valuable starting points.[15]

The scaffold candidates were then compared to our internal screening compound collection as well as to the eMolecules database, in order to assess complementarity, originality and novelty. To ensure an optimal diversity, we only kept scaffolds with at least two exit vectors. The selected scaffolds were then virtually decorated. The resulting virtual compounds were filtered as previously described, and a final diversity selection was performed. To maximize the diversity, we limited the number of derivatives per scaffold to 200. The rules applied were the same as for the non-exclusive compounds, with a weak and a strong exit vector subsequently enumerated to obtain maximal diversity.

In a further effort toward obtaining exclusive compounds, 15 natural-product like, novel, sp3-enriched scaffolds were sourced and a CRO was approached for the production of the final exclusive compounds by applying maximum building block diversity; this time on both exit vectors.

2.2.3 *Semi-exclusive Screening Compounds*

We directed our efforts toward target-focused compound libraries. These are collections of compounds which are designed to interact with an individual protein target or, frequently, target class (such as kinases, voltage-gated ion channels, serine- or cysteine proteases, or GPCRs). The design of such libraries generally utilizes structural information about the target or family of interest. In the absence of such structural information, a chemogenomic model that incorporates sequence and mutagenesis data to predict the properties of the binding site is often used. Another approach uses the information about known ligands of the target, to generate focused libraries through scaffold hopping.[16]

We acquired ion channel and Protein–Protein Interaction (PPI) focused library compounds on a semi-exclusive basis (compounds sold to a limited number of competitors).

2.2.4 *Compound Exchanges to Acquire Screening Compounds*

Agrochemical compounds are potentially interesting as starting points for pharmaceutical drug discovery, as they are distinct from medicinal chemistry compounds yet, like medicinal chemistry compounds, they are designed to be active on a biological target.[17] A compound exchange can be an efficient way to get access to premium compounds, but it depends on an active commitment from various corporate functions. In our case, this included the compound library committee (CLC), chemistry management, the legal department, computational chemistry, compound management, and logistics functions.

To be eligible for exchange with a partner, proprietary compounds need to be available in sufficient quantities and must belong to the exclusive part of the compound collection. At Actelion, a compound exchange is initiated by the CLC and approved by the Drug Discovery Chemistry management. To ensure a successful outcome, a project manager is assigned. A solid and clear contract is negotiated with the partner. Once the contract is signed and the lists of eligible compounds are exchanged, our computational chemistry team reviews, evaluates and selects the compounds of interest.

The most labor intensive step is the manual weighing and transfer of a defined amount of each compound requested by the exchange partner, from an Actelion vial to a destination vial provided by the partner; and receiving the compounds from the partner, dissolving them followed by LC-MS quality control, and integrating them into the screening compound collection.

Shipping and handling needs to be according to the partners' varying requirements, and compliant with national and international safety and legal regulations.

Finally, once the partner discovers hits among the provided compounds, resupply requests need to be fulfilled according to the terms agreed.

In addition to the logistical efforts that were initially underestimated, some additional lessons were learned in the course of several compound exchange projects:

- As the agrochemical compounds were selected from the partners' historical collections, about 33% of them did not pass Actelion's stringent quality requirements (see Fig. 6).
- The exchange program required clear and transparent internal communication to counter medicinal chemists' worries about "giving away their compounds" to another company.
- To facilitate the hit validation of exchange compounds, it is essential to consider compound resupply, synthesis protocol supply, and hit expansion in general, when setting up the terms of the initial exchange contract.
- As compounds obtained from, or provided to, a partner through an exchange program are now subject to contractual obligations, it is important to electronically keep track of those (compound flagging).

Despite these challenges, the several compound exchanges organized proved very successful to Actelion, as they lead to hits with chemical structures usually not observed in the pharmaceutical drug discovery chemical space.

2.2.5 *Macrocycles as Screening Compounds*

Macrocycles belong to a unique chemical class that helps to fill the gap between conventional small molecules and large

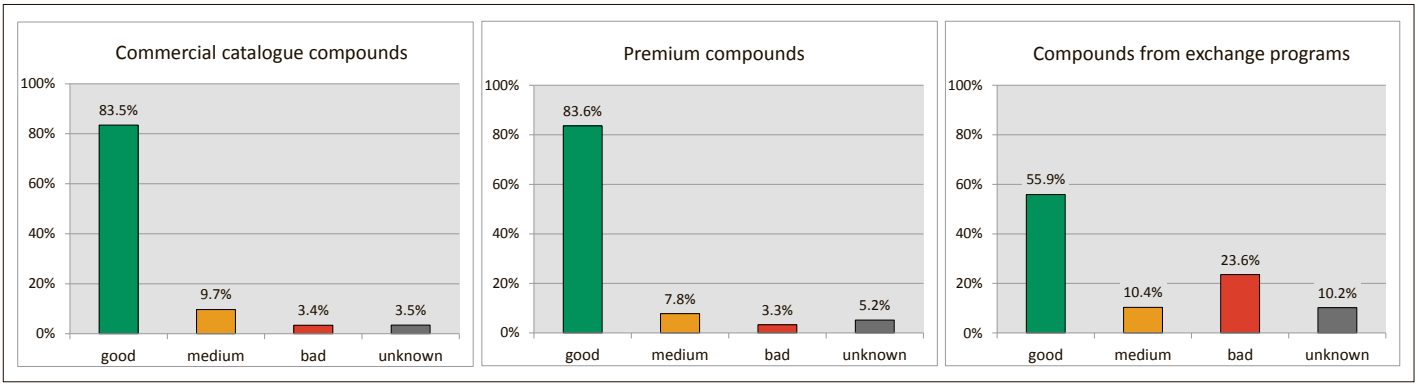

Fig. 6. Comparative QC data showing the lower quality of compounds coming from historical collections in compound exchange programs.

biomolecules.[18] They represent spatially pre-organized ring structures with special conformational features and behavior, with molecular weights ranging from 500 to 2000 Daltons. Macrocycles might offer a different tactic to tackling previously non-druggable targets, and represent a different intellectual property space. The tremendous progress in macrocyclization methodologies (*i.e.* click-chemistry, ring-closing metathesis, and palladium-catalyzed chemistry) in the last years, and the increasing number of commercially offered macrocycles render them valid starting points for chemistry programs. Our screening compound collection did not cover any macrocyclic chemical space in 2012. Therefore, the CLC decided to expand the collection in this direction. Over the past two years, commercially available macrocycles were selected and acquired based on scaffold and decoration diversity. Importantly, as most macrocycles do not fulfill the RO5, other selection criteria need to be developed for this type of compounds.[19]

2.2.6 *Niche Supplier Screening Compounds*

Niche supplier compounds are compounds offered by small compound suppliers. These compounds are less visible and less easily accessible than compounds available from large suppliers. Nevertheless, such compounds remain interesting as they are likely to be different from compounds found in large collections. In order to increase the scaffold diversity of the Actelion screening compound collection, a database collected by a commercial aggregator was used to select and acquire more than 1'000 niche supplier compounds.

2.2.7 *Commercial Virtual Library (CVL)*

The CVL is a virtual collection of about 7 million novel lead-like compounds constructed from more than 3'500 innovative cores (bridged ring-systems, spirocycles, medium-sized or annelated ring systems, and the like) and a huge collection of substituents, proposed by one of our suppliers. Such compounds can be synthesized at affordable prices and delivered within two to three weeks. More than 2'000 compounds were selected from the CVL based on scaffold novelty and diversity, and 2'000 were actually synthesized for the Actelion screening compound collection.

2.2.8 *University Screening Compounds*

To further explore the chemical space and in order to get access to innovative compounds, we tried to acquire compounds from academic laboratories. Direct contacts with individual academic investigators turned out to be time consuming and unproductive. Therefore, an aggregator compiling available university compounds was again used for the selection and acquisition of about 300 compounds from university labs. Due to the limited number, it is so far difficult to evaluate the success of such compounds in HTS campaigns. We also observed that the physical quality of these compounds was comparably low. This approach is presently not pursued any further.

2.2.9 *Natural Product and Derivative Compounds*

Natural products and their derivatives exhibit different physicochemical properties compared to standard synthetic compounds, yet they account for 39% of drugs approved by the authorities. Natural products can offer very valuable starting points to explore new chemical space. Moreover, by chemical modification, novel compounds with different biological activity from the starting natural product can be discovered.[20] For these reasons, Actelion has acquired 6'456 diverse compounds selected by cherry picking from two providers' catalogs, to prepare its unique Natural Product and Derivative Compounds Library.

2.2.10 *Approved Drugs Collection and Bioactive Compounds Library*

The Approved Drugs Collection assembles 1'586 small molecule active pharmaceutical ingredients (APIs) that have been approved by the FDA, EMA and other agencies. The compounds in this collection were selected to represent a broad chemical and pharmacological diversity. However, highly peculiar structural classes were excluded (*i.e.* contrast agents for diagnostic use).

By their nature, a useful bioavailability of these compounds is given, and the safety profile in humans is known. The annotations indicate the main biological targets and mechanisms of action. Drug repositioning or repurposing can be an important part of a drug discovery program and has led to several blockbuster drugs (*i.e.* Viagra and Rogaine).[21] Phenotypic screens, new biomarkers and non-invasive imaging techniques have created new opportunities for pursuing novel indications for approved compounds.

Discovering one or several known drugs being active in a phenotypic assay can help in assay validation, target identification, or identification of (novel) mechanisms of action. Of course known drug compounds can be immediately used in animal models, possibly accelerating drug discovery programs.

The 'Bioactive Compounds Library' is a unique collection of 2'100 biologically active chemical compounds for high throughput screening (HTS) and high content screening (HCS). The molecular mechanism of action of these compounds is known, and often also pharmacokinetic and safety data from preclinical research and clinical trials. The library includes inhibitors, APIs, natural products, and chemotherapeutic agents. The compounds are structurally diverse and cell permeable. The collection is associated with a rich documentation including IC_{50} data on the primary target. The library in combination with the approved drugs collection is used as a tool to validate new drug discovery assays and characterize orphan receptors.

2.3 *Comparative Analysis of Libraries L12 and L15 to Show the Impact of the CLC Guided Actions*

Based on the outcome of the analysis of the Actelion screening compound col-

lection described above, the actions previously summarized were implemented over the last three years to optimize the screening compound selection process. In order to visualize the effect of our efforts, we compare the analysis of the part of the screening compound collection purchased in 2012 (L12), with the part of the screening compound collection acquired in 2015 (L15).

L12 consists of 58'335 compounds and L15 consists of 65'580 compounds. The following parameters were compared:

- Molecular Weight distribution
- Fraction of sp3 carbons distribution (Fsp3)
- clogP distribution
- Premium versus Catalogue compound count
- Ring system count
- Self-Nearest Neighbor Analysis

Fig. 7 displays the Molecular Weight (MW) distribution of L12 and L15. A shift toward lower MWs is obvious. This is a consequence of the CLC strategy to increase the fraction of smaller, leadlike compounds in the screening compound collection. Leadlike hits help to decrease the high rate of compound attrition in drug development.[22] The tendency of screening hits to gain in size during lead optimization – jokingly referred to as 'molecular obesity' – consists a development risk that may contribute substantially to the limited productivity of drug discovery programs.[23]

Fig. 8 displays the distribution of the fraction of sp3 (Fsp3) carbons per molecule. This parameter shifts up from L12 to L15, reflecting the aim to increase the fraction of non-flat compounds in the screening collection. More complex molecules, as measured by carbon saturation, have the capacity to access greater chemical space. This increases the potential to identify compounds that better match the surface of target proteins.

Fig. 9 displays the clogP value of a compound, *i.e.* the logarithm of its calculated partition coefficient between n-octanol and water, a well-established measure of the compound's hydrophilicity/lipophilicity. Overall, compounds in L15 are slightly more hydrophilic than those in L12.

Fig. 10 displays the proportion of premium compounds versus catalogue compounds for L12 and L15. As intended, the fraction of premium compounds has increased over the last three years.

Fig. 11 displays the distribution of (plain) 'ring-type' system counts for each library. It can be seen that the chemical space in L15 is significantly wider as compared to that in L12. L15 contains 4'812 unique ring systems distributed over its 65'580 compounds, whereas L12 contains only 1'540 unique ring systems distributed over its 58'335 member compounds.

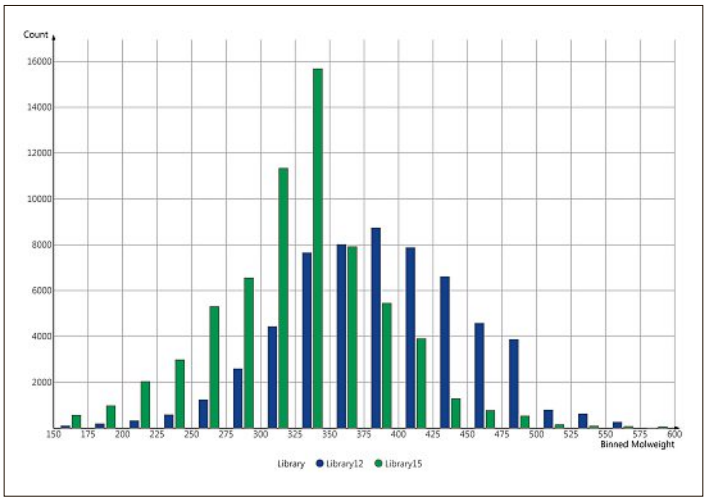

Fig. 7. Molecular weight distribution of L12 (blue) compared to L15 (green).

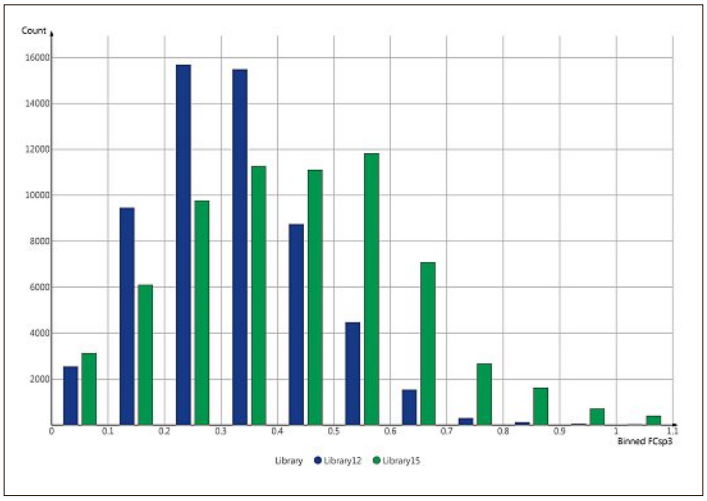

Fig. 8. Fraction sp3 of L12 (blue) compared to L15 (green).

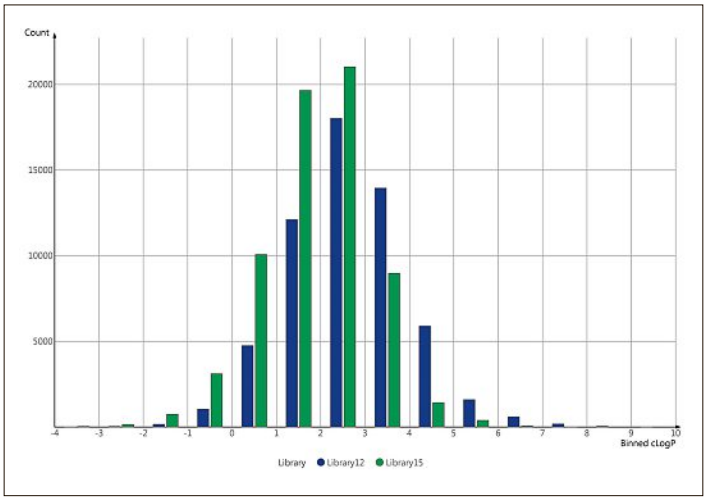

Fig. 9. Calculated logP distribution of L12 (blue) compared to L15 (green).

Fig. 12 shows the Self-Nearest Neighbor Analysis (SNNA) histograms of L12 and L15. In a SNNA, each individual library compound is compared to all other compounds from the same library in order to identify its nearest neighbor. This approach was performed using Schrödinger Canvas and MolPrint2D fingerprints and allows the evaluation of the self-diversity of a library. A common measure to assess similarity/diversity is the Tanimoto-coefficient.[24] The Nearest Neighbor Tanimoto coefficients of all library compounds are depicted as histograms. The trend from L12 to L15 toward lower Tanimoto coefficient values reflects an increased diversity of L15.

Based on this comparative analysis, we conclude that the section of the screening compound collection acquired in 2015 is leadlike, highly diverse, and enhanced with natural product-based and novel sp3-enriched scaffolds. It contains medicinal chemistry project compounds,

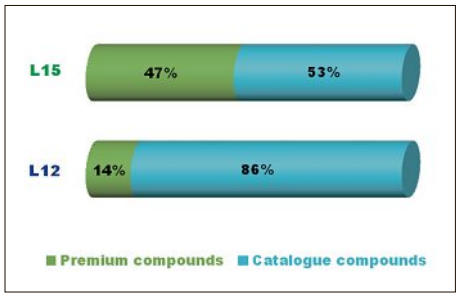

Fig. 10. Premium compounds vs catalogue compounds of L12 (bottom) compared to L15 (top).

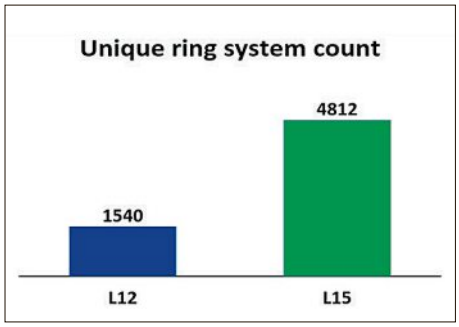

Fig. 11. Unique ring system count of L12 (blue) compared to L15 (green).

macrocyclic compounds, natural products and natural product-derived compounds, agrochemical compounds, designed non-exclusive, semi-exclusive and exclusive compounds, target focused screening compounds and university group compounds. Our efforts had the intended effects. We continue to optimize the Actelion screening compound collection by following the same strategy and consider to implement further measures.

One of these is the TNT-Library described in section 3, which we have just started to implement. We are now validating it in several early projects.

3. The TNT-Library: A Virtual Library of Choice for Virtual Screening Campaigns

To complement lead discovery through physical HTS, we also support drug discovery projects with virtual screening campaigns. For virtual approaches, the same screening compound selection criteria apply.

While it may be feasible to virtually screen any imaginable structure, in the end the hits still need to be physically obtained to confirm the predicted activity *in vitro*. Since not all virtual structures (and in fact, not even all published catalog structures) can be synthesized in reality, this is often the limiting step of a virtual campaign.

In order to gain access to virtual libraries the members of which will be physically available with a high probability and at a reasonable effort, we developed the concept of the TNT-library.

The TNT (Tractable aNd Tangible) library is a virtual library enumerated from in-house available building blocks and proprietary cores. The proprietary cores used to generate the compounds ensure immediate access to in-stock material to quickly generate a large number of synthetically accessible virtual compounds with a favorable IP situation. The building blocks, the cores and the final compounds are carefully selected in order to bring on leadlike and chemically meaningful matter. The synthetic accessibility is ensured by restricting the synthetic steps to twelve well established and robust chemical reactions routinely used in high throughput medicinal chemistry (*i.e.* amide couplings, reductive aminations, Suzuki cross-coupling reactions).

The system has the flexibility to generate several tens of millions of compounds but of which only approximately 5 million are virtually generated. This approach can be used as an idea generator for library design to enrich the screening collection as well as to identify novel hits on a specific target. As the final compounds can be obtained in one to three steps from in-stock material, they can be synthesized by our chemists within two weeks.

The TNT-library is increasingly used at Actelion for 2D and 3D virtual screenings. It is also used to rapidly expand hit structures identified in HTS campaigns.

A 2D molecular fingerprint by a Nearest Neighbor Approach (NNA) was chosen to compare the physical and the virtual collections and measure their complementarity.[7] All the compounds of the query collection (the TNT library) are compared to all the compounds of the reference collection (the Actelion screening compound collection). For each query compound, the nearest neighbor in the reference collection is reported (*i.e.* the structure with the highest Tanimoto similarity index)). Fig. 13 shows that despite using proprietary cores, the TNT library chemical space is highly complementary to the Actelion screening compound collection chemical space.

4. The Mathematics Behind the Scenes

4.1 *The Algorithms beyond Compound Selection*

Four in-house implemented algorithms were used for most of the compound selection process.

A ***chemical descriptor*** encodes molecules as vectors of equal length which are well suited for fast comparisons by a computer. Also, structural clustering based on such descriptors generally appears plausible to a medicinal chemist.

A virtual ***screening algorithm*** was developed to perform the similarity calculations. This algorithm has to be very efficient to allow billions of descriptor comparisons at a useful speed. Furthermore, a ***sampling algorithm*** is needed to draw representative subsamples from a set of molecules.

To visualize the selected molecules and their physico-chemical properties the in-house developed and open-source tool ***DataWarrior*** was used. A multi parameter optimization function (MPO) was implemented to calculate a single score from multiple physico-chemical properties. All implementations were done in Java to be platform independent and run on different operating systems including Windows, MacOS, and Linux.

4.2 *The SkeletonSpheres Descriptor*

This descriptor was developed by Actelion. It is a vector of integers which represents the occurrence of different substructures in a molecule. Five circular layers with increasing bond distance are located for each atom in the molecule. Hydrogen atoms are not considered. This results in five fragments starting with the

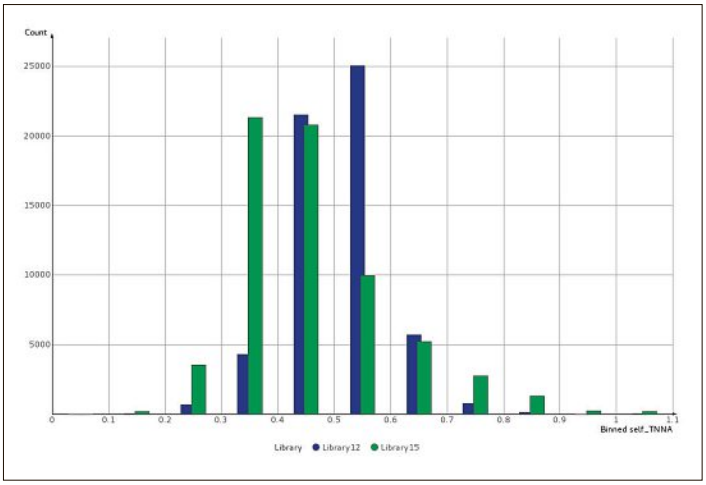

Fig. 12. Self Nearest Neighbor Analysis (SNNA) histograms of L12 (blue) and L15 (green).

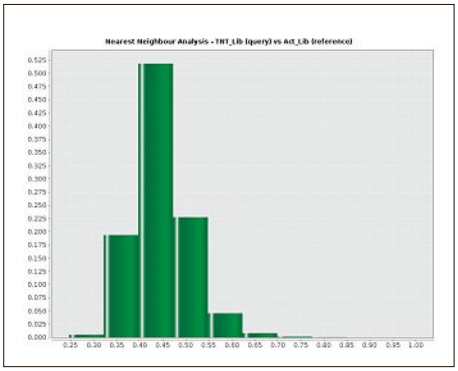

Fig. 13. Displays the nearest neighbor similarity distribution of the TNT collection taken as query collection. It is remarkable that both collections are fully complementary as both libraries are entirely dissimilar to each other.

naked central atom, adding one layer at a time. Every fragment is encoded as a canonical string (id-code), similar to the generation of canonical SMILES.[25] The canonical id-code includes the stereochemistry of the encoded fragment, which is a feature missing in other molecular descriptors. The string is then assigned to one of 1024 fields *n* in a vector. Therefore, the hash value of the id-code is calculated and the corresponding value in the vector is increased by one. The Hashlittle algorithm from Jenkins[26] is used as a binning function which takes a text string as input and returns an integer value between 0 (inclusive) and 1024 (exclusive). In preliminary experiments this hash function showed a good uniform distribution of the generated hash values. To consider the molecular scaffold without the influence of the hetero atoms, the whole calculation is repeated while replacing the hetero atoms with carbon. The resulting hash values are used to increment the corresponding fields in the vector. By adding this skeleton information to the descriptor vector the similarity calculation between two descriptor vectors becomes a bit insensitive to the exact position of the hetero atoms in two molecules. This directs the similarity value toward the perception of similarity by medicinal chemists. For them the exact position of a hetero atom is not as discriminating as it would be for the spheres descriptor without the skeleton coding part. The additional consideration of the scaffold information and the use of a histogram instead of a binary vector distinguishes the SkeletonSpheres descriptor from circular fingerprints.[27] The similarity calculation between two SkeletonSpheres vectors is straightforward: A calculation of *e*=5 spheres is done for all atoms *g* in the molecule serving once as a center atom, which results in *b*=*e g* 2 number of id-codes. The number of fields in the descriptor with a value above *0* is ≤*b*, depending on the number of unique id-codes and hash collisions. As similarity value for the comparison of two descriptor vectors, their overlapping fraction *o* is used. This calculates for two vectors v_1 and v_2 in:

$$o = \frac{\sum_{n}^{i=0} \min(\mathbf{v}_{1,i}, \mathbf{v}_{2,i}) / \max(\mathbf{v}_{1,i}, \mathbf{v}_{2,i})}{n}$$

The similarity values of the Flexophore and of the SkeletonSpheres descriptor were mapped to a comparable value by a scaling function. This was done to ease the use of the descriptors for the scientists in drug discovery. A similarity of 0.85 for a pair of SkeletonSpheres descriptors means that 85% of the chemical structures are overlapping.

4.3 *Molecular Flexibility Calculation*

The conformational flexibility of a molecule is an important parameter influencing induced fit and the entropy of the ligand to protein binding. A frequently used surrogate for molecular flexibility is the number of rotatable bonds. However, this simple approach neglects the fact that the amount of conformational change depends on whether a rotating bond is located in the periphery or center of a molecule. Furthermore, bond rotatability is not a binary property. While some *rotatable* bonds mainly populate one rotation state, others are rather flexible with low rotation energy barriers and multiple energetically equivalent minima. We have developed a computable molecular flexibility measure that takes these effects into account. The algorithm is built into the DataWarrior application[10] and, hence, its source code is freely available as part of the DataWarrior source code. The molecular flexibility is represented by a value ranging from 0.0 to 1.0. Individual bond rotatability values are assessed from torsion statistics of similar bonds derived from the CSD/COD databases (torsion maxima, frequencies, and 50% intervals). The contribution weight of any individual bond rotatability value to an overall molecular flexibility is then determined from the topological bond location, *i.e.* central bonds are weighted higher than those in the periphery. In more detail, the following steps are done: First, all relevant rotatable bonds are determined. These are non-aromatic single bonds, which are not in a ring with less than six members, where both atoms are sp2 or sp3 hybridized and carry at least one more non-hydrogen neighbor, and where a torsion change modifies the relative location of at least one non-hydrogen atom. If there are multiple rotationally redundant bonds, then only one of the redundant bonds is considered. For instance, in chains of conjugated triple bonds the following applies: If at least one terminal sp2/sp3 atom has no external non-hydrogen neighbor, then no single bond is considered rotatable. Otherwise that terminal single bond connecting the smaller substituent is considered the only rotatable bond of the linear atom strand. Then the local environment of any rotatable bond is characterized by its first and second shell of neighbor atoms plus various atom properties like ring membership, aromaticity, and stereo configuration. A canonical representation of the characterizing fragment is created to serve as a bond type specific key into a bond torsion statistics table. This bond angle statistics table was compiled earlier by processing all purely organic, high-resolution X-ray structures of the Cambridge Structural Database (CSD) this way: Any rotatable bond was characterized as described above and its torsion angle added to a torsion histogram associated to this particular rotatable bond type. All bond type specific histograms were smoothened to reduce artefacts. Then all distribution peak maxima and peak widths at a height of 50% were determined. Since the CSD is not an open database and its license prohibits publishing derived statistics data, our public source code contains statistics information from the Crystallography Open Database[28] (COD) instead. In order to calculate the bond rotatability r_b the respective bond key is used to get the associated torsion angle histogram. In rare cases without sufficient bond precedents in CSD or COD, a simple torsion histogram is predicted. An algorithm evaluates the histogram and assigns a value close to 1.0 if the histogram contains three equally distributed wide peaks of similar heights, while histograms with one narrow single peak are considered rather inflexible with a value close to 0.0 (r_b=0 for bonds that are considered non-rotatable when applying above conditions). Then a weighting factor w_b is assigned to every rotatable and non-rotatable bond as follows: For ring bonds w_b=0.33, since ring bonds cannot be changed without typically affecting two other ring bonds. For other bonds w_b=sqrt(2*ssAC/mAC) with ssAC being the number of non-hydrogen atoms on the smaller side of the bond and mAC being the number of non-hydrogen atoms in the molecule. From the bond rotatabilities r_b and bond weights w_b a raw molecule flexibility f_{raw} is calculated as:

$$f_{raw} = \frac{\sum_{b=1}^{n}(w_b \cdot r_b)}{\sum_{b=1}^{n} w_b}$$

with n = number of all hydrogen bonds in the molecule.

For typical molecules the raw flexibility values tend to be well below 0.5, because often many individual bond rotatabilities are 0.0 or small values. To compensate for this effect and to use the desired range from 0.0 to 1.0 more evenly, we apply a scaling function to calculate the final molecular flexibility f_m as:

$$f_m = 1 - (1 - {f_{raw}}^c)^{\frac{1}{c}} \quad \text{with c = 0.7}$$

4.4 *The Virtual Screening Algorithm*

Virtual screening with SkeletonSpheres descriptors means the comparison of integer vectors. Integer calculations are fast on modern CPUs. However, a comparison of a supplier library with one million molecules and an in-house library with 400 k molecules needs 400 billion vector similarity calculations. A single SkeletonSpheres descriptor is defined by 1024 integer numbers. This increases the number of necessary integer operations to more than 400 trillion. Every integer operation uses a min and a max function as it was given in the equation above. Consequently, a number of 800 trillion comparisons, smaller and bigger are needed together with 800 trillion summations. Summing up, a number of 1.6 quadrillion integer operations which are needed to compare the two compound libraries. Fortunately, the vector similarity calculations are independent from each other. So they can be easily parallelized. Our virtual screening algorithm detects the number of processor cores on the computer and creates as many similarity calculation threads as processor cores are available. The descriptors for the two libraries L_1 and L_2 are read bulk-wise from the hard-drive to prevent blockage of a large part of the RAM. The SkeletonSpheres descriptors from the bulks $L_{1,1}$ and $L_{2,1}$ are compared and only results exceeding a pre-defined threshold are written to the output. A server with four CPU sockets was used for the following performance test. Every socket was equipped with an Intel Xeon processor with ten cores. A Xeon processor core is capable of hyper-threading, which results in a total number of 80 cores. The processor clock cycles were specified with 2.4 GHz. After 16.5 h the virtual screening succeeded for the 400 billion vector comparisons. This equals 6.7 million compound similarity calculations per second.

4.5 *Sampling with the OptiSim Algorithm*

Subset selection is daily business in compound acquisition. Therefore, a reliable and sufficiently fast algorithm is needed to perform this task. Almost any algorithm that is used for classification, clustering or multidimensional scaling can be used for sampling. However, an important restriction for our choice of the sampling algorithm, was a time complexity less than quadratic. We decided to implement the OptiSim algorithm, which was developed by R. Clark.[29] This sampling algorithm has a moderate time complexity and works with any similarity metric. The OptiSim algorithm needs two parameters which are fairly robust. Additionally, it was possible to parallelize the OptiSim partially. The parallelization enabled the sampling of large descriptor sets with up to 100 k SkeletonSpheres descriptors.

4.6 *Compound Visualization with DataWarrior*

At the end of the compound selection process with virtual screening and sampling it is desirable to efficiently visualize the selected compounds, to obtain a quick overview of the success of the process. The in-house developed DataWarrior is a visualization tool for all different types of data and has been described recently.[10] Initially, the DataWarrior was developed to visualize data related to chemical structures. Because of its condensed storage of molecular information, more than one million compound structures can be visualized on a desktop computer. Physico-chemical properties of the molecules can be calculated and their distribution visualized. If the distribution has the desired form the compounds are ready for ordering. DataWarrior is a free and open-source tool and can be obtained at *openmolecules.org*.

4.7 *Knime Analytics Platform*

The Knime Analytics Platform is an integral part of our compound selection and library design processes. It is a free, user friendly graphical workbench for data analytics including data management, data transformation, investigation, visualization and reporting. KNIME consists of a series of pieces of program codes called nodes that can be connected in such a way that the input of one node is the output of the previous one. Each node has a dialog in which the user can configure the operation of the node. Mainly, generic Knime nodes, internally developed Actelion nodes as well as RDKit nodes (*http://www.rdkit.org/*) contribute to our Knime workflows.

5. The Reality Check: Biological Interrogation of the New Library Strategy

Today, the Actelion screening compound collection provides a library of chemically diverse small molecules with leadlike properties. Its latest addition of approx. 66'000 compounds, acquired in 2015 (L15), was used to judge the success of the *new library strategy*.

We indeed found several selective, cellular active hits with potencies in the low nM range in different screening projects.

In one case, a hit series with twelve derivatives from a designed library, in which the most potent hit displayed a 16nM activity in two different, orthogonal assays of an oncology target, was identified. Today, this scaffold is explored and optimized through a parallel chemistry approach within the Hit-to-Lead (H2L) team.

In another cellular screening project, the most potent and selective hit was obtained from the macrocycle collection. Remarkably, this macrocycle showed potent activity in a 3D cell culture model constructed from different primary cell types. Even though Actelion, to that date, had only limited experience with the synthesis of macrocycles, the H2L team immediately succeeded to synthesize a considerable number of analogs, resulting in more potent and efficacious derivatives.

After several screening campaigns with L15, it can be stated that the different sublibraries of L15 have different success rates that are target-dependent. As Actelion's therapeutic targets belong to very different protein classes, it is therefore essential to include chemical space as diverse as possible in our library collection. For the screening campaigns performed to date, the highest hit rates are observed among the in-house medicinal chemistry molecules, as well as the agrochemical compounds (validated hit rates around 0.15%); followed by validated hit rates around 0.07% for focused and designed libraries, macrocycles and natural-product derived compounds (Fig. 14). The lowest hit rate is observed among the commercial catalog compounds and the Commercial Virtual Library (CVL) (around 0.03%).

6. Conclusions

The size of the Actelion screening compound collection with its 300'000 compounds is well aligned with a mid-sized company's compound management capabilities. It has previously been shown that the commercially available leadlike space can be covered with a library of this size.[9]

The strategy of the Actelion Screening Compound Collection foresees an annual compound turn-over, *i.e.* every year the oldest 20% are removed from the collection and replaced with new compounds. The advantages of this rolling mode, as compared to a cumulative or static collection, are a low ratio of chemically decomposed or precipitated compounds, a higher availability for re-ordering, and a screening collection that reflects new developments

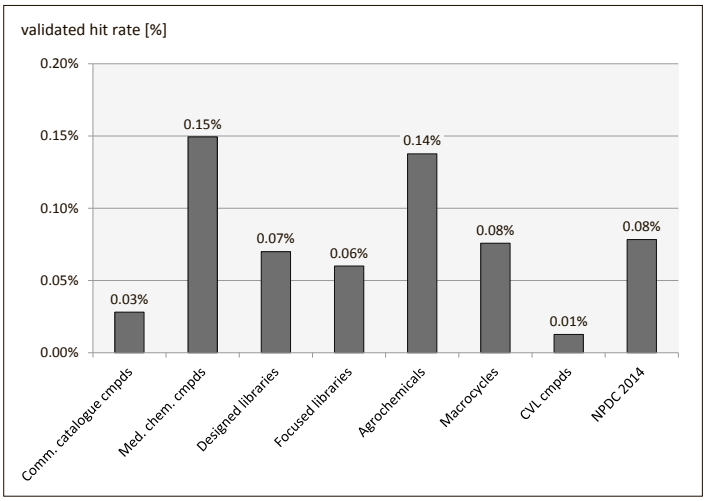

Fig. 14. Validated hit rates for the different sub-libraries of the 2015 annual library (L15) on a panel of different therapeutic targets.

and trends in chemistry. Re-screening the newly added sets can regularly provide novel chemical starting points for follow-up projects within ongoing drug development programs.

A library analysis in 2012 has shown that the screening compound collection is, though chemically diverse, more druglike than leadlike, contained a high proportion of flat compounds and was characterized by a high percentage of commercial *versus* proprietary compounds. As a consequence, the *compound library committee (CLC)* was created, composed of 2 computational scientists, 1 medicinal chemist, 1 Hit-to-Lead chemist and 1 HTS biologist. This committee formulated the *New Library Strategy* with the major aim to enhance the structural library quality. From 2013 to 2016 the *CLC* implemented novel computational tools and methods to rate the value of a compound (MPO score, REOS and PAINS filters). It opened novel sources for compound acquisition to enter new chemical space. In particular, exchanges of proprietary compounds with agrochemical companies were organized.

The commitment of the *CLC* to be transparent in its processes, decisions and actions resulted in broad acceptance among the stakeholders in the Drug Discovery Chemistry and Biology departments.

An analysis of the compounds added in 2015 demonstrated that they are more leadlike than druglike. The L15 set covers previously unexplored chemical space through macrocycles, natural-product derived compounds, agrochemical compounds, designed and focused libraries which are now included as *premium compounds* in the screening compound collection. The increased number of Fsp3 centers in compound structures proves that we "escaped flatland". Importantly, the L15 set of compounds contains a much higher percentage of proprietary compounds than the older L12.

Regular LCMS measurements provide evidence for the high physical integrity of our screening compound collection with >80% of our compounds showing >85% purity (see Fig. 2). The screenings performed so far provided several highly promising hits that are currently followed-up in different *Hit-to-Lead* and *Lead Optimization* programs. It is important to state that the goal of the *CLC* was not to achieve higher HTS hit rates, but to increase the chances of identified hits to serve as the basis of successful early drug discovery programs.

We conclude that the screening results obtained so far vindicate the current strategy of the annual compound turn-over approach, the exploration of novel chemical space, and the change of properties from druglike to leadlike with a higher percentage of proprietary compounds.

Acknowledgements

The authors thank Thomas Weller and Markus Riederer for their trust and support of the *Compound Library Committee*. We also thank the Compound Management team with Joao Silva, Sébastien Guigue, Dora Vieira, Danielle Steiner, Evangelin Baskar and Cyril Bruyère for turning our vision into reality, the HTS team with Geoffroy Bourquin, Alexandre Peter, Serge Brand, Raphael Lieberherr, Laksmei Siek and Solange Meyer for screening the newly acquired compound sets, the LC-MS team with Francois Le Goff and Marco Caldarone for all mass spectrometry measurements, and the Lead Discovery team with Hamed Aissaoui, Sylvia Richard, Shuguang Yuan and Geoffroy Bourquin for developing screening hits into leads. And we thank Joel Freyss and Tobias Fink for developing software tools and Naomi Tidten for stimulating discussions.

Received: May 31, 2017

[1] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug. Deliv. Rev*. **2001**, *46*, 3.
[2] M. M. Hann, *Med. Chem. Commun.* **2011**, *2*, 349.
[3] M. M. Hann, T. I. Oprea, *Curr. Opin. Chem. Biol.* **2004**, *8*, 255.
[4] a) T. T. Wager, R. Y. Chandrasekaran, X. Hou, M. D. Troutman, P. R. Verhoest, A. Villalobos, Y. Will, *ACS Chem. Neurosci.* 2010, *1*, 420; b) T. T. Wager, X. Hou, P. R. Verhoest, A. Villalobos, *ACS Chem. Neurosci.* **2010**, *1*, 435.
[5] a) J. Baell, M. A. Walters, *Nature* **2014**, *513*, 481; b) J. B. Baell, *ACS Med. Chem. Lett.* **2015**, *6*, 229; c) C. Aldrich, C. Bertozzi, G. I. Georg, L. Kiessling, C. Lindsley, D. Liotta, K. M. Merz, Jr., A. Schepartz, S. Wang, *ACS Chem. Biol.* **2017**, *12*, 575; d) S. Jasial, Y. Hu, J. Bajorath, *J. Med. Chem.* **2017**, *60*, 3879; e) J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719.
[6] F. Lovering, J. Bikker, C. Humblet, *J. Med. Chem.* **2009**, *52*, 6752.
[7] T. Kogej, N. Blomberg, P. J. Greasley, S. Mundt, M. J. Vainio, J. Schamberger, G. Schmidt, J. Huser, *Drug Discov. Today* **2013**, *18*, 1014.
[8] R. Mullin, *Chem. Eng. News* **2015**, *93*, 4.
[9] J. B. Baell, *J. Chem. Inf. Model* **2013**, *53*, 39.
[10] T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model* **2015**, *55*, 460.
[11] T. J. Ritchie, S. J. Macdonald, *Drug Discov. Today* **2009**, *14*, 1011.
[12] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615.
[13] a) W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discov. Today* **2002**, *7*, 903; b) W. P. Walters, M. Namchuck, *Nat. Rev. Drug Discov.* **2003**, *2*, 259.
[14] F. Lovering, *Med. Chem. Commun*. **2013**, *4*, 515.
[15] Y. Zheng, C. M. Tice, S. B. Singh, *Bioorg. Med. Chem. Lett.* **2014**, *24*, 3673.
[16] a) Y. C. Martin, J. L. Kofron, L. M. Traphagen, *J. Med. Chem.* **2002**, *45*, 4350; b) C. J. Harris, R. D. Hill, D. W. Sheppard, M. J. Slater, P. F. W. Sotouten, *Comb. Chem. High Through. Screen.* **2011**, *14*, 521.
[17] J. Delaney, E. Clarke, D. Hughes, M. Rice, *Drug Discov. Today* **2006**, *11*, 839.
[18] L. You, R. An, K. Liang, B. Cui, X. Wang, *Curr. Pharm. Des*. **2016**, *22*, 4086.
[19] a) E. A. Villar, D. Beglov, S. Chennamadhavuni, J. A. Porco, Jr., D. Kozakov, S. Vajda, A. Whitty, *Nat. Chem. Biol.* **2014**, *10*, 723; b) F. Giordanetto, J. Kihlberg, *J. Med. Chem.* **2014**, *57*, 278; c) B. C. Doak, B. Over, F. Giordanetto, J. Kihlberg, *Chem. Biol.* **2014**, *21*, 1115.
[20] a) D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2012**, *75*, 311; b) B. L. DeCorte, *J. Med. Chem.* **2016**, *59*, 9295.
[21] J. Aube, *ACS Med. Chem. Lett.* **2012**, *3*, 442.
[22] M. J. Waring, J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett, J. Wang, O. Wallace, A. Weir, *Nat. Rev. Drug Discov.* **2015**, *14*, 475.
[23] M. M. Hann, G. M. Keseru, *Nat. Rev. Drug Discov.* **2012**, *11*, 355.
[24] D. Bajusz, A. Racz, K. Heberger, *J. Cheminform.* **2015**, *7*, 20.
[25] a) Y. C. Martin, E. B. Danaher, C. S. May, D. Weininger, *J. Comput. Aided Mol. Des.* **1988**, *2*, 15; b) J. H. Van Drie, D. Weininger, Y. C. Martin, *J. Comput. Aided Mol. Des.* **1989**, *3*, 225.
[26] *http://burtleburtle.net/bob/c/lookup3.c.*
[27] N. Wale, I. A. Watson, G. Karypis, *J. Chem. Inf. Model* **2008**, *48*, 730.
[28] S. Grazulis, D. Chateigner, R. T. Downs, A. F. Yokochi, M. Quiros, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Crystallogr.* **2009**, *42*, 726.
[29] R. D. Clark, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.