Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-019-00215-x

Diversifying chemical libraries with generative topographic mapping

Arkadii Lin[1,2] · Bernd Beck[2] · Dragos Horvath[1] · Gilles Marcou[1] · Alexandre Varnek[1]

Received: 10 April 2019 / Accepted: 15 July 2019

Abstract

Generative topographic mapping was used to investigate the possibility to diversify the in-house compounds collection of Boehringer Ingelheim (BI). For this purpose, a 2D map covering the relevant chemical space was trained, and the BI compound library was compared to the Aldrich-Market Select (AMS) database of more than 8M purchasable compounds. In order to discover new (sub)structures, the “AutoZoom” tool was developed and applied in order to analyze chemotypes of molecules residing in heavily populated zones of a map and to extract the corresponding maximum common substructures. A set of 401K new structures from the AMS database was retrieved and checked for drug-likeness and biological activity.

Keywords Generative topographic mapping · Chemical library diversity enrichment · Big data

Abbreviations

GTM	Generative topographic mapping
FS	Frame set
LLh	Likelihood
AD	Applicability domain
RBF	Radial basis function
DB	Database
AMS	Aldrich Market Select
BI	Boehringer Ingelheim
MCS	Maximum common substructure

Introduction

Structural library enrichment is an important task for pharmaceutical industry. The number of hits selected in screening campaigns depends on drug-likeness and diversity of the underlying screening set. To be efficient in drug-discovery, the existing screening pool needs to be regularly updated in order to include new chemotypes.

One can suggest two different scenarios of the screening pool enrichment with new chemical matter: computer-aided enumeration of virtual structures under some constraints (e.g. molecular weight, LogP, etc.), or selection of existing structures from an external database. Recently, several attempts were made to create a workflow for an efficient molecular de novo design [1–5]. However, synthetic feasibility of virtual structures including synthetic routes and optimization of reaction conditions still need to be assessed. The second scenario is more practical because new structures selected as a result of comparison of two data sets (a reference set and an external set) do exist and can be purchased or synthesized following the reported in the literature procedure.

Different approaches of chemical databases comparison were reported so far: cell-based clustering [6], pairwise distance analysis [7], and some dimensionality reduction methods (principle component analysis (PCA) [8], self-organizing maps (SOM) [9], generative topographic mapping (GTM) [10]) providing with the visualization support. GTM is a method of choice in this study because of its clear advantage over PCA and SOM approaches. GTM approximates data probability distribution functions both in the initial D-dimensional space of molecular descriptors and in the 2D latent space [11] and, thus, represents a fuzzy-logics generalization of Kohonen maps, supporting predictive modeling of continuous or categorical property landscapes.

✉ Bernd Beck
bernd.beck@boehringer-ingelheim.com

✉ Alexandre Varnek
varnek@unistra.fr

[1] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal Str., 67081 Strasbourg, France

[2] Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88397 Biberach an der Riss, Germany

Published online: 12 August 2019

Recently we demonstrated that GTM represent an efficient tool for comparison of large chemical libraries FDB-17 and PubChem-17 [12]. Hierarchical zooming technique [13] was successfully applied in [12] in order to analyze the chemotypes of molecules populated selected zones and to highlight the scaffolds present exclusively in FDB-17.

In this study, the zooming technique was automatized and coupled to a maximum common substructure (MCS) extraction protocol ("AutoZoom" tool). The developed tool was used for enrichment of the in-house collection of Boehringer Ingelheim (further on referred to as the "BI Pool") by the compounds from the commercial Aldrich-Market Select (AMS) database. A drug-likeness and an activity profile of selected AMS compounds against 749 biological targets were assessed using the ChEMBL data-driven predictor based on Universal GTMs [14, 15].

Data

Boehringer Ingelheim (BI) is steadily committed to innovation in medicinal chemistry and is hence interested in new compounds featuring new scaffolds. At the same time, new structures have to be synthesizable and should have the potential to be active.

As a basis in this work we used an in-house collection of drug-like compounds provided by BI (BI Pool) which contained more than 1.7M structures. The source for novel compounds was the publicly available AMS collection of purchasable compounds containing more than 8.2M items (http://www.aldrichmarketselect.com). The data was standardized by ChemAxon's standardizer tool using a list of rules, such as aromatization, removing isotopes, removing stereo, standard representation of N-oxides including nitro group, etc. [16].

Method

The computational workflow consists of three parts. First, the mapping of AMS chemical space was undertaken by calibrating a pertinent GTM manifold, followed by projection of entire AMS and BI Pool collections. Then, hierarchical zoom was performed for selected areas of the map followed by MCSs extraction. The most of interest represented some zones exclusively populated by AMS compounds. The latter were extracted and profiled using universal GTMs described in our previous papers [14, 15]. To this purpose, the publicly available virtual screening webserver of the Laboratory of Chemoinformatics (http://infochim.u-strasbg.fr/webserv/VSEngine.html) was employed. In addition, simple molecular properties, like LogP, number of H-bond donors and acceptors, molecular weight, and TPSA, were computed using ChemAxon's JChem engine [17].

GTM training

The GTM method relates the data points positions in the initial N-dimensional space and in the latent 2D space. The GTM algorithm is described in a range of publications [10–12, 18]. Briefly speaking, GTM injects a 2D hypersurface (*manifold*) into a multidimensional data space populated by a set of representative items (the Frame Set, FS). The algorithm fits the manifold to the FS data distribution by changing the positions of Radial Basis Function centers and, hence, maximizing the data log likelihood (*LLh*). At the next stage, the data points are projected on the manifold followed by the manifold unbending. Each compound in the latent space is represented by a vector of normalized probabilities (*responsibilities*) computed in the nodes of a square grid superposed with the manifold. In turn, the entire data set can be characterized by a vector of cumulative responsibilities. This enables the user to perform an efficient data sets comparison as well as QSAR/QSPR studies [10, 11, 19].

In our early study [12], frame set compounds were randomly selected from large chemical libraries. Here, a FS containing 25K AMS compounds of controlled diversity (featuring no two compounds more similar than a given threshold) was prepared. To measure the dissimilarity, Soergel distance [20] basing on Morgan fingerprints [21, 22] of radius 4 was computed. FS compounds are expected to represent a non-redundant, representative "core" of spanned chemical space. They are not subjected to any other specific constraints, meaning that any state-of-art molecular descriptor/dissimilarity metric can be equally well used for selection.

The GTM manifold was trained using an incremental algorithm described by Gaspar et al. [23]. The parameters were taken from the previous study [12]. The experience of previous projects [12, 24, 25] showed that the usage of ISIDA descriptors is a good choice for GTM training. The initial descriptor space features ISIDA counts of sequences of 2 and 3 atoms, colored by their CVFF [26] force field types and including formal charge information (IA-FF-FC-2-3) [27, 28]. Fragmentation of the FS compounds produced 6142 distinct fragments. However, the vast majority thereof is sparsely populated: only 798 terms were considered for actual manifold construction (the descriptors for which standard deviation over the FS compounds exceeds 2% of their value range width). This (or closely related) fragmentation schemes were often selected by evolutionary [29] map tuning procedures [12, 15]. Other adopted map parameters include resolution (841 nodes), the number of RBFs (324), the regularization coefficient (3.236), RBF width (0.4), and incremental block size (10K compounds).

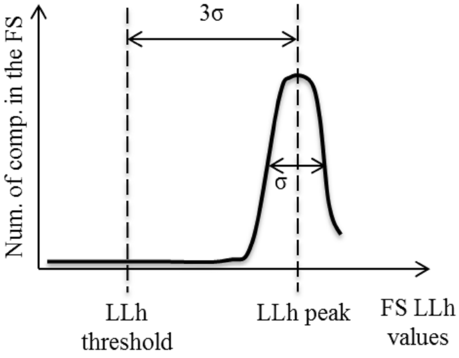


Fig. 1 GTM Applicability Domain is identified by log likelihood threshold $LLh_0 = LLh_{peak} - 3\sigma$. Here, LLh_{peak} and σ are, respectively, a position and with of a Gaussian function which fits the LLh distribution

When the Expectation–Maximization algorithm used to train the manifold has achieved a certain level of convergence ($LLh_{new} - LLh_{prev} \leq 0.001$), the entire data was projected, and the outliers (the structures positioned far away from the manifold) were removed. To do so, a new strategy for GTM applicability domain (*AD*) identification was suggested where a Gaussian is fitted to the FS compounds distribution minimizing the root mean square error. Once the fitting is done, the LLh threshold is determined as the LLh value with the highest population (peak) minus three Gaussian widths ("3σ" rule, Fig. 1).

For visualization and analysis purposes, property and fuzzy class landscapes are used to "color" the map. To this goal, the mean class/property value in each node is taken as responsibility-weighted means of class labels/property values of resident items [11]. In consequence, areas of interest (for example, clusters of nodes exclusively populated by AMS compounds) can be easily highlighted.

Zooming

GTM landscape analysis is the following step in the library comparison process. The goal is to bind a certain chemotype to a particular area on the map. In simple cases, map zones (square clusters of nine nodes) do indeed contain structurally quite homogeneous populations of residents. If so, it is straightforward to search for common scaffolds or maximum common substructures (MCSs). However, if too many compounds (e.g. more than 1000 items) reside in one zone, searching of common scaffolds or MCSs is not efficient. Therefore, since the algorithm detects highly populated zones, zooming is automatically applied. For this purpose, the compounds for which the sum of its responsibilities within the zone is higher than 0.95 are selected and used as frame set source for the fitting of a new GTM manifold (using the same setups as those of the global map). For this purpose, the FS—of minimal 1000, but maximal 10% of the local compound pool size—is randomly selected. The "submap" is likewise checked for the zones with population exceeding 1000 items. If necessary, the procedure is repeated (multi-level zooming). If a zone contains less than 1000 compounds, it will be analyzed as such, without further zooming.

Maximum common substructure (MCS) searching

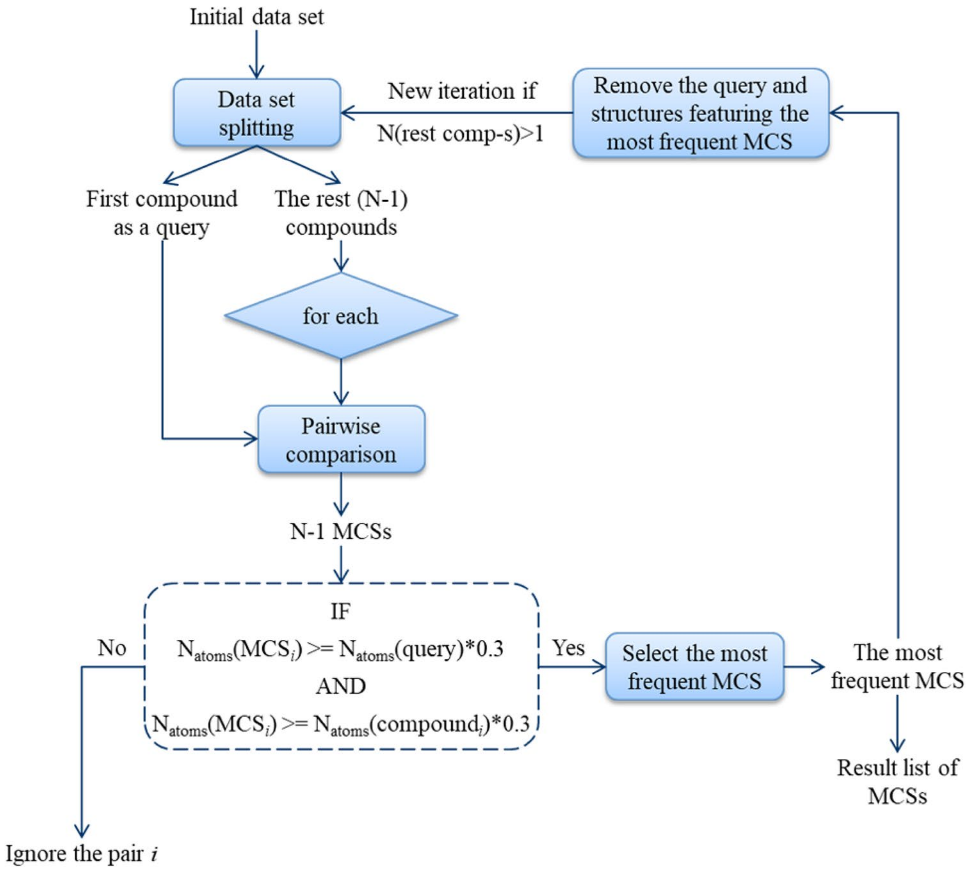
Responsibility patterns (RP) have been used to identify the shared underlying features (scaffolds, substructures, pharmacophore patterns) for a chosen area on the map [19, 30]. Compounds sharing a same RP will typically share some common structural features that are further manually processed to annotate the map. This is a tedious and error prone task. As an alternative, it is proposed here to exploit MCS search to automatically highlight shared features. Our solution is based on ChemAxon's JChem engine [17].

The problem of MCS searching for a set of compounds was already discussed earlier by Hariharan et al. [31]. The authors showed that in some situations, the intersection of pairwise MCS search is empty or results in small, non-specific substructure, while the molecules in a given set share large and complex substructures. The problem is that such common substructure of a compound set is not the maximum common substructure of any compounds pair. As a solution, Hariharan et al. enumerated all maximal cliques for each pair of molecules, and then intersected the generated lists. The so called multi-MCS is the largest of the identified substructure that is common to all compounds in the set.

However, when the molecule set is very large, the idea to return a single multi-MCS does not work anymore. In this case, we aimed at identifying lists of frequent substructures. In our approach, an arbitrary selected structure in the list of N items is compared to the other N-1, resulting into N-1 connected MCS (Fig. 2). Since we are working with large sets, this already result in a large list of chemically relevant substructures, although the list might not be exhaustive. Additionally, a size filter keeps only the MCS covering at least 30% of the heavy atoms in both structures of a pair. Then, duplicate MCSs are removed from the list and sorted according to their occurrence in the list. The most frequent MCS is selected. Structures featuring the selected MCS are removed from the list, and a new iteration is started. In contrast with the previous scenarios, the new strategy returns a list of MCSs which is more relevant in the context of BigData.

The entire workflow is implemented in Python3 language using NumPy [32, 33] and Plotly [34] libraries. When the MCSs absent in the BI pool were found, the structures containing these MCSs were retrieved from the AMS collection,

Fig. 2 MCS extraction protocol


Initial data set
Data set splitting
New iteration if
N(rest comp-s)>1
Remove the query and structures featuring the most frequent MCS
First compound as a query
The rest (N-1) compounds
for each
Pairwise comparison
N-1 MCSs
IF
N_atoms(MCS_i) >= N_atoms(query)*0.3
AND
N_atoms(MCS_i) >= N_atoms(compound_i)*0.3
No
Yes
Select the most frequent MCS
The most frequent MCS
Result list of MCSs
Ignore the pair i

and their biological profile was predicted using previously developed universal GTMs [13].

Virtual profiling of novel compound candidates

The approach supported on the public property prediction server (http://infochim.u-strasbg.fr/webserv/VSEngine.html) utilizes consensus prediction of the activity class (active or not) of a compound with respect to 749 biological targets for which structure–activity records found in ChEMBL v.24 were considered to be sufficiently robust to provide for meaningful activity class landscapes on the seven distinct "universal" GTMs of drug-like space. Each candidate is iteratively projected onto each of the seven universal maps [15], and its projection is then placed in the context of the map-specific activity landscapes of each of the 749 targets. For each target, the compound is assigned a probability to belong to the "active" class, which corresponds to the relative excess of "active" population in its residence zone (or zero if the target-specific data from ChEMBL do not occupy at all this residence area). Herewith, a consensus probability $\bar{P}$ to be active on a target is taken as the mean of the seven predictions of the complementary universal maps. This mean is penalized by the standard deviation of the seven estimations (Eq. 1), to signal that mutual agreement of predictions enhances the trustworthiness of consensus.

$$P_{corrected} = \bar{P} - \sqrt{\frac{1}{6}\sum_{i=1}^{7}\left(P_i - \bar{P}\right)^2} \tag{1}$$

where $\bar{P}$ is the mean probability over the 7 universal maps; P_i is the probability to be active on a map i; $P_{corrected}$ is the corrected consensus probability.

The tool supports processing of up to a few million compounds, operating on the HPC cluster of the University of Strasbourg, in order to return a virtual profile matrix of input compounds × 749 predicted consensus probabilities.

Results and discussion

In order to train the GTM manifold, a Frame set (FS) of 25K compounds needed for the manifold construction was diversity-picked from the AMS library with the dissimilarity threshold equal to 0.4. At the next stage, the log likelihood threshold LLh = − 2501.52 was determined as described in Fig. 1 in order to delineate the GTM

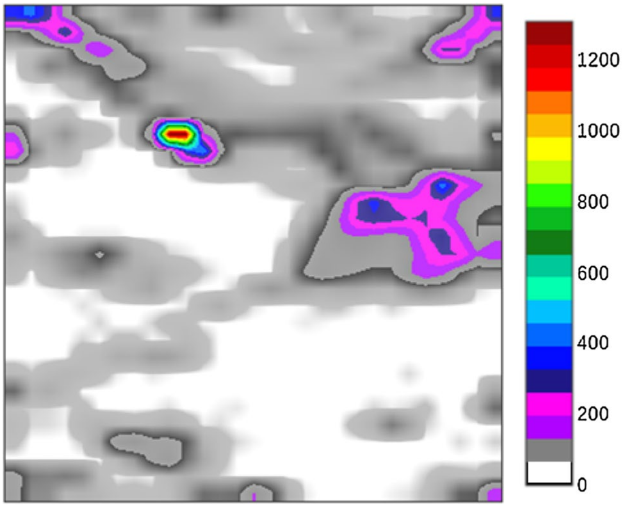


Fig. 3 Frame set density landscape. Here, the white space means non-populated areas. Both color intensity (transparency) and color choice are associated to local density values (red areas have no transparency)

applicability domain (AD). With this threshold, 95.5% of the FS items passed the AD criteria (23.9K compounds out of 25K). Figure 3 visualizes the distribution of the FS compounds over the map. The density landscape shows that the FS covers most parts of the map, and the maximal population of compounds in each node doesn't exceed 5% of the entire FS.

To understand how the two chemical collections relate to each other, they were projected on the map and rendered as individual density landscapes and a fuzzy classification landscape, respectively (Fig. 4). Some 94.1% of the BI Pool and 95.8% of the AMS collections passed the LLh threshold which means that the frame set extracted from AMS is diverse enough to describe both databases. In general, as far as the frame set is diverse enough to span the relevant chemical space zone, its explicit composition is of rather little importance—a recurrent conclusion in all our GTM studies, notably the creation of "universal" maps [24] where a frame set of the order of 10K random compounds was shown to suffice for the coverage of ChEMBL chemical space and

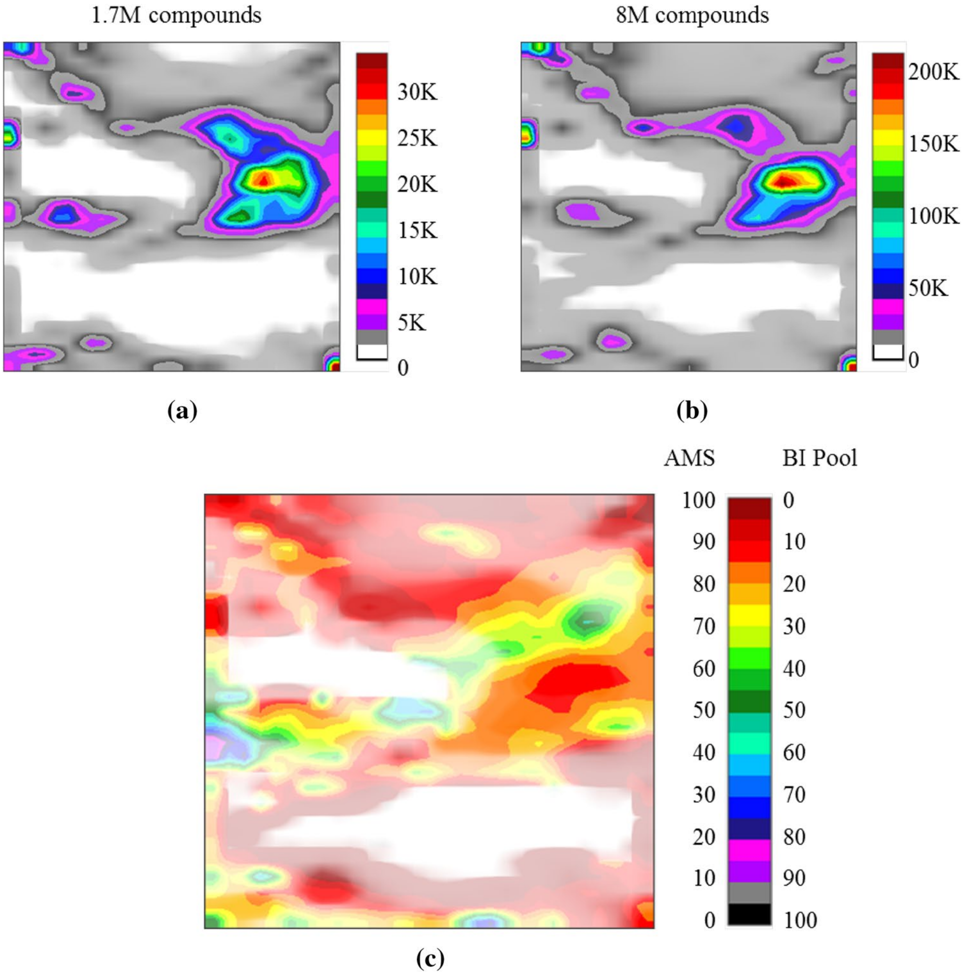


Fig. 4 BI Pool versus AMS comparison: **a** BI Pool density landscape, **b** AMS density landscape, and **c** fuzzy class landscape. Here, the white space means non-populated areas, and the transparency corresponds to the density

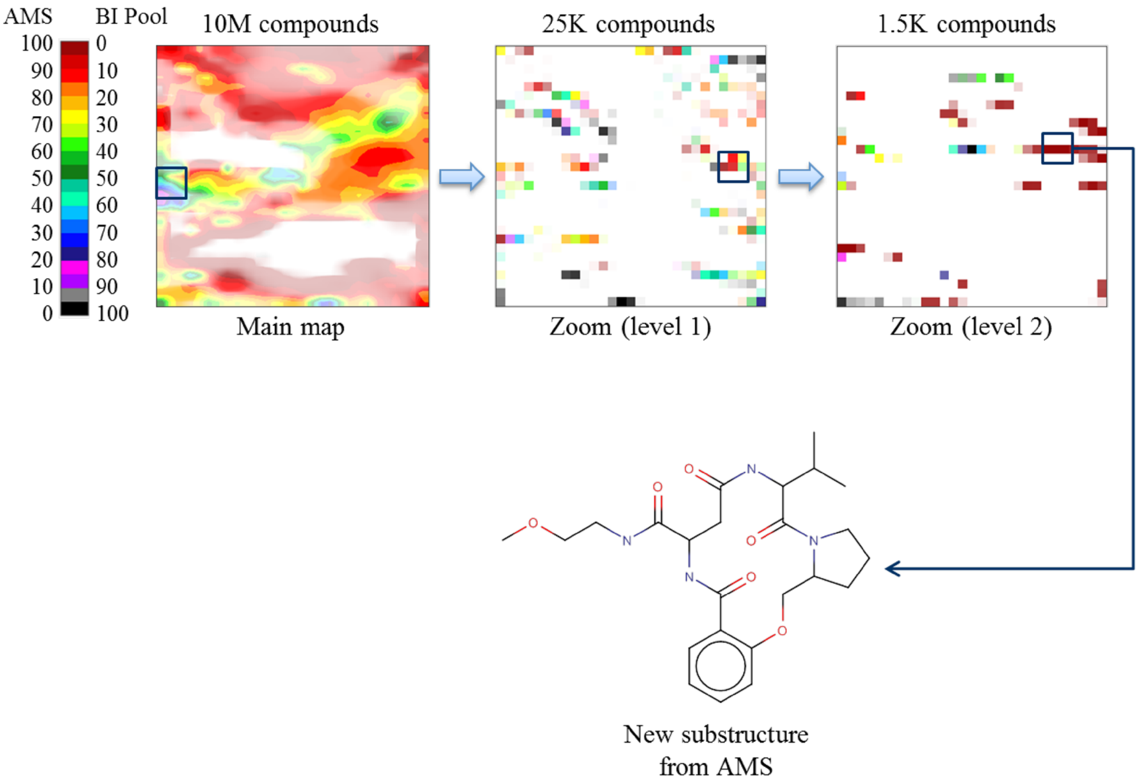


Fig. 5 An example of zooming analysis. Here, a new substructure from AMS collection was discovered using 2-levels zooming. The white space means non-populated areas, and the transparency corresponds to the density of population

supporting robust predictive activity models for hundreds of independent targets.

The density landscapes in Fig. 4a, b show that the libraries are globally similar since they both reside mostly the same areas. However, there are some areas where the AMS library has a strong presence and even fills some “holes” of the BI Pool. In the fuzzy class landscape, AMS-dominated areas are dark red (Fig. 4c). It is obvious that the dark-red areas can serve as a source of new chemotypes for the BI collection. However [12], even mixed zones might also contain some structural patterns not shared by both libraries. To investigate this possibility, 187 zones were checked whereby 151 zones were zoomed (the maximal level of zooming was up to 4). The procedure took approximately 7 days using 48 CPUs. An example of multi-level zooming is given in Fig. 5.

In total, more than 222K substructures were processed. This set included some 45.5K MCS present only in AMS collection. More than 401K structures containing these MCSs were extracted from the AMS collection and projected onto the map. The density landscape with some examples of the most popular new AMS substructures is given in Fig. 6.

Comparing the density landscape from Fig. 6 and the fuzzy class landscape from Fig. 4, we see that most of compounds came from the areas where AMS dominated. At the same time, several thousands of structures also came from the mixed areas (green and yellow). This was achieved by the application of zooming. In order to check the drug-likeness of the extracted structures, simple molecular properties, namely the number of H-bond donors and

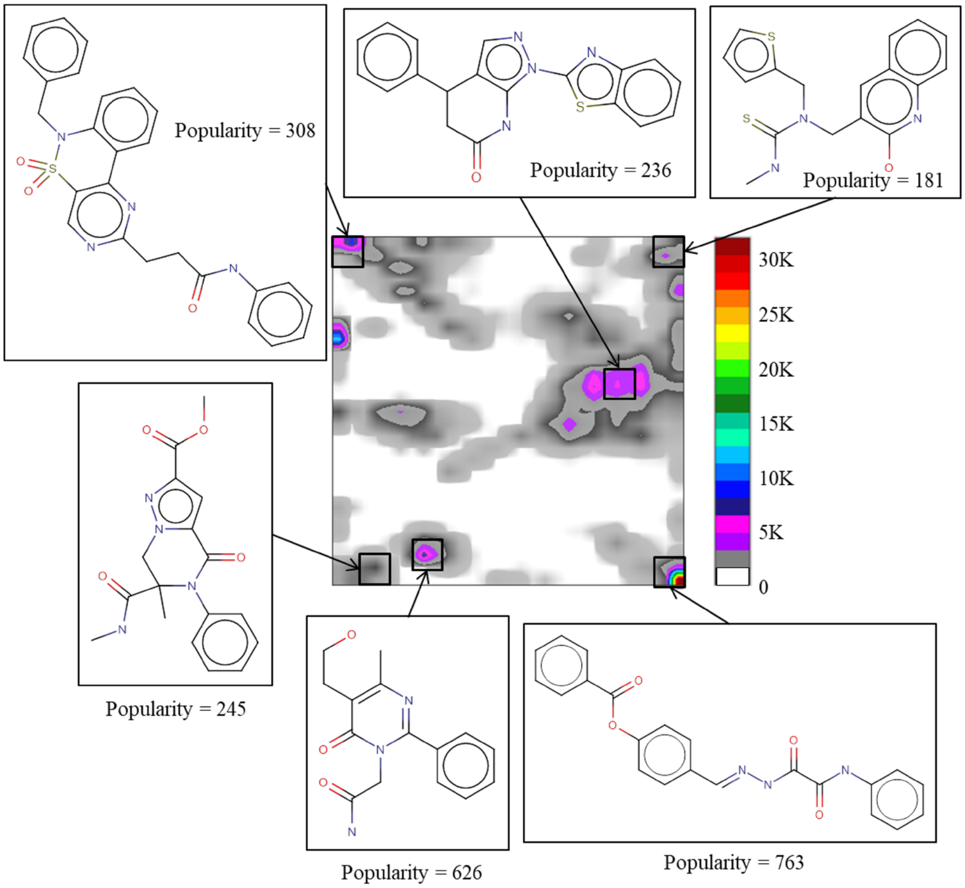


Fig. 6 Density landscape for the new 401K structures. Here, several most popular (within the particular zone) new substructures are shown. The number of corresponding compounds is presented here as a popularity score

acceptors, LogP, molecular weight, and TPSA were computed (Fig. 7).

Accordingly to Lipinski's rule of five [35], most of the extracted compounds can be classified as drug-like. These structures were also virtually profiled against 749 ChEMBL targets. 109.5K compounds were predicted as active against at least one out of 749 ChEMBL targets with a probability score $P_{corrected} > 0.5$. About 1.2K compounds out of it were predicted according to Eq. 1 as active with $P_{corrected} > 0.8$ and passed BRENK [36], PAINS [37] and NIH [38, 39] filters. The four examples with the highest corrected consensus probability to be active in one of the CHEMBL targets are shown in Fig. 8, where the compounds are predicted as active against Photoreceptor-specific nuclear receptor (CHEMBL4374), Cholecystokinin B receptor (CHEMBL3508), Muscarinic acetylcholine receptor M4 (CHEMBL317), and Pyruvate dehydrogenase kinase isoform 1 (CHEMBL4766) [40].

The type of the source of the structures (a chemical online store) allows us to say that these compounds are potentially synthesizable or even purchasable (the real synthesizability depends on a supplier, since some suppliers just claim that it can be synthesized if a client asks). This and the number of predicted actives demonstrate that the revealed substructures are new and useful for the pharma company. Also, it supports the statement that GTM is a powerful method for the efficient library comparison and enrichment (in terms of structural diversity).

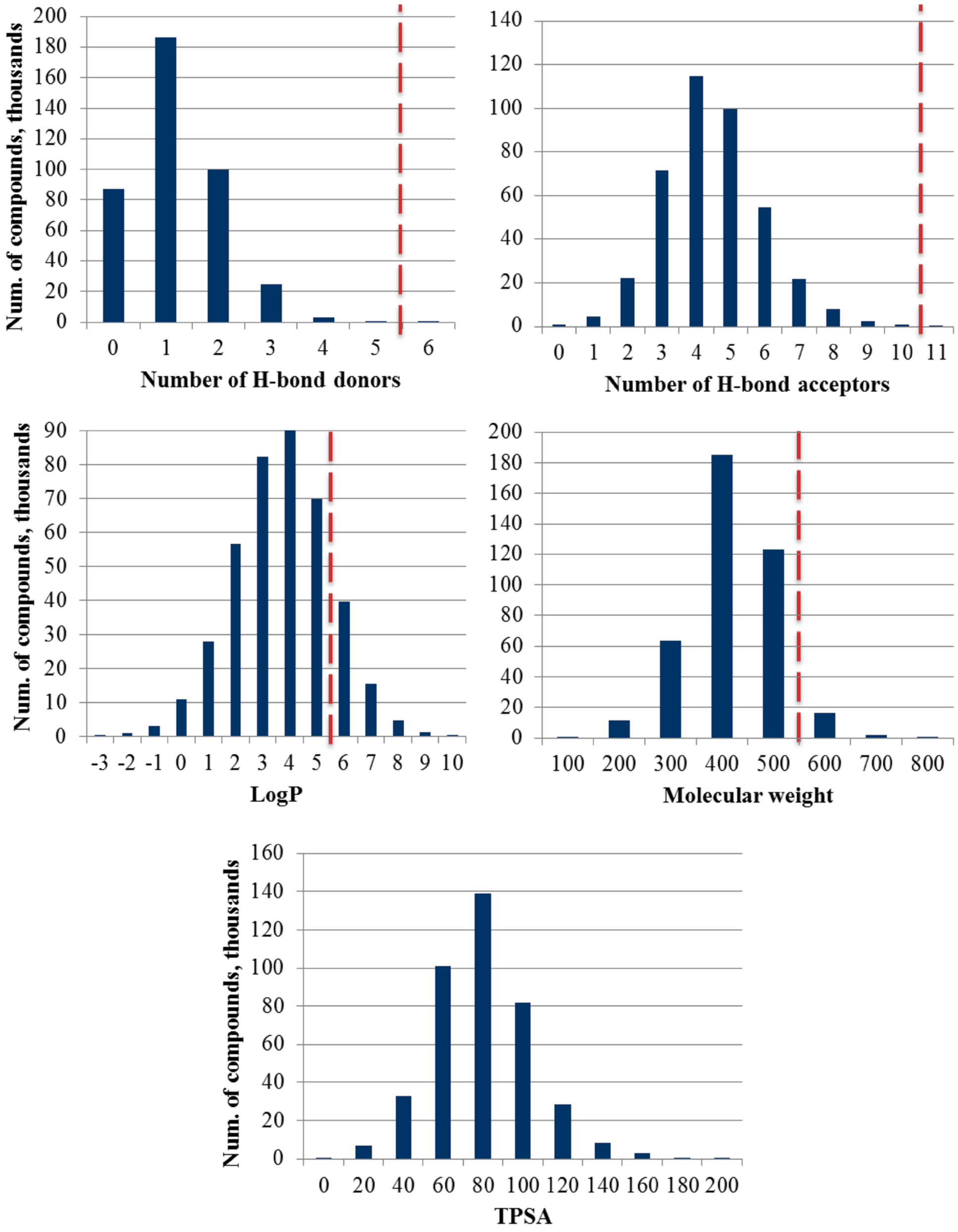
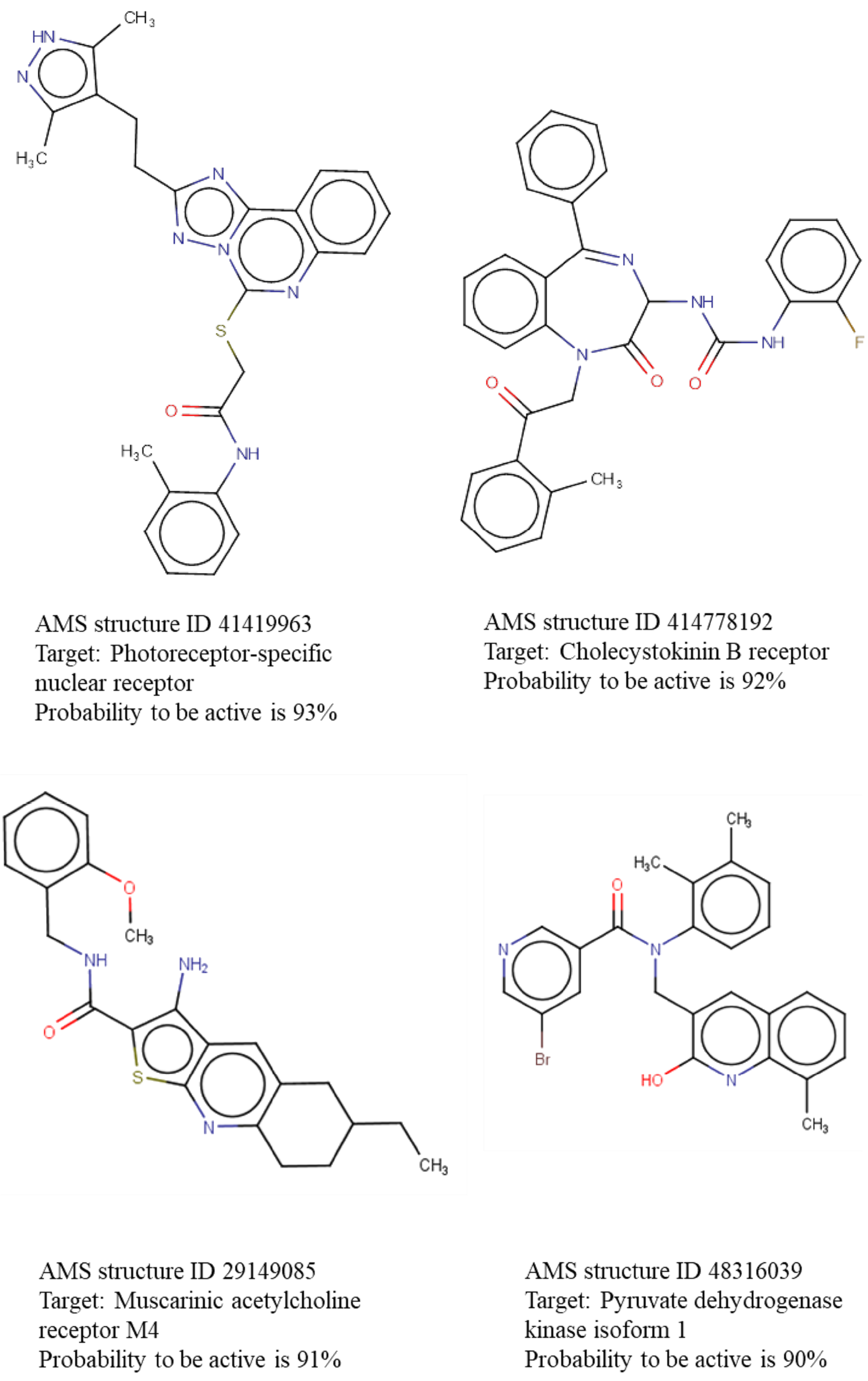


Fig. 7 Histograms represent the number of H-bond donors and acceptors, LogP, molecular weight, and topological polar surface area (TPSA) computed for the extracted 401K AMS compounds. Here, the red dashed line represents Lipinski's thresholds [35]

Conclusion

Generative topographic mapping was enabled to provide automated hierarchical analysis of large libraries, by means of the herein described "AutoZoom" tool. This integrates automated zooming and a new MCS extraction protocol and was successfully applied to diversify the in-house collection of Boehringer Ingelheim (BI). Some 45.5K substructures were found to be absent in the BI collection. The corresponding structures (401K items) were checked for Lipinski's rule compliance and classified as drug-like. In addition, they were virtually profiled against 749 ChEMBL targets. More than 1.2K compounds were predicted active against different targets with a corrected consensus probability (removing a standard deviation) higher than 80%. The discovered structures were recommended to the company to be imported as novel chemical matter that would be useful in diversifying the in-house collection.

Fig. 8 Examples of structures predicted as actives taken from the 1.2K AMS compounds. Here, the probability to be active returned by the web server is computed according to Eq. 1



AMS structure ID 41419963
Target: Photoreceptor-specific nuclear receptor
Probability to be active is 93%

AMS structure ID 414778192
Target: Cholecystokinin B receptor
Probability to be active is 92%

AMS structure ID 29149085
Target: Muscarinic acetylcholine receptor M4
Probability to be active is 91%

AMS structure ID 48316039
Target: Pyruvate dehydrogenase kinase isoform 1
Probability to be active is 90%

Acknowledgements The authors thank Boehringer Ingelheim Pharma GmbH & Co KG for the provided data.

Author contributions The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding The project leading to this article has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu).

References

1. Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A et al (2018) Reinforced adversarial neural computer for de novo molecular design. J Chem Inf Model 58:1194–1204. https://doi.org/10.1021/acs.jcim.7b00690
2. Kang S, Cho K (2019) Conditional molecular design with deep generative models. J Chem Inf Model 59:43–52. https://doi.org/10.1021/acs.jcim.8b00263
3. Schneider P, Schneider G (2016) De novo design at the edge of chaos: miniperspective. J Med Chem 59:4077–4086
4. Sattarov B, Baskin II, Horvath D et al (2019) De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. J Chem Inf Model 59:1182–1196. https://doi.org/10.1021/acs.jcim.8b00751
5. Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 52:2864–2875. https://doi.org/10.1021/ci300415d
6. Chang J-W, Jin D-S (2003) A new cell-based clustering method for large, high-dimensional data in data mining applications. In: Proceedings of the 2002 ACM symposium on Applied computing. ACM, p 503
7. Medina-Franco JL, Maggiora GM, Giulianotti MA et al (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Des 70:393–412. https://doi.org/10.1111/j.1747-0285.2007.00579.x
8. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330
9. Bernard P, Golbraikh A, Kireev D et al (1998) Comparison of chemical databases: analysis of molecular diversity with self organising maps (SOM). Analusis 26:333–341. https://doi.org/10.1051/analusis:1998182
10. Kireeva N, Baskin II, Gaspar HA et al (2012) Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. Mol Inform 31:301–312. https://doi.org/10.1002/minf.201100163
11. Gaspar HA, Baskin II, Marcou G et al (2015) GTM-based QSAR models and their applicability domains. Mol Inform 34:348–356. https://doi.org/10.1002/minf.201400153
12. Lin A, Horvath D, Afonina V et al (2018) Mapping of the available chemical space versus the chemical universe of lead-like compounds. ChemMedChem 13:540–554. https://doi.org/10.1002/cmdc.201700561
13. Tino P, Nabney I (2002) Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. IEEE Trans Pattern Anal Mach Intell 24:639–656. https://doi.org/10.1109/34.1000238
14. Lin A, Horvath D, Marcou G et al (2019) Multi-task generative topographic mapping in virtual screening. J Comput Aided Mol Des 33:331–343. https://doi.org/10.1007/s10822-019-00188-x
15. Casciuc I, Zabolotna Y, Horvath D et al (2019) Virtual screening with generative topographic maps: how many maps are required? J Chem Inf Model 59:564–572. https://doi.org/10.1021/acs.jcim.8b00650
16. ChemAxon Standardizer. https://docs.chemaxon.com/display/docs/Standardizer. Accessed 1 Feb 2019
17. ChemAxon JChem. https://chemaxon.com/products/jchem-engines. Accessed 1 Feb 2019
18. Bishop CM, Svensén M, Williams CKI (1998) GTM: the generative topographic mapping. Neural Comput 10:215–234. https://doi.org/10.1162/089976698300017953
19. Sidorov P, Viira B, Davioud-Charvet E et al (2017) QSAR modeling and chemical space analysis of antimalarial compounds. J Comput Aided Mol Des 31:441–451. https://doi.org/10.1007/s10822-017-0019-4
20. Monev V (2004) Introduction to similarity searching in chemistry *. Match-Commun Math Comput Chem 51:7–38
21. (2019) RDKit: Open-source cheminformatics. http://www.rdkit.org. Accessed 1 Feb 2019
22. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
23. Gaspar HA, Baskin II, Marcou G et al (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. J Chem Inf Model 55:84–94. https://doi.org/10.1021/ci500575y
24. Sidorov P, Gaspar H, Marcou G et al (2015) Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. J Comput Aided Mol Des 29:1087–1108. https://doi.org/10.1007/s10822-015-9882-z
25. Volochnyuk DM, Ryabukhin SV, Moroz YS et al (2019) Evolution of commercially available compounds for HTS. Drug Discov Today 24:390–402. https://doi.org/10.1016/j.drudis.2018.10.016
26. Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ et al (1988) Structure and energetics of ligand binding to proteins: escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. Proteins Struct Funct Bioinform 4:31–47. https://doi.org/10.1002/prot.340040106
27. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA property-labelled fragment descriptors. Mol Inform 29:855–868. https://doi.org/10.1002/minf.201000099
28. Marcou G, Solov'ev VP, Horvath D, Varnek A (2017) ISIDA fragmentor—user manual
29. Horvath D, Brown J, Marcou G, Varnek A (2014) An evolutionary optimizer of libsvm models. Challenges 5:450–472
30. Klimenko K, Marcou G, Horvath D, Varnek A (2016) Chemical space mapping and structure-activity analysis of the ChEMBL antiviral compound set. J Chem Inf Model 56:1438–1454. https://doi.org/10.1021/acs.jcim.6b00192
31. Hariharan R, Janakiraman A, Nilakantan R et al (2011) Multi-MCS: a fast algorithm for the maximum common substructure problem on multiple molecules. J Chem Inf Model 51:788–806. https://doi.org/10.1021/ci100297y
32. Oliphant TE (2006) A guide to NumPy. Tregol Publishing, USA
33. Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9:10–20. https://doi.org/10.1109/MCSE.2007.58
34. Inc. PT (2015) Collaborative data science. In: Plotly Technol. Inc. https://plot.ly. Accessed 1 Feb 2019
35. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 64:4–17. https://doi.org/10.1016/j.addr.2012.09.019
36. Brenk R, Schipani A, James D et al (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem Chem Enabling Drug Discov 3:435–444
37. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740

38. Doveston RG, Tosatti P, Dow M et al (2015) A unified lead-oriented synthesis of over fifty molecular scaffolds. Org Biomol Chem 13:859–865
39. Jadhav A, Ferreira RS, Klumpp C et al (2009) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. J Med Chem 53:37–51
40. Gaulton A, Hersey A, Nowotka ML et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954. https://doi.org/10.1093/nar/gkw1074