CHEMOGRAPHY: SEARCHING FOR HIDDEN TREASURES

Yuliana Zabolotna, Arkadii Lin, Dragos Horvath, Gilles Marcou, Dmitriy M Volochnyuk, Alexandre Varnek

CHEMOGRAPHY: SEARCHING FOR HIDDEN TREASURES

*Yuliana Zabolotna[1], Arkadii Lin[1], Dragos Horvath[1], Gilles Marcou[1], Dmitriy M. Volochnyuk[2,3], Alexandre Varnek[1]**

[1] University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France)

[2] Institute of Organic Chemistry National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine

[3] Enamine Ltd. (www.enamine.net), Chervonotkatska Street 78, Kyiv 02094, Ukraine

ABSTRACT

The days when medicinal chemistry was limited to a few series of compounds of therapeutic interest are long gone. Nowadays, no human may succeed to acquire a complete overview of more than a billion existing or feasible compounds within which the potential "blockbuster drugs" are well hidden, and yet only a few mouse clicks away. To reach these «hidden treasures», we adapted Generative Topographic Mapping to enable efficient navigation through the chemical space, from a global overview to structural pattern detection, covering, for the first time, the complete ZINC library of purchasable compounds, relative to 1.6 million biologically

relevant ChEMBL molecules. About 40 000 hierarchical maps of the chemical space were constructed. Structural motifs inherent to only one library were identified. Roughly 20 000 off-market ChEMBL compound families represent incentives to enrich commercial catalogs. Alternatively, 125 000 ZINC-specific compound classes, absent in structure-activity bases are novel paths to explore in medicinal chemistry. The complete list of these chemotypes can be downloaded using the link https://forms.gle/B6bUJj82t9EfmttV6.

INTRODUCTION

Nowadays, the number of molecules available to medicinal chemists is huge. The ZINC database merges commercial catalogs proposed by numerous chemical suppliers and contains more than 1.4 billion compounds[1]. It includes both already synthesized or in-stock compounds, and tangible molecules. Despite being just a tiny fraction of the estimated number of possible drug-like molecules (around 10^{33} structures)[2], the currently known chemical space is far from being fully studied and apprehended by medicinal chemists. For example, ChEMBL[3], containing biologically studied compounds extracted from the scientific literature is a thousand times smaller than ZINC. Thus, while chemical suppliers compete to enumerate the higher number of new virtual molecules[4], already existing compounds are largely unexplored from a drug discovery perspective.

Within the two last decades, the usefulness of purchasable screening libraries playing the role of a source of potential drugs has been evaluated in numerous reports[5-12]. These studies typically rely on a statistical analysis of chemical collections in terms of four groups of characteristics: physicochemical properties (e.g. molecular weight, logP, polar surface area, etc.), molecular complexity, diversity and novelty (usually based on a simple scaffold analysis[13]). All these reports provide an important insight into the evolution of medicinal chemistry relevant properties

of commercially available compounds and their distribution across screening libraries of different chemical suppliers. Yet, the scope of the mentioned works does not cover the entire chemical market, but only up to 2% of the purchasable compounds (16M out of 800M unique ZINC molecules). Moreover, there is a lack of chemical analysis of commercially available libraries. Indeed, the direct references to molecular structures were limited to the typical scaffold population analysis - a convenient and yet biased way to comprehend structural diversity[14]. The same scaffold may be adorned with radically different pharmacophore patterns and, hence, have completely different biological effects. On the other hand, the same pharmacophore may be "incarnated" by radically different scaffolds and yet exhibit similar activity[15].

In the meantime, all those works aim to analyze only the current state of the chemical market without trying to identify and, if possible, to fill the gaps in the purchasable chemical space. One way to evaluate such possible incompleteness is a comparison of commercial catalogs with a reference subset of molecules possessing desired properties. Such an approach was previously adopted by Shelat and Guy in their study of the biological relevance of screening libraries[16]. They compared some purchasable chemical collections (≈2M unique structures) with a set of known drugs (≈8K compounds). The results have shown that there is only a 14% scaffold overlap between analyzed subsets which brings us to the conclusion that commercial chemical space at that time was not sufficiently covering biologically relevant compounds. The challenging goal of increasing that coverage can hardly be achieved by unguided compounds enumeration. It requires a deep understanding of the main features of both purchasable and biologically relevant chemical space.

In this context, our study focuses on two goals: (i) commercial chemical space enhancement and (ii) its exploration. The first one consists in the identification of biologically relevant

compounds that are absent from the current chemical market. Such molecules, being synthesized in academic laboratories, small start-ups, big pharmaceutical companies or coming from natural product-based programs[17], are also entering biological assays and results of these tests eventually become publicly available. Those biologically relevant compounds and especially their untested analogs, if added to the commercial catalogs, could be highly useful in further screening campaigns and SAR studies, and, thus, become good starting points for the development of new «best-sellers» of the chemical market. At the same time, not all commercially available compounds have been tested before in biological studies. The compound classes that have been overlooked by medicinal chemists can be used for expanding the scope of the biological exploration of the commercially available chemical space.

In order to find such "hidden treasures", we performed thorough chemical analysis of the drug discovery-oriented commercial chemical space, featuring (after standardization and duplicate removal) 800M ZINC compounds, versus 1.6M molecules that have already attracted the attention of medicinal chemists and were therefore captured in the ChEMBL database together with their observed biological activities. Both ZINC and ChEMBL compounds were split into four groups depending on the type of biological tests and selected drug design strategy, resulting in fragment-like[18], lead-like[19], drug-like[20] and PPI-like[21] subfamilies. The purchasability of ZINC molecules was also assessed: they were further split into ZINC-Real - in-stock compounds directly available for purchase, and ZINC-Tangible - compounds that can be synthesized upon request .

Thousands of chemotypes, specific only to ChEMBL or ZINC libraries, were detected for each of the mentioned subspaces. It was done using one of the most efficient chemography methods of dimensionality reduction - Generative Topographic Mapping (GTM)[22], that has already proven

to be a successful approach for visualization and versatile analysis of large chemical libraries.[23] It produces easily readable 2D maps of chemical space - a very convenient way for navigating through billions of compounds.

It was found that commercially available libraries are missing numerous compound families known to include biologically active members - highly potent inhibitors of important biological targets. Some examples of ChEMBL and ZINC specific chemotypes will be discussed in the text, while the full list of these structures – a potential source of inspiration for synthetic and medicinal chemists - can be downloaded using the link https://forms.gle/B6bUJj82t9EfmttV6. Notice that the identified in this work ZINC-specific MCSs that were absent in both ChEMBL and PubChem[24] (revealed by secondary substructure check), were then in Silico profiled against 749 ChEMBL targets. It was done with the help of GTM Profiler tool[25] used to evaluate their potential usefulness in drug design (http://infochim.u-strasbg.fr/webserv/VSEngine.html).

Chemography as a versatile tool for chemical space analysis

Both chemography, as an "art of navigating in chemical space[26], and activity/property prediction should be used for chemical space analysis. The first is needed to navigate through the complex structure of the chemical data, and the second might serve to set the landmarks (identify compounds potentially possessing desired properties, by predicting those properties, in absence of experimental data). Also, the chosen approach must be "Big Data"-compatible. Generative Topographic Mapping method, or GTM, conveniently fulfills all these requirements. Briefly speaking, it translates compounds from the initial multidimensional descriptor space to a 2D latent space, called a 2D map. In contrast to Self-Organizing Maps[27], GTM distributes molecule projection over the map with node-specific probabilities (responsibilities) instead of

unambiguously assigning each compound to only one point on the map. This smoothness enables creation of GTM landscapes – cumulated compound responsibility patterns, colored by average values of different properties, e. g. density, biological activity, assigned class, etc. (see examples in Figure 1 a). The details of the method are provided in Supporting Infomation.

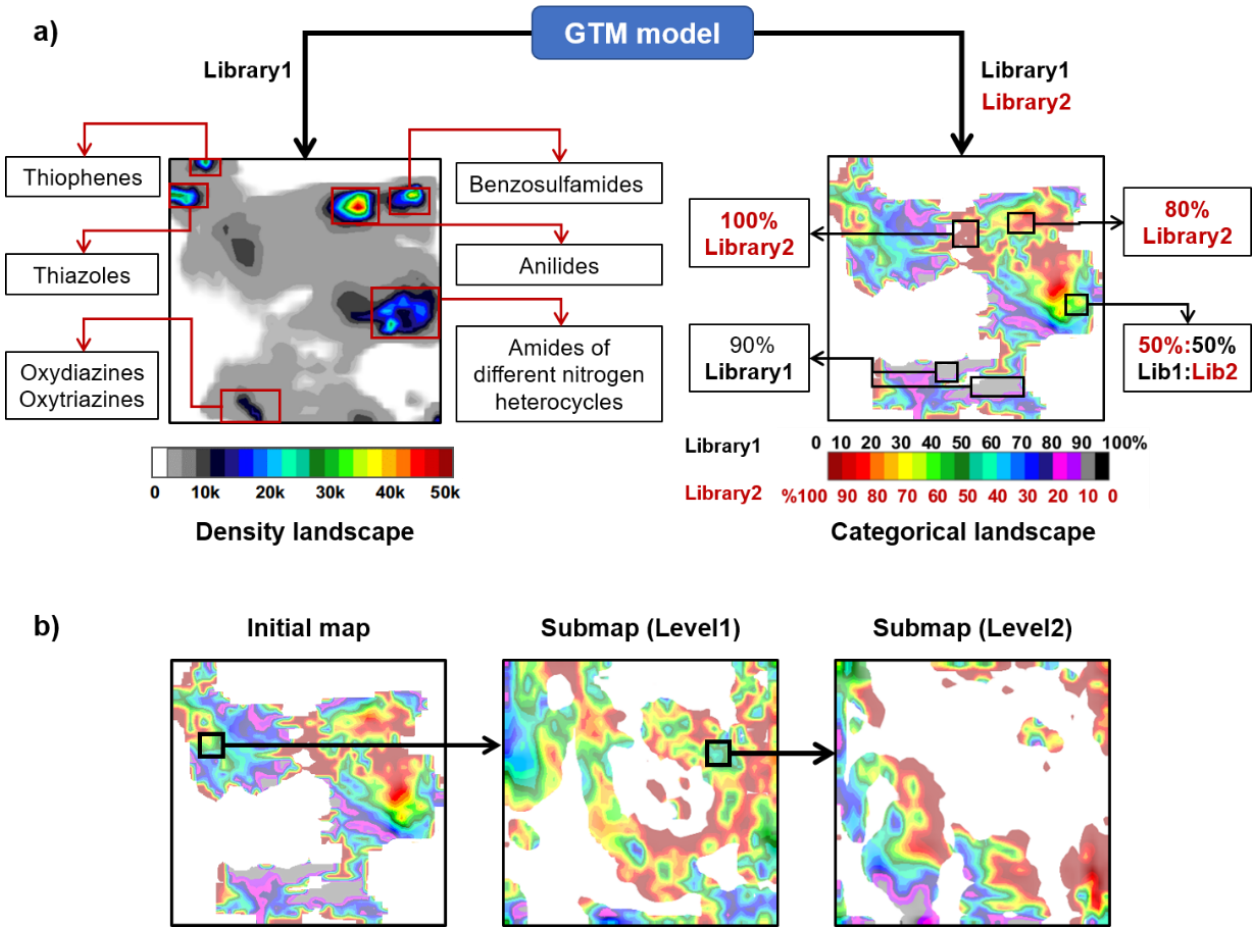


Figure 1. Generic scheme of library analysis and comparison with GTM: a) left - density landscape used to analyze the distribution of different compound classes across the chemical space (color spectrum matches the cumulated responsibility, corresponding to the number of resident compounds) versus right - a categorical landscape rendering chemical space regions occupied by two libraries (the color code matching the proportion of residents from each library); b) a schematic overview of the HGTM navigation through the highly populated areas of the

chemical space – compounds, extracted from the zone of interest, are used for constructing a new map, now focused only on this region of chemical space.

Walking over this map and performing an in-depth chemotype analysis of the residents of the local map zones is a rational and intuitive way to systematically "browse" the chemical space and get acquainted with the structural patterns it hosts. In this work those patterns were characterized by maximum common substructures (MCSs) – the largest structural fragments that aim to generalize common features of the group of molecules they represent[28]. These MCSs were defined as substructural fragments, that contain at least 30% of each molecule they represent. MCS was preferred over the widely used scaffold concept because it is open-ended and adaptive: it may coincide with the scaffold or be more specific by including key substituents (side-chains) if appropriate. The algorithm that combines both GTM and MCS detection was presented by Lin et al.[29] and is briefly discussed in Supporting Information.

Yet, 2D maps cannot accommodate a huge number of compounds all while capturing fine differences between close neighbors: a hierarchical zooming approach will be required to let the user capture details of the chemical population at any point of the global map, and reach down to "hidden treasures" buried beneath millions of compounds. Hierarchical GTM (HGTM)[29, 30], a.k.a "Zooming" is a technique that trains a new map on a set of compounds extracted from a given zone on the parent map, in order to ensure a locally optimal mapping (Figure 1b). The zoomed map is free to fit the local compound distribution, with no constraints to simultaneously match all the other compounds – which is the key benefit, beyond the obvious gain in resolution (the latter could have been easier achieved by imposing a finer grid mesh on the global map).

Last but not least, with a robust structure-activity set used to create an activity landscape (a landscape colored by activity values), the map can be turned into a potent QSAR/QSPR model[25,

[31-33]. Predictivity of those models can be quantitatively determined and serve as a guide in the search for "the best map" parameters configuration. In this way, our group built seven optimized "Universal" maps of the drug design-relevant chemical space, selected for their ability to host as many predictive activity landscapes, for different drug targets with enough structure-activity data reported in ChEMBL[25]. Those maps are the basis of the GTM Profiler - a virtual screening tool, that allows to predict compound activity against 749 biological targets. It is extremely time-effective for already mapped molecules. The previously reported "top" Universal map serves here as the principal tool for the biologically-biased analysis of the commercial compound space.

RESULTS AND DISCUSSION

Chemical analysis of the commercially available chemical space

Initially, 1.3 billion (out of total 1.5 billion) compounds from ZINC15, passing built-in "standard reactivity" filter and 1.8 million molecules from ChEMBL (version 25) were collected for this project. After structure standardization and stereoisomer "fusion" into a common, stereochemistry-depleted representation, 800 million ZINC and 1.6 million ChEMBL unique structures remained. Compounds with unwanted functionalities were filtered out (Table S1) and four subsets associated with different stages and strategies of drug discovery were defined (Table 1). Commercially available compounds were split according to their purchasability into ZINC-Real and ZINC-Tangible. The first group contains all compounds that have been already synthesized in a sufficient quantity and thus can be delivered in under 2 weeks to the buyer with a 95% acquisition success rate. The second one, in contrast, contains compounds that were designed by suppliers as a result of the stock enhancement programs and have not been synthesized yet. Thus 8-10 weeks are needed for their delivery and acquisition success rate is about 70%. [1] Tangible libraries are considered as the source for the chemical enhancement of the

Real ones. They can be readily made from existing building blocks according to the well-defined procedures[34], approved by the synthetic chemists. Therefore ZINC-Tangible compounds were used in this study rather than de-novo generated molecules[35, 36] of uncertain chemical feasibility. Further details about data preparation and filtering rules can be found in Supporting Information.

The present analysis employs Universal map #1 as the best one out of the previously built general-purpose chemical space maps[25]. It was constructed in a way to be able to predict 618 biological activities present in ChEMBL database. Being multitarget-oriented, this map can be considered as a generalized framework for biologiacally-biased chemical space visualization. It is based on one of the ISIDA fragment descriptors – atom sequences with a length from 2 to 3 atoms labeled by CVFF Force Field types and Formal Charges labels[37]. See more details about the construction of the Universal map #1 in Supporting Information.

Table 1. Size of the medicinal chemistry-relevant subsets after standardization & appropriate filtration.

	ChEMBL	ZINC-Real	ZINC-Tangible
Fragment-Like	15 398	103 530	2 772 851
Lead-Like	361 051	3 253 343	329 893 210
Drug-Like	668 222	5 158 676	516 492 788
PPI-Like	229 570	1 248 875	63 632 835

First, each of the abovementioned ZINC subsets were projected onto the universal map. Density landscapes of the subsets were built in order to obtain a general overview of the structural features of the purchasable chemical space (Figure 2). Interestingly, the commercial compounds are distributed in a highly imbalanced manner: the major part of the map area is

rather sparsely populated (gray zones), by contrast to a few outstanding density peaks (multicolored regions). In Figure 3, the structural analysis of the densest regions of the lead-like ZINC-Real part of the chemical space is provided: characteristic MCS of some zones are shown. The density imbalance goes in correspondence with the previously reported unequal compound distribution across different compound classes[11, 12]. An overrepresentation of synthetically accessible benzensulfonamides, anilids and other amides is noticed (Figure 3: regions R3, R4, and R5). These chemical subfamilies echo, firstly, the extreme popularity of combinatorial chemistry methods in the 20th century. Based on limited sets of building blocks and simple reactions, they allowed to synthesize large numbers of compounds at the cost of limited chemical diversity. At the same time, the complexity of the synthetic path for some compounds prevented the mass production of their analogs.

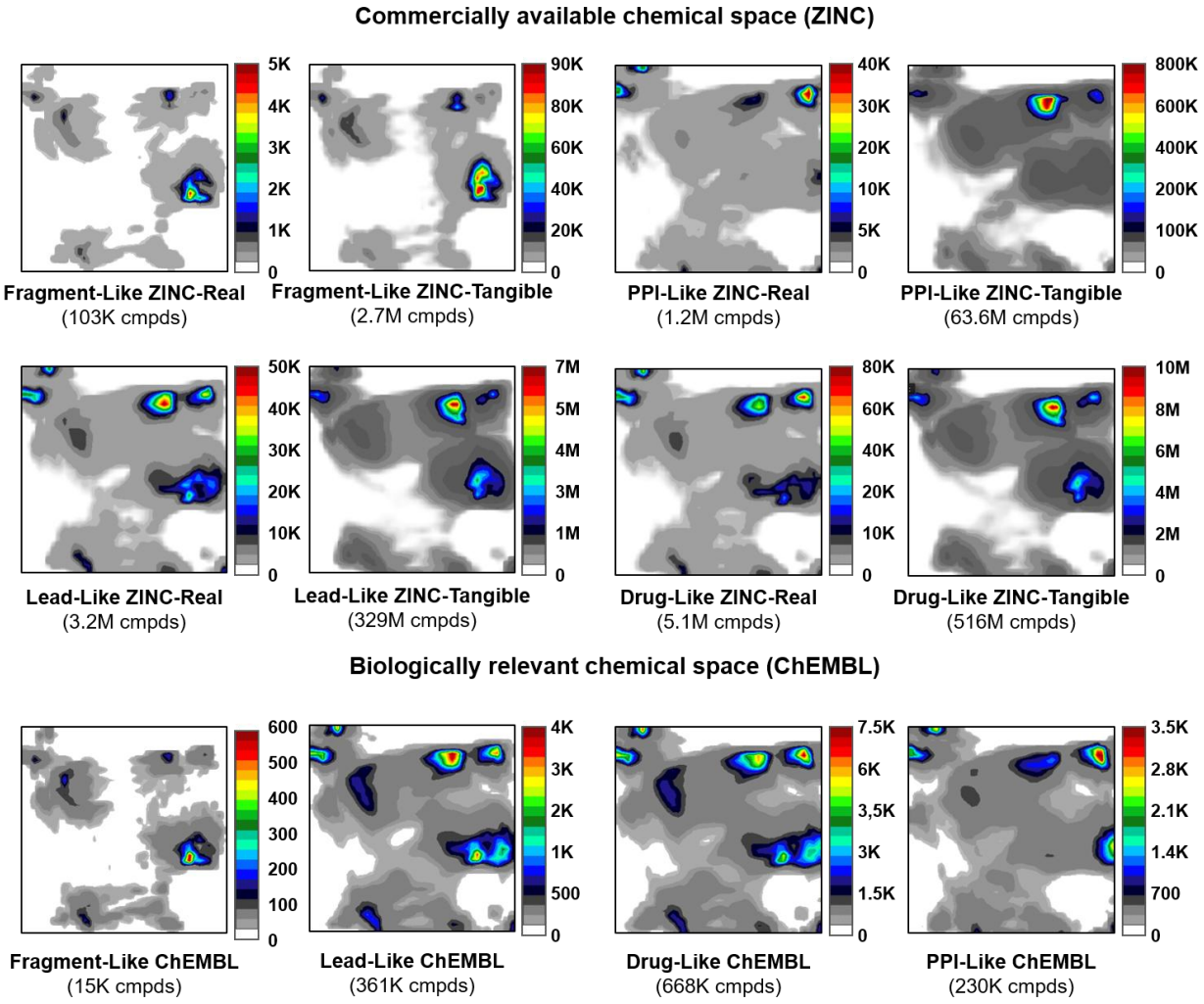


Figure 2. Density landscapes of commercially available (ZINC) and biologically relevant (ChEMBL) subsets. The color scale renders the corresponding number of compounds residing in each colored node of the map.

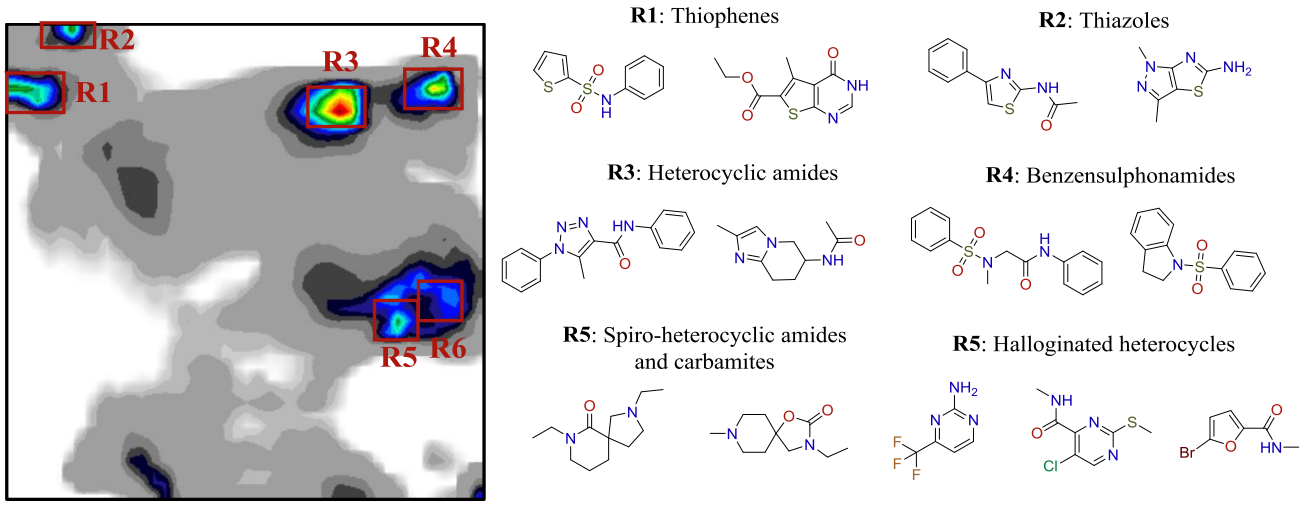


Figure 3. Examples of the most frequent structural motifs from the densest regions of the Lead-Like ZINC-Real map.

The second reason is medicinal chemistry demand, which has also reshaped purchasable libraries significantly. For example, sulfonamides - the main inhabitants of the R4 region, are known for their anti-bacterial properties for almost 100 years. Back in time, together with antibiotics, they revolutionized the medicinal approach for various infections treatment, moving it from immuno- to chemotherapy[38]. Another examples are thiophene-containing compounds (region R1), that possess diverse therapeutic properties - antimicrobial, anticancer, anti-inflammatory activity etc.[39] In addition, the thiophene cycle is highly popular in medicinal chemistry due to its bioisosteric correspondence with phenyl.

The previous century's synthetic methods and medicinal chemistry demands are still influencing the current chemical market[40]. This historical bias can be a dangerous limitation for discovering new valuable patterns in medicinal chemistry - novel chemotypes with a specific activity. Since tangible ZINC libraries have been designed rather recently, in theory, their compound distribution should be more balanced than those of in-stock collections. In practice,

all the analyzed subsets of ZINC-Real and ZINC-Tangible are very similar. The same shape of the occupied areas on the map as well as the position of the high-density regions can be observed (Figure 2). Although tangible libraries increase the total number of compounds on the market, they still tend to sample the same areas of the chemical space that are already overpopulated by in-stock libraries. That means that current strategies of the commercial library enhancement do not provide a uniform chemical space sampling and thus there is an urgent need for their improvement.

In search of the «hidden treasures»

Commercial chemical space is huge and thus expected to include novel chemotypes that were never subjected to biological testing so far. Moving them from the chemical store onto a shelve of the medicinal chemistry lab might open new opportunities in drug discovery. At the same time, suppliers might miss some important types of compounds - highly potent drug-design candidates, that were developed and tested in small companies or academic laboratories. These compounds are of high interest for medicinal chemists, and their presence in the commercial catalogs will certainly enrich the latter.

In search of these “hidden treasures”, a detailed comparison of ZINC and ChEMBL libraries was performed. From “bird’s-eye” perspective, the ChEMBL and ZINC chemical spaces coincide fairly well: in Figure 4, for each of the landscapes, there are only a few small zones in which the extremes of the color spectrum (local population exclusively stemming from one of the libraries) can be observed. However, this resolution level is certainly not sufficient, as one single node of the map may contain up to several millions of compounds (Figure 2), forcing dissimilar compounds to share common zones. The HGTM approach has been used to further navigate through highly populated areas. Up to five zooming levels were used to build about

40 000 “child” maps (Figure 4). All zones, containing in total more than 1 000 compounds were zoomed, others – subjected directly to the MCS detection protocol[29]. For example, in the landscape hosting 3.6M lead-like [ChEMBL+ZINC-Real] compounds (Figure 5), Zone1 is equally frequented by both libraries and contains more than 82 000 compounds. Two zooming iterations of this zone reveals a detailed landscape where areas with unique substructures (and, hence, chemotypes) can be found for each library (Zone3 and Zone4).

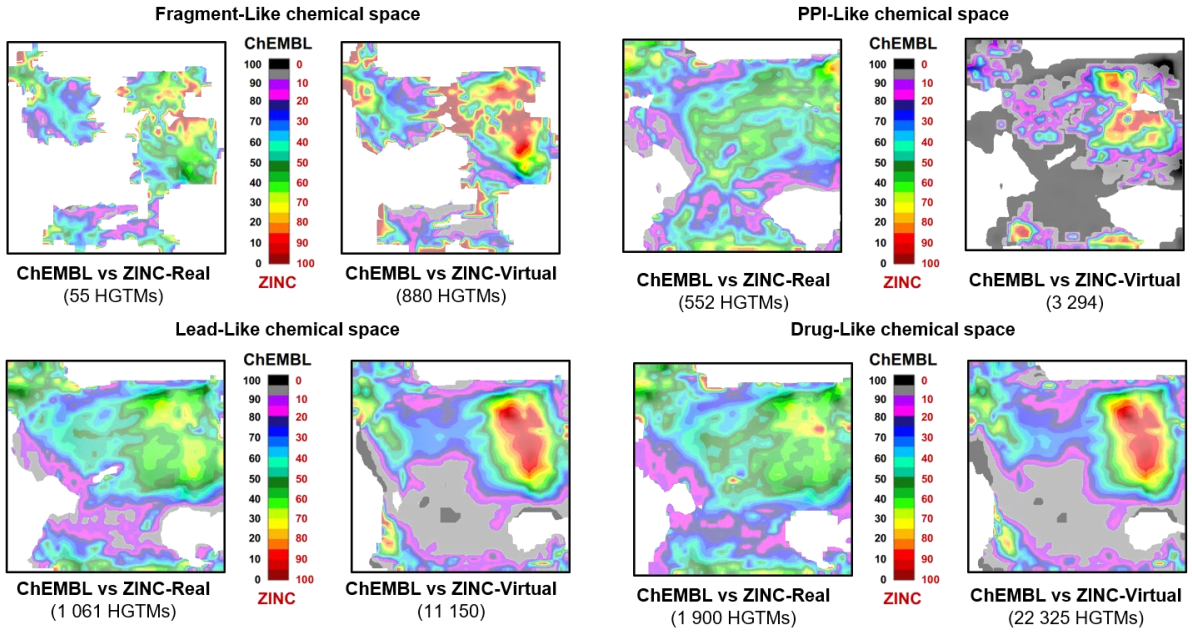
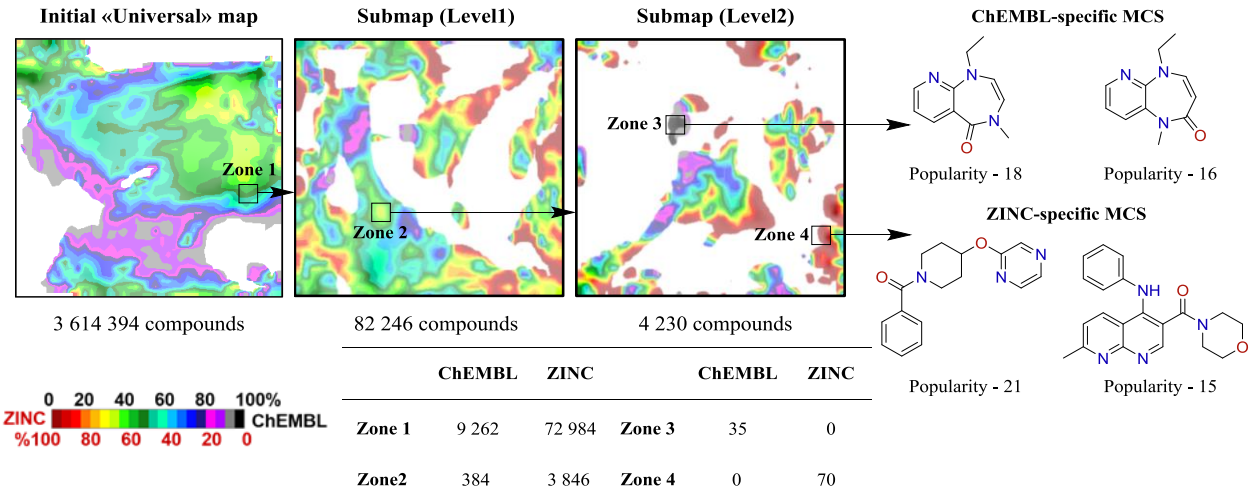


Figure 4. Categorical landscapes of the medicinal chemistry relevant subsets of commercially available chemical space. Each map visualizes compounds both from ChEMBL (zones colored in black) and ZINC (colored in red). White regions correspond to the empty areas. All colors in between correspond to the various normalized proportion of compounds from different subsets, projected into a particular node of the map (see Supporting Information). Numbers of parenthesis shows how many subsidiaries or «zoomed» GTMs were built.



	ChEMBL	ZINC		ChEMBL	ZINC
Zone 1	9 262	72 984	Zone 3	35	0
Zone2	384	3 846	Zone 4	0	70

Figure 5. HGTM navigation of the highly populated areas of the chemical space: Lead-Like ChEMBL vs ZINC-Real example. The Table provides the composition of each highlighted zone. Starting from the dense mixed Zone1, through the two levels of zoom, small purely ChEMBL (Zone3) and ZINC(Zone4) subareas are detected. Corresponding MCS and their popularity (number of compounds that contain each structural fragment) are also reported.

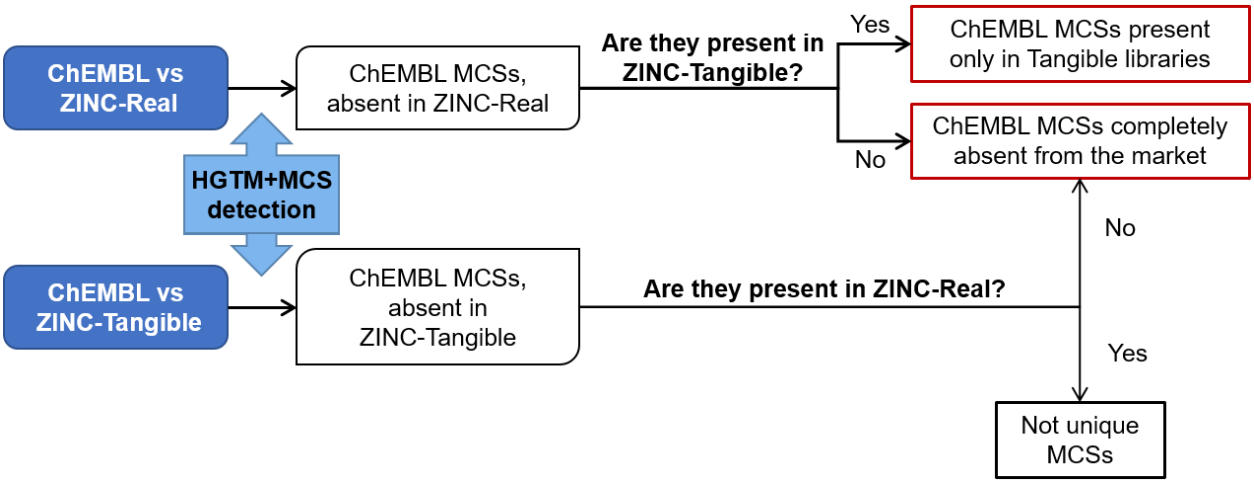


Figure 6. Schematic workflow: searching for ChEMBL-specific MCS with no commercial coverage.

First, we focused on MCSs present in ChEMBL but not in ZINC. The workflow of their search is depicted in Figure 6. ChEMBL subsets (fragment-like, lead-like, drug-like and PPI-like) were

compared pairwise to ZINC-Real and ZINC-Tangible. The ChEMBL-specific MCSs, locally discovered as a result of such comparison, were used as queries in a substructure search against the corresponding ZINC-Real and ZINC-Tangible subsets. The absence of substructure hits means that these MCSs are not only zone-specific but unique to the respective subspace of biologically tested compounds, and absent from the supplier libraries. Several examples of the potent nanomolar inhibitors containing some of the specific substructures for each of the analyzed subsets are shown in Figure 7. For more examples of ChEMBL-specific MCSs, see Table S2.

Most of the new ChEMBL substructures are much more complex than simple Bemis-Murcko scaffolds. For some substructures, it is the side chains that make them unique - the corresponding scaffolds with different decorations can be present on the market. This is the key advantage of our MCS-based search for characteristic substructures over a rigid scaffold-based approach. Figure 7 includes compounds active against therapeutically important targets. Those compounds and especially their analogues can be useful not only in the context of their known activities, but also (and more so) in other drug design campaigns featuring other biological targets.

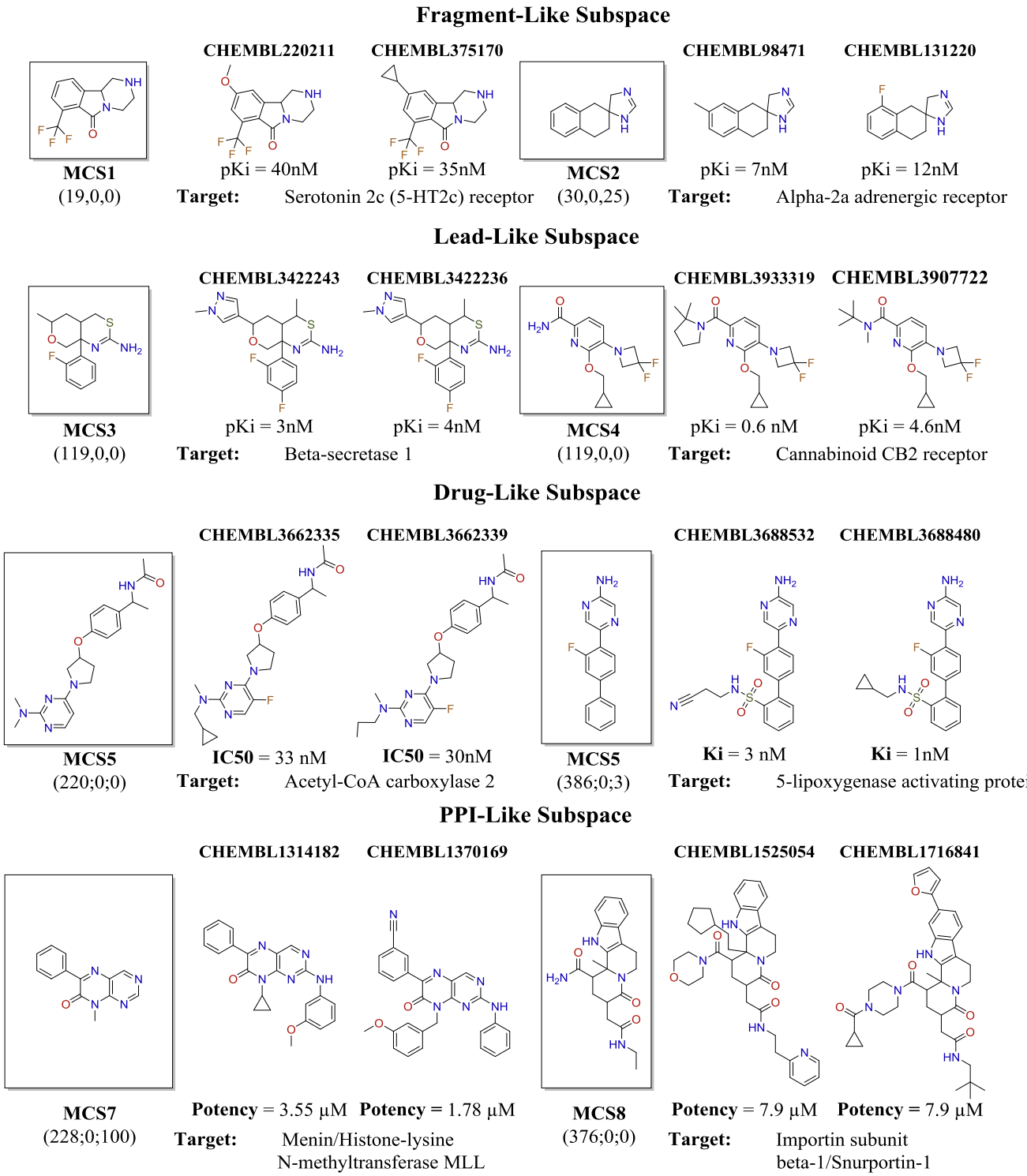


Figure 7. Examples of the highly potent inhibitors, incarnating one of the reported unique ChEMBL substructures, recommended for the chemical space enhancement. Numbers in

parenthesis under each MCS identify the number of corresponding compounds containing this MCS in ChEMBL, ZINC-Real, and ZINC-Tangible libraries respectively. All reported targets are Homo Sapiens proteins with high therapeutic importance.

The absence from the commercially available chemical space of so many potentially very important compound families, known to include biologically (very) active members, is somehow intriguing – after all, those molecules were produced and tested, but somehow left no trace of precursors or analogs in commercial space. Several plausible explanations may exist - the "unique" MCS may emerge during the reaction, thus not be present in commercial building blocks, the compound was produced from proprietary building blocks, etc. Some of the ChEMBL-specific chemotypes can be missing from the vendors' libraries because they are part of the intellectual property space, which covers compounds protected by the patents. Unfortunately, the analysis of the intellectual property chemical space is not straightforward. A majority of patented structures are represented in a form of Markush structures, making these libraries impossible to cartograph (as prerequisite individual enumeration and molecular descriptor calculation for the combinatorically enumerated structures covered by a Markush formula may be too costly, or outright unfeasible). Furthermore, not all of mechanically enumerable Markush substituent combinations stand for chemically stable compounds – and even less represent confirmed actives. Specific tools for Markush-targeted substructure querying and even (connectivity-driven) similarity search tools exist, but more sophisticated approaches involving information-rich descriptors, such as topological pharmacophore patterns, cannot be applied. Users will be free to submit any species of interest highlighted by our tool to a state-of-the-art check against patent libraries, but in our opinion no closer integration can be envisaged –

the rigorist, connectivity-centric legal status of a compound is not easy to reconcile with its fuzzy-logics-based responsibility patterns.

It should also be noted that the presence of the particular chemotype in the patents libraries, yet does not mean that respective compounds cannot be synthesized or used in drug design campaigns. The point is that some patents protect only compound usage against a specific biological target or family of targets, leaving the freedom to operate outside of the specified research area. Such compounds can still be used in primary screening campaigns against novel biological targets.

The entire list of concerned MCS is freely available and, in our opinion, is an interesting source of enrichment of the purchasable in-stock libraries enhancement.

Biological exploration of the currently available chemical space

The complementary application of this work is the detection of biologically unexplored regions of chemical space, e.g. ZINC-specific MCS. The same approach highlighted two sets of ZINC-Real and ZINC-Tangible-specific substructures derived from compounds not found in ChEMBL. Table S3 shows a diverse set of examples.

One might argue that some of those compounds could have been not "overlooked" by medicinal chemists, but rather intentionally discarded from the screening campaigns. However, the herein employed standardization and filtering procedure should have eliminated most of the obviously reactive compounds or potential PAINS from the 800M filtered pool of ZINC compounds (albeit there is no absolute consensus of what precisely "unwanted" structures are). Thus, in order to dispel remaining doubts, additional analysis of the key substructures as a potential source of the highly potent hits was performed.

The ultimate pertinence of herein highlighted ZINC-specific MCSs for biological exploration of the chemical space will only be completely validated by actual experimental screening of those compounds, by MedChem groups pursuing specific drug discovery projects. This path is beyond the present work, which limits itself to present some indirect hints of the usefulness of these compounds, notably by (i) investigating whether those types of compounds have been tested already, without being reported yet in ChEMBL database or (ii) predicting biological properties of the compounds of interest using the same universal map-based property landscapes – a fast, robust and intuitive approach directly emerging from the chemographic concept.

Not being present in ChEMBL is not yet synonymous with being "off the beaten path". ChEMBL focuses mostly on the higher-level (dose-response) biological data, but some of the ZINC-specific MCS might have served in HTS campaigns reported elsewhere. PubChem, the largest collection of structure-activity data including high-throughput screening (HTS) reports, has been chosen in this study as an alternative external subset. 101M compounds, 80M of which are unique structures (stereoisomers were considered duplicates) were collected after analysis of ChEMBL- and ZINC-specific maximum common substructures was finished (December 2019). 3.1M of those compounds have been tested in at least one biological assay, while only 1.1M compounds were labeled as "active" by PubChem.

In a search for the potential drug candidates out of ZINC-specific subspace, around 24K of lead-like ZINC-Real unique MCS (absent not only in lead-like subset but also in the unfiltered version of ChEMBL) were used as queries against 3.1M biologically tested PubChem compounds, but only molecules marked as «active» were reported as hits. The lead-like real subset was selected as the most relevant with respect to the HTS demands and instant purchasability of corresponding compounds.

As a result, 9 575 ZINC MCSs were found in PubChem. For 1 628 of those MCSs, there were 4 520 PubChem compounds labeled as actives in 1 772 different biological assays. Among them one of the recent studies of natriuretic polypeptide receptor (hNpr1) antagonism[41]. It was published in July 2019 and therefore could have not been included in used here ChEMBL version 25, that was released in March 2019. Using HTS, the authors identified potent hNPR1 inhibitors. One of these compounds (JS-11) was further tested *in vivo* in mouse, causing decrease of the behavioral response. Interestingly, this molecule contains one of the ZINC-specific substructures identified earlier - MCS12. Figure 8 shows examples of MCSs, that were found in the «active» PubChem subset, including MCS12 and corresponding compound JS-11. These examples prove that previously unexplored regions of chemical space may contain "hidden treasures" – potential drug candidates or at least starting points for their design.

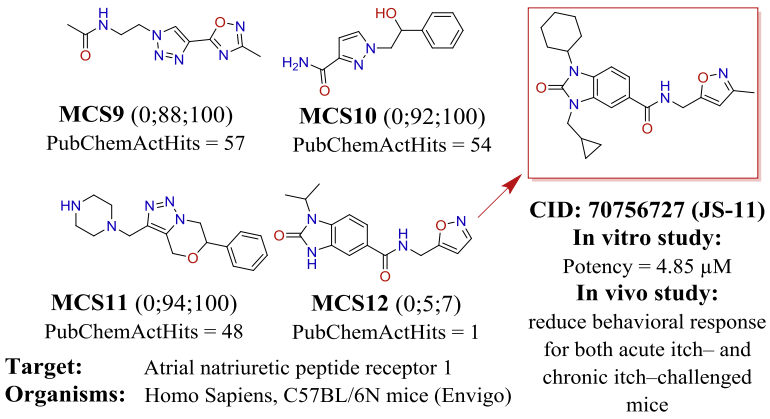


Figure 8. Examples of the ZINC-specific MCS, generalizing compound classes, tested in hNpr1 antagonism studies. Compound on the right (JS-11) has been ranked as the best inhibitor and was tested in vivo model, showing a decline in the behavioral response for itch-challenged mice.

Remaining 13 891 ZINC-specific MCSs absent from PubChem, were considered as "overlooked" by medicinal chemists and, thus, suggested as a guide for the more efficient

exploration of the purchasable chemical space. In order to assess their potential biological activity, 149K lead-like ZINC-Real compounds incarnating those MCSs were profiled against 749 ChEMBL biological targets, using the in-house GTM-based Profiler[25]. These results not intended to represent any specific "virtual screening campaign" pending experimental validation, but are shown as an illustration of the power of this multifaceted tool – both a chemical space map and activity profile predictor, at the same time. Their accuracy is, of course, essential, but that issue was already addressed in many other publications – both benchmarking studies[33] and prospective virtual screens[42, 43]. The conclusion is that they are slightly less accurate than machine-learned models, but acceptable because unlike the former "black box" models they are visual and intuitive.

As a result, 41K compounds (around 30% of the virtually screened molecules) were marked as potentially active against 525 ChEMBL biological targets. Half of the hits (Table 2) were predicted to be active only against a single target, another 21% - against two targets, and remaining compounds are predicted to be highly promiscuous (cumulating up to 18 activities). The MCS with the highest number of compounds predicted as actives are shown in Figure 9.

Table 2. Target specification of the profiled compounds.

Type of target	Number of targets	Number of predicted actives
Receptors	181	25 395
Enzymes	148	30 300
Kinases	108	5 860
Other targets	88	14 453

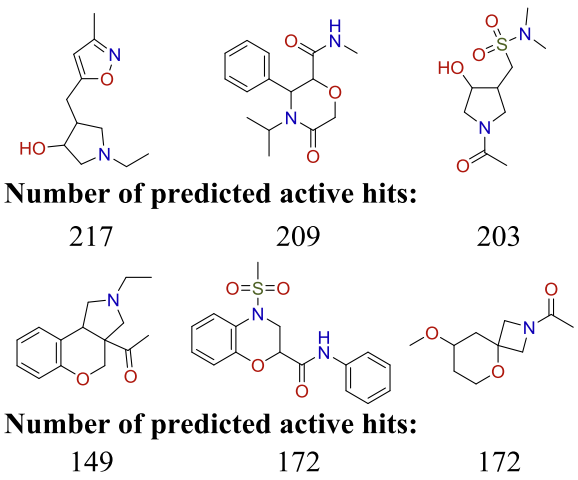


Figure 9. ZINC-Specific PubChem absent MCS, that had the higher number of corresponding compounds, predicted as actives using GTM-based Profiler

CONCLUSION

This HGTM analysis of the chemical space has provided a better understanding of the structural features of the purchasable chemical space. For the first time, all commercially available compounds have been taken into consideration, focusing on the detection of specific "open-ended" chemotypes (by contrast to scaffolds, maximum common substructures can be more specific by containing side chain substituents). It was shown that the chemical market is highly unbalanced, with a bias towards sulfonamides, amides, etc. Comparison of the main features of the in-stock and tangible compounds distribution demonstrated that tangible libraries still sample the same areas of the chemical space that were already overrepresented by in-stock molecules. Thus, there is a need for novel strategies of commercial library enhancement, which can provide a uniform chemical space sampling, avoiding the synthesis of a large number of close analogs. It goes without doubt that chemoinformatics and machine learning methods will be of paramount importance for the development of such strategies in the future.

At the same time, the biological relevance of the purchasable chemical space was assessed in this work. On one hand, it was discovered that a lot of compound families, known to include biologically active members, are absent from the in-stock catalogs of chemical suppliers. Some of them can be conveniently found in the tangible libraries, the most straightforward source of compounds for the in-stock enhancement campaign, while others are completely unavailable. In both cases, those substructures represent a potential source of inspiration for synthetic chemistry in search of enriching the commercial compound portfolio. On the other hand, the high number of ZINC-specific substructures demonstrates the limited extent of the biological exploration of purchasable libraries. Tens of thousands of such chemotypes encountered in neither ChEMBL nor PubChem can be used as a "novelty" guide for the further screening campaigns. More than 40.000 HGTMs generated in this work can be used in the future investigations of chemical space of any other library.

Finding library-specific substructures by comparing a 1.6M to an 800M-compound library is only rendered possible by means of the combination of the fast, zone-based clustering of compounds on GTMs and hierarchical zooming, allowing to focus on detailed chemical space zones within which the Maximum Common Substructure detection algorithm can be technically applied. A smooth and comprehensive link is herewith established between the universal map, providing bird's-eye view of the "Big Data" library and the specific substructures found in the particular chemical space zones.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge via the Internet at http://pubs.acs.org.

- Computational methods details
- Examples of the unique biologically relevant MCS for the commercially available libraries enhancement
- Examples of the unique ZINC MCS for the biological exploration of the commercially available chemical space

Link to the complete list of unique MCS for Fragment-Like, Lead-Like, Drug-Like and PPI-Like subsets - https://forms.gle/B6bUJj82t9EfmttV6

AUTHOR INFORMATION

Corresponding Author

Prof. A. Varnek, E-mail: varnek@unistra.fr.

ORCID

Yuliana Zabolotna: 0000-0001-9068-612X

Arkadii Lin: 0000-0002-9546-0012

Dragos Horvath: 0000-0003-0173-5714

Gilles Marcou: 0000-0003-1676-6708

Dmitriy M. Volochnyuk: 0000-0001-6519-1467

Alexandre Varnek: 0000-0003-1886-925X

Notes

The ISIDA/GTM software used in this work is available from the authors upon the request.

ABBREVIATIONS

HTS, high throughput screening; MCS, maximum common substructure; GTM, generative topographic mapping; HGTM, hierarchical generative topographic mapping; QSAR, quantitative structure-activity relationship; QSPR, quantitative structure-property relationship; PPI, protein-

protein interaction; ACC2, acetyl-CoA carboxylase 2; PAINS, pan-assay interference compounds.

REFERENCES

1. Sterling, T.; Irwin, J. J., ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, 55, 2324-2337.
2. Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* **2013**, 27, 675-679.
3. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, 40, D1100-D1107.
4. Walters, W. P., Virtual Chemical Libraries. *J. Med. Chem.* **2019**, 62, 1116-1124.
5. Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E., Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 643-651.
6. Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A., Drug- and Lead-likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers. *J. Chem. Inf. Model.* **2010**, 50, 470-479.
7. Lucas, X.; Grüning, B. A.; Bleher, S.; Günther, S., The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, 55, 915-924.
8. Petrova, T.; Chuprina, A.; Parkesh, R.; Pushechnikov, A., Structural enrichment of HTS compounds from available commercial libraries. *MedChemComm* **2012**, 3, 571-579.
9. Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K.-C., Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, 29, 55-67.
10. Verheij, H. J., Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Divers.* **2006**, 10, 377-388.
11. Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B., Evolution of commercially available compounds for HTS. *Drug Discov. Today* **2019**, 24, 390-402.
12. Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T., Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J. Cheminformatics* **2017**, 9, 25.
13. Bemis, G. W.; Murcko, M. A., The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.
14. Hu, Y.; Stumpfe, D.; Bajorath, J., Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2016**, 59, 4062-4076.
15. Schneider, G.; Schneider, P.; Renner, S., Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, 25, 1162-1171.
16. Shelat, A. A.; Guy, R. K., Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, 3, 442-446.
17. Beutler, J. A., Natural Products as a Foundation for Drug Discovery. *Current Protocols in Pharmacology* **2019**, 86, e67.
18. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **2003**, 8, 876-877.
19. Gleeson, M. P., Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, 51, 817-834.
20. Lipinski, C. A., Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, 44, 235-249.
21. Morelli, X.; Bourgeas, R.; Roche, P., Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011**, 15, 475-481.
22. Bishop, C. M.; Svensén, M.; Williams, C. K. I., GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, 10, 215-234.
23. Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A., Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, 13, 540-554.
24. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E., PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2018**, 47, D1102-D1109.
25. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A., Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J. Chem. Inf. Model.* **2019**, 59, 564-572.
26. Oprea, T. I.; Gottfries, J., Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, 3, 157-166.
27. Kohonen, T., Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, 43, 59-69.
28. Cao, Y.; Jiang, T.; Girke, T., A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **2008**, 24, i366-i374.
29. Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A., Diversifying chemical libraries with generative topographic mapping. *J. Comput. Aided Mol. Des.* **2019**.
30. Tino, P.; Nabney, I., Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE PAMI* **2002**, 24, 639-656.
31. Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A., GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **2015**, 34, 348-356.
32. Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A., Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012**, 31, 301-312.

33. Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A., Multi-task generative topographic mapping in virtual screening. *J. Comput. Aided Mol. Des.* **2019**, 33, 331-343.
34. Hann, M. M.; Leach, A. R.; Green, D. V. S., Computational Chemistry, Molecular Complexity and Screening Set Design. *Chemoinformatics in Drug Discovery* **2005**, 43-57.
35. Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J., De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **2020**, 11, 10.
36. Reymond, J.-L.; Awale, M., Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, 3, 649-657.
37. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, 29, 855-868.
38. Chast, F. Chapter 1 - A History of Drug Discovery: From first steps of chemistry to achievements in molecular pharmacology. In *The Practice of Medicinal Chemistry (Third Edition)*, Wermuth, C. G., Ed.; Academic Press: New York, 2008, pp 1-62.
39. Jha, K. K.; Kumar, S.; Tomer, I.; Mishra, R., Thiophene: the molecule of diverse medicinal importance. *J. Pharm. Res* **2012**, 5.
40. Grygorenko, O. O.; Volochnyuk, D. M.; Ryabukhin, S. V.; Judd, D. B., The Symbiotic Relationship Between Drug Discovery and Organic Chemistry. *Chem. Eur. J.* **2020**, 26, 1196-1237.
41. Solinski, H. J.; Dranchak, P.; Oliphant, E.; Gu, X.; Earnest, T. W.; Braisted, J.; Inglese, J.; Hoon, M. A., Inhibition of natriuretic peptide receptor 1 reduces itch in mice. *Sci. Transl. Med.* **2019**, 11, eaav5464.
42. Orlov, A. A.; Khvatov, E. V.; Koruchekov, A. A.; Nikitina, A. A.; Zolotareva, A. D.; Eletskaya, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Horvath, D.; Osolodkin, D. I.; Varnek, A., Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol. Inform.* **2019**, 38, 1800166.
43. Casciuc, I.; Horvath, D.; Gryniukova, A.; Tolmachova, K. A.; Vasylchenko, O. V.; Borysko, P.; Moroz, Y. S.; Bajorath, J.; Varnek, A., Pros and cons of virtual screening based on public “Big Data”: In silico mining for new bromodomain inhibitors. *Eur. J. Med. Chem.* **2019**, 165, 258-272.

SYNOPSIS TOC