ELSEVIER

feature

An approach towards enhancement of a screening library: The Next Generation Library Initiative (NGLI) at Bayer — against all odds?

Markus Follmann[1], markus.follmann@bayer.com, Hans Briem[2], Andreas Steinmeyer[2], Alexander Hillisch[1], Monika H. Schmitt[2], Helmut Haning[1] and Heinrich Meier[1]

Pharmaceutical companies often refer to 'screening their library' when performing high-throughput screening (HTS) on a corporate compound collection to identify lead structures for small-molecule drug discovery programs. Characteristics of such a library, including the size, chemical space covered, and physicochemical properties, often determine the success of a screening campaign. Therefore, strategies to maintain and enhance the overall quality of screening collections are crucial to stay competitive and to cope with the 'novelty erosion' that is observed gradually. The Next Generation Library Initiative (NGLI), the enhancement of Bayer's HTS collection by 500 000 newly designed compounds within 5 years, is addressing exactly this challenge. Here, we describe this collaborative project, which involves all internal medicinal chemists in a crowd-sourcing approach, as well as selected external partners, to reach this ambitious goal.

Introduction

HTS of large compound collections has proven to be of great value in the search for starting points in medicinal chemistry optimization programs [1–4]. Despite recent advances in alternative approaches, such as virtual screening [5], *de novo* structure-based design [6,7], fragment-based lead finding [8], and screening of DNA-encoded libraries [9,10], HTS of corporate compound libraries continues to have the highest impact on small-molecule lead discovery [11] and was the origin for numerous marketed drugs [3], including DPP IV inhibitor sitagliptin (Januvia) and factor Xa inhibitor rivaroxaban (Xarelto), to name just two examples.

It is widely accepted among medicinal chemists that size (number of compounds), quality (structural attractiveness [12] and purity), molecular properties (drug-like or lead-like), and diversity (chemistry space covered [13]) of a screening library are key determining factors for the success of a screening campaign [14,15]. Another important aspect for drug discovery conducted in a competitive environment, such as the pharmaceutical industry, is achieving patentable novelty for a lead series. This can pose a considerable challenge when working on hits found in a HTS collection that has grown over time, including acquisition campaigns for largely nonexclusive, commercially available compound sets.

Here, we describe our approach towards enhancement of our corporate screening collection by 500 000 newly designed compounds.

Conclusions derived from analysis of current screening collection and project goal

Over the past 20 years, the Bayer HTS screening library has grown from $<1 \times 10^6$ compounds to its current size of $\sim 4 \times 10^6$ compounds. The historic growth was mainly driven by adding project compounds and in-house parallel synthesis libraries. On top, commercial screening libraries were purchased containing nonexclusive or semiexclusive substances. These oppor-

https://doi.org/10.1016/j.drudis.2018.12.003

tunistic expansions were complemented by systematic enrichment with compounds from our agricultural unit and, last but not least, the incorporation of 870 000 compounds from the Schering HTS set that became available through the Schering acquisition in 2006 [16]. This one-time boost of the Bayer screening deck with lead- and drug-like compounds had a positive impact on HTS success in several cases where the existing collection had failed to deliver any useful novel starting points. To give two examples, a screen for stimulators of soluble guanylate cyclase delivered a previously unknown series of imidazo[1,2-a]pyridines ultimately leading to a clinical candidate. In addition, a novel series of 1,4-dihydropyridine c-Met kinase inhibitors was discovered. Thus, increasing the size of a screening deck with high-quality lead- and drug-like compounds dissimilar to the original set [16] will result in a higher probability of success in HTS campaigns. Consequently, when we set out to develop our NGLI, we had to clearly define the quality and differentiation criteria for compound inclusion and had to provide tools and establish processes to ensure adherence to such guidelines.

When we analyzed the novelty of our screening collection in 2014, we were surprised that 18% (0.64×10^6) of all compounds were commercially available from various sources and 54% (1.88×10^6) were represented in a generic sense in PubChem [17]. This analysis substantiated and quantified the common perception of many in our drug discovery project teams that substantial and, over time, increasing, efforts were generally required to reach novelty when starting with hits from our HTS collection. To have an impact on the overall composition of our screening library, we proposed to add 500 000 compounds to compensate for the observed 'novelty erosion' and we defined novelty and full exclusivity as prime quality requirements for our NGLI compound sets.

Further important quality aspects to consider when enhancing screening libraries are physicochemical properties and the overall diversity of compounds. Calculation of the *in silico* oral phys-chem Score (oPCS, scores range from 0: oral drug like/small to 10: not drug like large/lipophilic) [18] distribution of our current screening pool showed that ~75% of all compounds had a score of 0–4, with an average score of 2.9 (corresponding to an average corrected molecular weight of 397 and calc $logD_{7.5}$ of 3.0). Although we were rather satisfied with the status quo, we decided to set ourselves even more ambitious goals for the enhancement set and, therefore, set a target of achieving an overall reduction of at least 1 unit in the average oPCS. In addition to improving calculated physicochemical properties, we wanted to implement some modifications affecting the covered chemical space. Given the unfathomable depth of drug-like structural space [13], it is apparent that closing of gaps in a compound collection is almost impossible. Nevertheless, we reasoned that one convincing approach to increase diversity without increasing molecular weight would be through the introduction of more saturation and three-dimensionality [19]. Traditionally, medicinal chemistry structures and, consequently, screening collections are strongly biased towards flat molecules assembled around (hetero-)aromatic scaffolds. Calculation of the average Fsp^3 [19] for our HTS-pool showed a value of ~0.3. Thus, we planned to increase this fraction significantly for the novel part of the library. In summary, we set ourselves the following project goal for the NGLI: enhancement of the HTS pool by 500 000 novel compounds over a period of 5 years and that all compounds should have the following average properties: oPCS <2, molecular weight <400, calc $logD_{7.5}$ <3.0, and Fsp^3>0.4.

Overview of set-up and design strategies

The framework of NGLI requires us to design and procure the prospected 500 000 new compounds within a period of 5 years. To achieve an optimal balance of overall diversity with practicability in actual syntheses, the guiding criteria were set as follows: each library should be based on a scaffold ideally having two to three handles for diversification. To achieve novelty, the scaffold and/or substitution pattern should provide two points of deviation from literature precedence. A maximum of 20 late-stage intermediates should, via final diversification, enable access to libraries comprising 400–600 compounds. The proposed synthesis scheme, though unprecedented, should not require more than 10–12 linear steps and should be plausible as judged by experienced medicinal chemists, including parallel synthesis experts, in our internal review panel. Realizing that our in-house community is a unique source for explicit (general rule-based, e.g., phys-chem properties) and implicit (individual preference and group experience, e.g., wanted and unwanted functionalities [20]) medicinal chemistry knowledge, we chose a 'crowd-sourcing' approach to collect and select the majority of NGLI libraries from this resource. In addition, through close collaboration with selected external partners, we complemented this approach by enhancing the HTS deck with libraries of peptidic and nonpeptidic macrocycles or other modalities incorporating specific know-how. The overall guidelines for library proposals are summarized in Box 1.

To facilitate the library design and enumeration process, thus allowing the chemists to focus on the creative process of idea generation, a tracking database was provided to which all library proposals were uploaded. We also developed a cheminformatics workflow for library enumeration, schematically outlined in Fig. 1.

Complementary to this crowd-sourcing approach and to increase the likelihood of finding pharmacologically active compounds, we made use of the vast amount of internal and published structure–activity relationship (SAR) and protein structural knowledge. We formed four target class focused 'design teams' each comprising medicinal and computational chemists. The task of those teams was to implement a strategy to make best use of the available protein structure and SAR data and apply that to the design of 10 000 compounds per year and team. The 'GPCR & Ion Channel' team approached the challenge with target-subfamily specific likeness models. Given that there is no common structural motif for all G-protein-coupled receptors (GPCRs) or ion channel ligands, internal and external SAR data (ChEMBL) for homologous target proteins were extracted and cleaned. Molecules were represented by extended connectivity fingerprints (ECFP)-4 fingerprints and physicochemical properties and a random forest algorithm was trained to distinguish actives from inactives. The likeness models were applied to selected novel compounds from virtual libraries (virtual compounds and possible bioisosteric databases). The 'Protease' design team primarily used the vast amount of co-crystal structures to set up an automated docking and structure-based design approach. Special emphasis was placed on the selection of the sidechains of the protease inhibitors given their importance for interaction with the target protein. The 'Kinase' design team primarily designed new cores and prioritized those by docking into

BOX 1

NGLI library proposals guidelines

Molecular properties (average): oral phys-chem score [18] (oPCS) <2, MW <400, calc logD7.5 <3.0, Fsp3 >0.4.
Novelty: scaffold and/or substitution pattern should provide two points of deviation from literature precedence.
Synthetic tractability: maximum of 10–12 linear steps, plausible synthetic scheme.
Structural features: no unwanted functionalities [20].

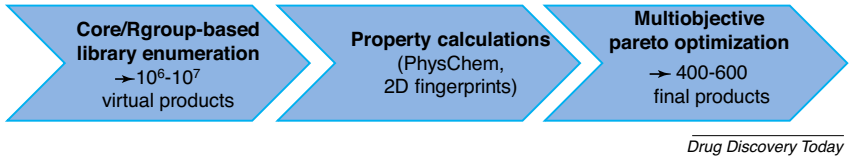

FIGURE 1

Library enumeration workflow: Step 1: a core scaffold representing the chemists' idea is decorated with chemically feasible reagents chosen from a set of ~15 000 building blocks preselected for desirability and easy availability from commercial sources. The size of virtual libraries initially generated by this procedure typically ranges from 10^6 to 10^7 virtual products. Step 2: for each virtual product, various physicochemical properties (e.g., corrected molecular weight, predicted solubility, calc $logD_{7.5}$ and the combined descriptor oPCS), as well as 2D functional connectivity fingerprints (FCFP-4) [21] are calculated. Step 3: to select the most appropriate building blocks for the final combinatorial matrix to be produced (e.g., 20 × 20 building blocks giving rise to 400 final products), we used a multiobjective evolutionary algorithm [22] to find building-block combinations lying on a Pareto-optimal front (i.e., offering the best possible compromise with respect to physicochemical properties and feature diversity of the library).

X-ray or homology models. In addition, in cases where the protein structural information was not sufficient, ligand-based pharmacophore approaches were applied. The fourth design team, 'Epigenetic Targets', used a mix of all the above-mentioned design approaches, because they were faced with nonhomologous targets and, as a result, ligand- and protein structure-based approaches were applied depending on the concrete targets to be addressed. All four design teams used selected representatives of the target class in which they had previous project background. Additional targets were added based on a survey of early and future targets for which we plan to initiate drug discovery projects. The overall workflow for proposal handling is depicted in Fig. 2.

Logistics and quality control

To guarantee quick value generation, a process was elaborated to enable the seamless flow of data and new compounds into the Bayer systems. The fast availability of compounds for HTS campaigns and for retrieval purposes is vital to keep active research projects on schedule. To incorporate more than 100 000 test compounds and 5000 final-stage intermediates per year, very robust and reliable workflows had to be created.

A major part of the compound management process was outsourced to external partners located worldwide. Considerable effort was invested to establish workflows for handling data and compounds with minimal manual intervention across languages, cultures, and time zones. Several research IT systems were modified, and new tools were built, to securely and efficiently support the workflows. As a result, newly delivered substances were incorporated into the screening file in less than 4 weeks.

The compounds made by our partner organizations were thoroughly analyzed (NMR, LC–MS, >90% purity required for 95% of compounds), weighed into vials and formatted into different types of plate according to Bayer's specification. A minimum of 30 mg of material per compound was requested to secure sufficient stock amounts. Upon arrival at Bayer, internal quality control was performed for all compounds before addition to the overall HTS screening deck. To our delight, it turned out that all plates delivered displayed good quality, and few errors occurred during the compound management process at the partner laboratories. Thus, efficient processes were successfully implemented with different partners allowing smooth incorporation of a large number of novel compounds into the Bayer HTS screening deck. In addition, the availability of final-stage intermediates in multigram quantities allowed efficient resynthesis and easy access to analogs.

Properties of the produced libraries

As outlined earlier, designed libraries should exhibit favorable physicochemical properties, as well as diversity with respect to our existing compound collection. With a significant number of new compounds in hand, we were able to compare these new molecules to the old part of the screening file. As can be deduced from Fig. 3, we were clearly able to achieve property distributions desirable for lead-like compounds. In addition, on average, we could significantly improve these properties compared with our historical Bayer HTS pool.

Furthermore, we performed a similarity analysis between each NGLI library compound

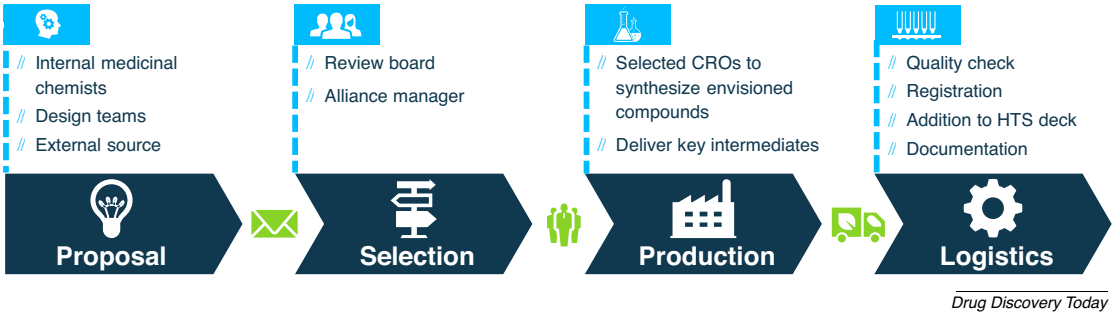

FIGURE 2

Library ideas from all sources are reviewed by an internal expert team comprising experienced medicinal and computational chemists. Proposals fulfilling our criteria outlined in the main text are selected for production by partner labs at selected contract research organizations (CROs), a process closely supervised by our dedicated alliance manager. Regular feedback loops between CROs and Bayer experts ensure timely adaptation of designs and synthesis routes to unforeseen synthetic challenges, leading to high overall success rates in the synthesis of library scaffolds and/or late-stage intermediates and individual target compounds. Once a library has been produced, the compounds are shipped according to different formats as described in the main text.

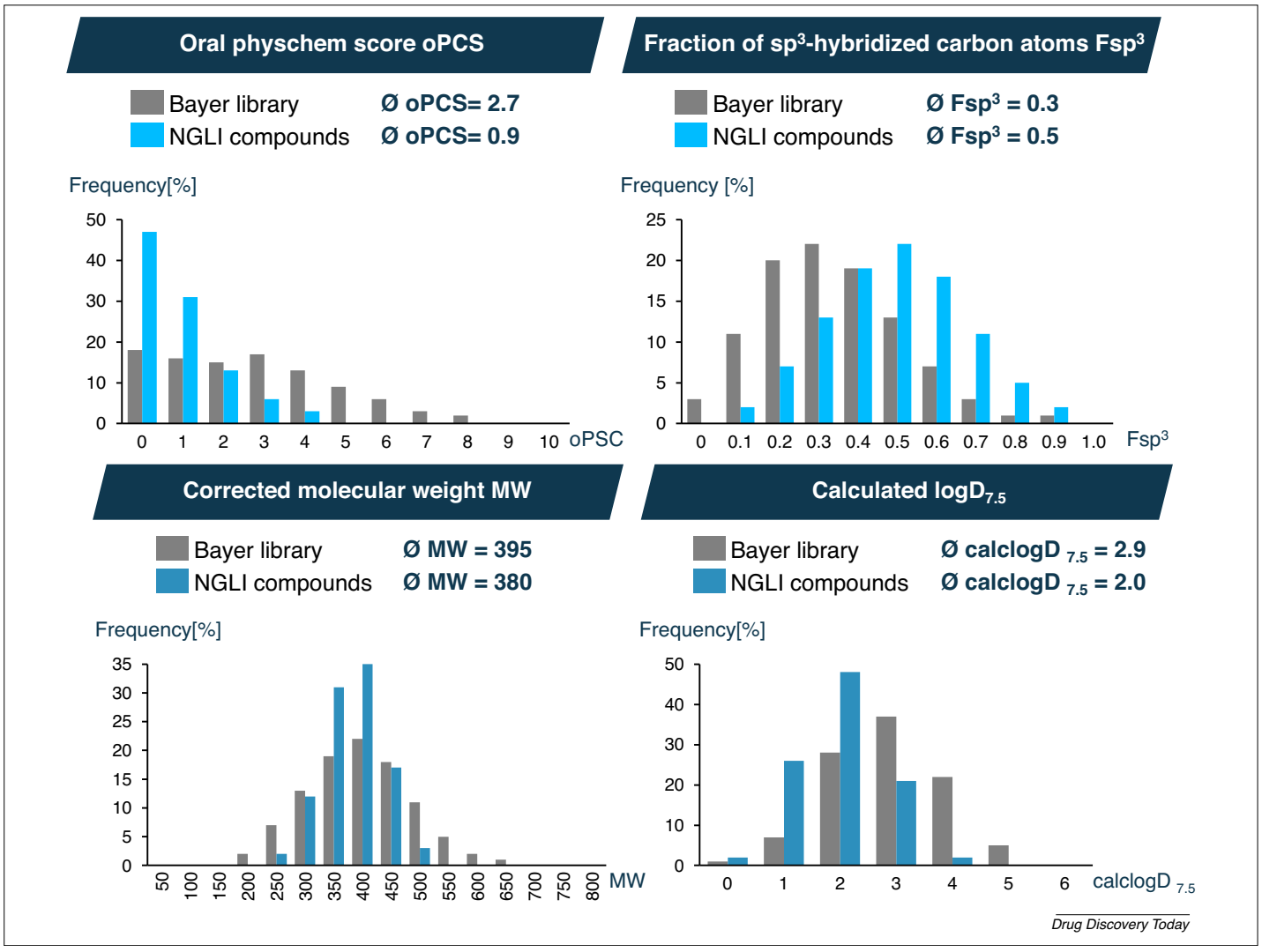

FIGURE 3

For various structural and physicochemical properties typically used to judge the lead-likeness of a given library, we compared the frequency distributions and mean values of the newly designed Next Generation Library Initiative (NGLI) compounds (in blue) with those derived from Bayer's current high-throughput sequencing (HTS) pool (in gray). For all four properties analyzed, the NGLI compounds showed favorable property distributions, on average superior to the existing HTS pool. Therefore, we conclude that future HTS hits derived from NGLI compounds should constitute good starting points for compound optimization.

and its closest neighbor in the Bayer HTS pool (Fig. 4). Overall, both compound pools are clearly distinct from each other, with few library compounds exhibiting similarity coefficients above 0.7 to their respective nearest neighbors in the HTS pool. This distribution clearly reflects the strict novelty criteria applied to each proposed library scaffold, thus expanding the chemical space covered by our future HTS pool.

We also analyzed the quality and success of the design approaches. After 2.5 years, ~250 000 compounds were delivered over time and added to the HTS library. Of those novel compounds, 6900 were found as primary HTS hits and further processed in IC_{50} testing; 4000 were active (<10 μM) and, of those, 1300 were highly active (<1 μM). Of the highly actives (<1 μM), 46% stemmed from design approaches (active on the target class they were designed for), 33% derived from design approaches but were active on a target class they were not designed for, and 23% of highly actives were from chemistry-driven approaches. Although statistically not significant and preliminary, the novel compounds on average had lower hit rates relative to our standard HTS compound collection. We assume that the lower hit rates reflect the better physicochemical profiles and diversity of the novel compounds. Although a library initiative of this magnitude with its focus on novelty and exclusivity is designed for long-term impact, it has been rewarding to see that NGLI-derived HTS hits have already entered SAR campaigns. A more detailed analysis of hit-rates and origin of hits will be published in due course.

Concluding remarks and outlook

The NGLI represents a unique exercise in the history of Bayer aiming at rejuvenating our screening library with 500 000 novel, lead-like compounds with a clear focus on further strengthening our capabilities in small-molecule lead discovery. A special feature of NGLI is its collaborative set-up, making best possible use of in-house knowledge involving hundreds of scientists by a crowd-braining approach complemented by select external partners. In addition to providing access to novel compounds and intermediates in multigram quantities, the initiative provides a rich source of exclusive information, such as a virtual chemistry space enabling *in silico* screening of rapidly accessible compounds. Furthermore, designing libraries of compounds populating unprecedented chemi-

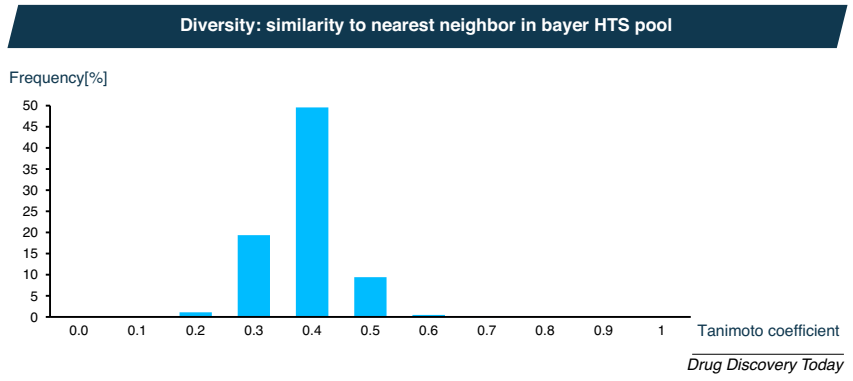

FIGURE 4

For each Next Generation Library Initiative (NGLI) compound, the closest analog in Bayer's high-throughput sequencing (HTS) pool was determined by calculating Tanimoto similarity coefficients using circular extended connectivity fingerprints (ECFP)-4 fingerprints [20]. The frequency distribution of the Tanimoto coefficients is centered around a low similarity value of 0.4. There were only a few NGLI compounds exhibiting significant similarity (>0.7) to an existing compound from the HTS pool. Thus, it can be concluded that the NGL initiative clearly broadens the chemical space of Bayer's HTS collection.

cal space requires the invention and exploration of novel synthetic routes, representing a plethora of synthetic chemistry knowledge that we plan to use for machine-learning approaches to further enable chemical synthesis with artificial intelligence.

Proposing novel chemistry-driven compound libraries with excellent physicochemical properties and target-family based design approaches by best possible integration of structural biology knowledge and computational design strategies should ideally provide numerous quality hits in future HTS campaigns. We can already conclude that compounds from NGLI have had a positive impact on the outcome of our latest screens. Nevertheless, a more detailed and/or comprehensive analysis can only be shared at a later date on the basis of sufficient data.

Acknowledgments

We thank Jens Ackerstaff, Hartmut Beck, Andreas Beckmann, Kristin Beyer, Niels Böhnke, Anne Bonin, Michael Brands, Nico Bräuer, Thorsten Blume, André Dieskau, Jan Dreher, Knut Eis, Sebastian Essig, Pascal Ellerbrock, Stefan Gradl, Alexey Gromov, Michael Hahn, Simon Herbert, Stefan Jaroch, Sarah Johannes, Johannes Köbberling, Florian Kölling, Jorg Kroll, Ingo Kühlborn, Lara Kuhnke, Ngoc Mai Le Cong, Niels Lindner, Mario Lobell, Anne Mengel, Katharina Meier, Stefanie Mesch, Hideki Miyatake-Ondozabal, Jeffrey Mowat, Steffen Müller, Stefan Mundt, Thomas Neubauer, Adam Nitsche, Duy Nguyen, Olaf Panknin, Karsten Parczyk, Hartmut Rehwinkel, Ulrike Röhn, Susanne Röhrig, Jens Schamberger, Hartmut Schirok, Carsten Schmeck, Holger Siebeneicher, Stephan Siegel, Rene Spang, Timo Stellfeld, Alexander Straub, Ton ter Laak, Pierre Wasnaire, Jens Willwacher, Lars Wortmann, Hans-Peter Wrona-Metzinger, Greta Ziemann, and Dmitry Zubov for valuable contributions to the NGLI-project. Special thanks to Lisa Candish for checking the manuscript.

References

1 Gong, Z. *et al.* (2017) Compound libraries: recent advances and their applications in drug discovery. *Curr. Drug Discov. Technol.* 14, 216–228
2 Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today* 11, 277–279
3 Macarron, R. *et al.* (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195
4 Karawajczyk, A. *et al.* (2015) Expansion of chemical space for collaborative lead generation and drug discovery: the European Lead Factory. *Drug Discov. Today* 20, 1310–1316
5 Haga, J.H. *et al.* (2016) Virtual screening techniques and current computational infrastructures. *Curr. Pharm. Des.* 22, 3576–3584
6 Hillisch, A. *et al.* (2015) Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem* 10, 1958–1962
7 Hartenfeller, M. and Schneider, G. (2011) De novo drug design. *Methods Mol. Biol.* 672, 299–323
8 Lamoree, B. and Hubbard, R.E. (2017) Current perspectives in fragment-based lead discovery (FBLD). *Essays Biochem.* 61, 453–464
9 Zimmermann, G. *et al.* (2016) DNA-encoded chemical libraries: foundations and applications in lead discovery. *Drug Discov. Today* 21, 1828–1834
10 Franzini, R.M. and Randolph, C. (2016) Chemical space of DNA-encoded libraries. *J. Med. Chem.* 59, 6629–6644
11 Brown, D.G. and Bostrom, J. (2018) Where do recent small molecule clinical development candidates come from? *J. Med. Chem.* 61, 9442–9468
12 Bickerton, G.R. *et al.* (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98
13 Reymond, J.-L. (2015) The Chemical Space Project. *Acc. Chem. Rev.* 48, 722–730
14 Villar, H.O. and Hansen, M.R. (2009) Design of chemical libraries for screening. *Expert Opin. Drug Discov.* 4, 1215–1220
15 Kogej, T. *et al.* (2013) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. *Drug Discov. Today* 18, 1014–1024
16 Schamberger, J. *et al.* (2011) Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. *Drug Discov. Today* 16, 636–641
17 Kim, S. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.* 44, 1202–1213
18 Lobell, M. *et al.* (2006) In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* 1, 1229–1236
19 Lovering, F. *et al.* (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756
20 Wunberg, T. *et al.* (2006) Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* 11, 175–180
21 Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754
22 Zitzler, E. *et al.* (2000) Comparison of multiobjective evolutionary algorithms: empirical results. *Evol. Comput.* 8, 173–195

Markus Follmann[1,*]
Hans Briem[2]
Andreas Steinmeyer[2]
Alexander Hillisch[1]
Monika H. Schmitt[2]
Helmut Haning[1]
Heinrich Meier[1]

[1]Bayer AG, Pharmaceuticals, Aprather Weg 18a, 42113 Wuppertal, Germany
[2]Bayer AG, Pharmaceuticals, Müllerstr. 178, 13342 Berlin, Germany

**Corresponding author.*