Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery

Yuliana Zabolotna, Fanny Bonachera, Dragos Horvath, Arkadii Lin, Gilles Marcou, Olga Klimchuk, Alexandre Varnek

▶ **To cite this version:**

Yuliana Zabolotna, Fanny Bonachera, Dragos Horvath, Arkadii Lin, Gilles Marcou, et al.. Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery. Journal of Chemical Information and Modeling, 2022, 62 (18), pp.4537-4548. 10.1021/acs.jcim.2c00509 . hal-04244044

HAL Id: hal-04244044
https://hal.science/hal-04244044

Submitted on 16 Oct 2023

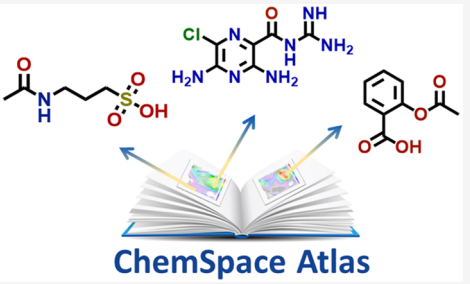
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery

Yuliana Zabolotna, Fanny Bonachera, Dragos Horvath, Arkadii Lin, Gilles Marcou, Olga Klimchuk, and Alexandre Varnek*

ABSTRACT: Nowadays, drug discovery is inevitably intertwined with the usage of large compound collections. Understanding of their chemotype composition and physicochemical property profiles is of the highest importance for successful hit identification. Efficient polyfunctional tools allowing multifaceted analysis of constantly growing chemical libraries must be Big Data-compatible. Here, we present the freely accessible ChemSpace Atlas (https://chematlas.chimie.unistra.fr), which includes almost 40K hierarchically organized Generative Topographic Maps (GTM) accommodating up to 500 M compounds covering fragment-like, lead-like, drug-like, PPI-like, and NP-like chemical subspaces. They allow users to navigate and analyze ZINC, ChEMBL, and COCONUT from multiple perspectives on different scales: from a bird's eye view of the entire library to structural pattern detection in small clusters. Around 20 physicochemical properties and almost 750 biological activities can be visualized (associated with map zones), supporting activity profiling and analogue search. Moreover, ChemScape Atlas will be extended toward new chemical subspaces (e.g., DNA-encoded libraries and synthons) and functionalities (ADMETox profiling and property-guided de novo compound generation).



INTRODUCTION

With the rapid growth of compound libraries over the last decades,[1] drug discovery campaigns resemble a search for the proverbial needle in the haystack. One of the most significant contributors to the expansion of chemical data is combinatorial chemistry.[2] However, many of the early combinatorial libraries are now considered far from the optimal chemical space of drug discovery.[3] The realization that unbiased library synthesis and screening cannot revolutionize drug discovery and overshadow natural products led to the "fall" of combinatorial chemistry.[4] In response, medicinal chemists turned to virtual (also called tangible) compound libraries in a search for higher diversity, quality, and novel chemotypes.[5] It became the state of the art to use tangible libraries for virtual screening (VS) in order to obtain a more extensive and diverse pool of virtual hits, out of which a subset would be selected for synthesis and experimental testing. This trend encouraged the creation of numerous and ever larger virtual libraries,[6−13] making it hardly possible to comprehend the entire scope of all available compounds.

Under such conditions, a deep understanding of the data currently available for medicinal chemists is of the highest importance and efficient computational techniques that allow analysis of the large amount of chemical data now play a crucial role in the early stages of drug design. Over the last decades, multiple standalone software tools[14−18] embodying the concept of chemical space[19]—an abstract space in which points represent compounds with clearly defined neighborhood relationships—has been developed. They provide a wide range of functionalities for chemical space visualization and analysis. However, they can be difficult to install and maintain.[20] Their usage may require technical coding or scripting skills, making them available mostly to chemoinformatics professionals.[21] Therefore, online resources can be a more convenient choice as soon as they usually are intuitive and relatively easy to use.

Freely Accessible Web Tools for the Interactive Chemical Space Visualization. At the moment, there are about a dozen freely accessible online servers that allow navigation and analysis of the chemical spaces defined by different MedChem relevant libraries (see Table S1 in the Supporting Information (SI)). Almost all of them rely on the vector-based chemical space representation methods: PCA[22] and t-SNE.[23] Only the tMap server[24] features a graphlike representation. Most of those tools visualize only previously processed libraries, but some servers allow users to project a limited set of user-defined compounds. However, the latter usually takes a long time and web site crashes are not

uncommon. The size of the precomputed data sets varies from 10^2 to 10^7, which appears to be a current limit for most of chemical space visualization techniques.

Moreover, the larger the covered collections, the fewer functionalites are supported. Indeed, the two implementations that enable navigation among up to 10 M compounds—tMap[24] and Faerun[25]—provide only simple visualization of physicochemical properties (and activity in case of tMap) without the possibility to project new data for analysis. In addition, interpretability and convenience of navigation expectedly drop for the largest chemical spaces, for all earlier reported tools provide only a global "bird's eye" view of chemical space.

At the same time, more focused navigators like PUMA[20] and ChemMaps[21] provide users with broader functionalities allowing new data set projections and their comparison with precomputed libraries. In the case of PUMA, diversity analysis (scaffold and fingerprint diversity plots, etc.) is also available, and ChemMaps has an option of toxicity prediction. Nevertheless, none of the existing web implementations support activity profiling, even though the activity maps can be displayed. Another significant shortcoming of existing tools is the availability of only one global view on the chemical space, without the possibility to analyze the local features of smaller clusters containing close analogues. It also explains the absence of structural functionality, such as scaffold and maximum common substructure (MCS) analysis.

The aforementioned weaknesses of existing tools are mainly caused by the limitations of the underlying chemical space representation techniques. For instance, PCA can process massive datasets only if they have linearly dependent features. The standard solution, in this case, would be to train the model using a representative subset and project the remaining data onto the 2D map. t-SNE is a nonlinear method and thus overcomes this drawback. However, because of the need to store a distance matrix for the entire dataset, t-SNE is limited, in terms of set sizes. Also, adding data to a precursory t-SNE map is not feasible—a drawback also inherent to the tMaps algorithm.

ChemSpace Atlas Navigator. Considering the main trends in drug discovery, chemical space navigation cannot be limited to a simple visualization of similarity relationships for compounds from predefined libraries. Various property and biological activity visualization, polypharmacological profiling, analogue search, and detailed structural analysis should also be available. Moreover, all of these must be "Big Data"-compatible in order to cope with hundreds of millions, and even billions, of compounds. We have repeatedly used Generative Topographic Mapping (GTM)[26] as a highly efficient dimensionality reduction method that possesses numerous advantages and overcomes many drawbacks of other approaches. In contrast to Self-Organizing Maps,[27] GTM distributes molecule projections over the map with node-specific probabilities (responsibilities) instead of unambiguously assigning each compound to only one point on the map (see Figure S1 in the SI). This smoothness enables the creation of GTM landscapes: maps, colored by average values of different (biological, physicochemical, etc.) properties (Figure S2 in the SI). These maps can be turned into potent quantitative structure−activity/property relationship (QSA/PR) engines.[28−31] A single GTM manifold is able to host more than one predictive landscape. Hundreds of properties/activities can be predicted simultaneously using correctly optimized "universal" GTM: a general-purpose map that can accommodate ligands of diverse biological targets on the same GTM manifold (for more details, see the SI).[28,32] However, not all activities are equally well predicted by a single universal map. In order to achieve better predictive performance for maximal number of biological activities, not one, but several universal maps based on different types of descriptors can be used (see Table S3 in the SI). Such maps represent complementary and strongly synergistic views of biologically relevant chemical space. They can be used not only as a predictive tool but also as frameworks for the analysis of large chemical libraries in the medicinal chemistry and drug design context.

There is no limitation of the number of items that can be hosted by a GTM, but a 2D map charged with Big Data-level libraries may only render generic "bird's eye" views of the common and specific chemical space zones covered. However, the hierarchical zooming approach (hGTM)[33,34] provides a "pile" of hierarchical maps connecting the bird's eye view down to detailed maps of specific neighborhoods that are easy to annotate by individual or common (sub)structures (see Figure S3 in the SI). It provides intuitive, easy-to-use, and highly interpretable global and local views of the chemical space and enables efficient structural analysis of the selected areas. Zooming is achieved by refitting a local manifold to optimally cover the residents of a chemical space zone defined by the 3×3 grid of nodes around the user-picked central node. Zones overlap, so the same node may participate to up to four zones (see Figure S3). "Residents" of a zone are defined as compounds with cumulative responsibility over the 9 nodes (sum of probabilities to be assigned to these nodes) of the zone above 0.85. Moreover, as new information emerges every day, it is a significant advantage that new data points can be easily projected onto the existing maps without retraining any GTM. All of that makes GTM one of the best choices for developing a new chemical space visualization tool with extended functionality.

The herein presented, intuitive web tool ChemSpace Atlas (https://chematlas.chimie.unistra.fr/) assembles tens of thousands of hierarchically related GTMs based on nine universal maps (see Tables S2 and S3), covering biologically relevant chemical (sub)spaces. Seven of them were trained and optimized using ChEMBL data in a way to host simultaneously ligands of diverse biological targets and serve as an efficient activity profiling platform.[28] The first universal map was further used to analyze chemical space defined by biologically tested compounds from ChEMBL, commercially available molecules for HTS from ZINC, and DNA-encoded libraries enumerated using purchasable BBs. Thus, in the ChemSpace Atlas, those respective sections are based on the first uMap. However, as soon as there is a limited number of NPs in ChEMBL, a specific NP-uMap was constructed using compounds from the COCONUT collection of NPs.[35] Similarly, a dedicated universal map of synthons was created (without considering the leaving groups in actual reagents).[36] This map was trained on synthons generated both from commercially available reagents and ChEMBL compounds (via their fragmentation).

The hierarchy, obtained as a result of "zooming" of the universal maps, enables convenient navigation through the hundreds of millions of compounds from a global bird's eye view to structural pattern detection. On each level, there are several landscapes colored according to the different properties of corresponding compounds (see Figure S2). ChemSpace Atlas is based on previously published research, interconnected by an easy-to-use web interface. It consists of six modules: "Fragment-Like", "Lead-Like", "Drug-Like", "PPI-Like", and "NP-like"

Table 1. Description of Eight Navigators Composing Chemspace Atlas[a]

navigator name	featured libraries	size of the analyzed chemical space (after standardization and filtration)	main uMap	number of hGTMs in the hierarchy
Natural Products Navigator	COCONUT	253K	NP-uMap[35]	241 hGTMs
	NP-Like ChEMBL (v26)	474K		
	NP-Like ZINC20	586K		
Fragment-Like Chemical Space Navigator	FL ChEMBL (v25)	15K	1st uMap of ChEMBL[28]	880 hGTMs
	FL ZINC15 (stock)	103K		
	FL ZINC15 (tangible)	2.7M		
Lead-Like Chemical Space Navigator	LL ChEMBL(v25)	363K	1st uMap of ChEMBL	11 150 hGTMs
	LL ZINC15 (stock)	3.2M		
	LL ZINC15 (tangible)	329M		
Drug-Like Chemical Space Navigator	DL ChEMBL(v25)	668K	1st uMap of ChEMBL	22 325 hGTMs
	DL ZINC15 (stock)	5.1M		
	DL ZINC15 (tangible)	516M		
PPI-Like Chemical Space Navigator	PPIL ChEMBL(v25)	229K	1st uMap of ChEMBL	3 294 hGTMs
	PPIL ZINC15 (stock)	1.2K		
	PPIL ZINC15 (tangible)	63M		
ChEMBL Activity Space Navigator and Activity Profiler	Visualization: ChEMBL (v26)	1.7M	1st uMap of ChEMBL	–
	Profiler: ChEMBL(v24)	1.6M	seven uMaps of ChEMBL[28]	

[a]Featured libraries, their size, underlying uMap, and the size of the hierarchy in case hierarchical zooming was performed.

chemical space Navigators,[35,37] and "ChEMBL activity space Navigator/Activity profiler".[28]

IMPLEMENTATION

Data. ChemSpace Atlas covers several libraries that are frequently used in drug discovery. Among them, there are ZINC,[38] ChEMBL,[39] and COCONUT[40] (Table 1). All those compounds were standardized according to the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics at the University of Strasbourg (infochimie.u-strasbg.fr/webserv/VSEngine.html) using the ChemAxon Standardizer.[41] That included:

- dearomatization and final aromatization (heterocycles such as pyridone were not aromatized)
- conversion to canonical SMILES;
- salts and mixture removal;
- neutralization of all species, except nitrogen(IV);
- major tautomer generation;
- stereochemical information removal.

Stereochemical information has been ignored due to the fact that ISIDA descriptors,[42] used in this work, would not capture it anyway. All stereoisomer IDs were assigned to only one stereochemistry-depleted chemical structure.

Biologically Relevant Chemical Space. The biologically relevant chemical space is represented by ChEMBL—a large-scale collection of bioactivity data from binding, functional, and ADMET assays.[39] ChemSpace Atlas will regularly update the implemented ChEMBL release—both in terms of adding new compounds to the pre-existing maps and updating the links between (stereochemistry-depleted) SMILES and the associated compound ChEMBL IDs. Predictive landscapes will improve as future ChEMBL releases provide more structure–activity data to "color" the maps, and "near-orphan" targets for which no such predictive landscapes could be generated, because of insufficient experimental data, may eventually change their status and be co-opted into the predictable polypharmacological profile of ChemSpace Atlas. Reconstruction of underlying manifolds of local zoomed maps may also be necessary for zones witnessing a significant population increase, while the refitting of the universal manifolds is not necessary unless the update of the activity landscapes based on the current manifolds fails to generate new highly predictive landscapes.

Commercially Available Chemical Space. ZINC collection (versions 15 and 20; see Table 1) was used in this work as a representation of the purchasable chemical space.[11,43] It is a publicly available database that collects commercially available compounds from various chemical vendors and annotated compounds from libraries such as PubChem and ChEMBL. All compounds in ZINC are grouped into several purchasability categories:[43]

- in stock — delivery in under 2 weeks, 95% typical acquisition success rate;

Table 2. Examples of Possible Medicinal Chemistry Questions, Popular in the First Stages of Drug Design and Chemspace Atlas Solutions

MedChem questions	ChemSpace Atlas solutions
I. Unbiased Exploration of the Chemical Space	
What kind of compounds populate the analyzed libraries? What are their common structural features? What are the most represented chemotypes?	Density landscapes of ZINC, ChEMBL, or COCONUT libraries + MCS/scaffold analysis
Where can a mapped compound be purchased?	Direct links to the ZINC Web site
What biological activity the selected compound is reported to have?	Direct links to the ChEMBL Web site
Can the selected compound be found in natural sources (and where)?	Direct links to the COCONUT Web site
II. Analysis of the Chemical Space in the Context of a Given Biological Target	
What kind of compounds (chemotypes) were tested biologically against a target of interest? What are the major chemotypes in compounds tested against a given biological activity?	Density landscapes of target-specific ligand series from ChEMBL + MCS/scaffold analysis
What is the structural difference between actives and inactives?	Binary class landscapes (actives versus inactives of a selected target) + MCS/scaffold analysis
Which actives have a desired property profile?	Property landscapes of ZINC, CHEMBL, or COCONUT libraries
What compounds should be tested next against a given biological target?	Consensus activity predictor (for one target)
III. Analysis of the Chemical Space in the Context of Given Compounds (Query Molecules)	
Analysis of the knowledge space (ChEMBL Navigator)	
Was there anything similar to my query molecules already reported in ChEMBL?	"ChemSpace Tracker" functionality for the whole ChEMBL density landscapes provided in different sections of the tool
Are there any actives of the selected biological target similar to the query molecules?	"ChemSpace Tracker" functionality for the target-specific binary class landscapes
Analysis of the commercially available chemical space	
Are there any commercially available compounds similar to the query molecules?	"ChemSpace Tracker" functionality for the ZINC density landscapes
Are there any "scaffold hopping" analogues for the query compounds?	"ChemSpace Tracker" functionality for the ZINC density landscapes + MCS/scaffold analysis
Analysis of the NP and NP-like chemical space (NP Navigator)	
Is there any naturally occurring compound similar to the query molecules?	"ChemSpace Tracker" functionality for the COCONUT landscapes
Is there any pseudo-NP similar to the query molecules?	"ChemSpace Tracker" functionality for the NP-like ZINC landscapes
Activity profiling of the given molecules	
To which targets may my compounds bind?	Consensus activity profiler

- procurement agent — in stock, delivery in 2 weeks, 95% typical acquisition success rate;
- make-on-demand — delivery typically within 8−10 weeks, 70% typical acquisition success rate;
- boutique — where the cost may be high but still likely less expensive than synthesis from scratch, 70% typical acquisition success rate.

In ChemSpace Atlas, the first two groups were combined as "in-stock commercially available" subset. All the rest formed the tangible subset. Note that ZINC compounds are not employed for the construction of universal map manifolds, but may serve to define zoomed local manifolds. The steady increase of the ZINC database may over time require additional hierarchical levels being added to ChemSpace Atlas.

Chemical Space of Natural Products. The COlleCtion of Open Natural prodUcTs (COCONUT) is the most complete up-to-date dataset of natural products (NPs), containing 406 076 unique compounds with no stereochemistry.[40,44] They were extracted from 53 various data sources, like Traditional Chinese Medicine database,[45] Marine Natural Products,[46] Collective molecular activities of useful plants,[47] Super Natural II,[48] etc. All compounds were curated, registered, and annotated with various precomputed molecular properties.

Structure of ChemSpace Atlas. ChemSpace Atlas was designed as a container for several subspace navigators, which can be accessed from the main page (https://chematlas.chimie.unistra.fr/):

- Fragmentlike Chemical Space Navigator
- Leadlike Chemical Space Navigator
- Druglike Chemical Space Navigator
- PPI-Like Chemical Space Navigator
- Natural Products Navigator.

In addition, the "ChEMBL activity space Navigator and activity Profiler" section uses series of compounds with reported biological activity against all biological targets for which large enough sets exist. In the current version, this applies to 749 distinct human, rodent, and parasitic targets (please refer to previous works[28,32] for the automated extraction, activity binning rules, and criteria for rejection or inclusion of such "large enough" sets in in the activity space Navigator). Fuzzy activity classs landscapes based on these series are used for pharmacological profiling using consensus activity class prediction on seven universal maps[28] (see details in the SI).

Each of the navigators listed above is focused on specific subspaces of the biologically relevant chemical space that differ in size (Table 1): from 10^5 in the case of natural products to 10^8 for druglike structures. Each of the six navigators is based on the separate hierarchy of maps developed and reported in previous publications. The universal GTMs were evolved with the help of a genetic algorithm,[49] which allowed optimal descriptor space and GTM parameters selection. Zoomed maps were then constructed using the parameters of the main map and frameset composed of compounds localized in a specific zone (see more details in the SI). The descriptors defining chemical spaces are different variations of ISIDA fragment descriptors[42] from simple atom sequencing to complex variations labeled by force-field constants, formal charges, and pharmacophoric features. Apart from the libraries that have been already projected onto the

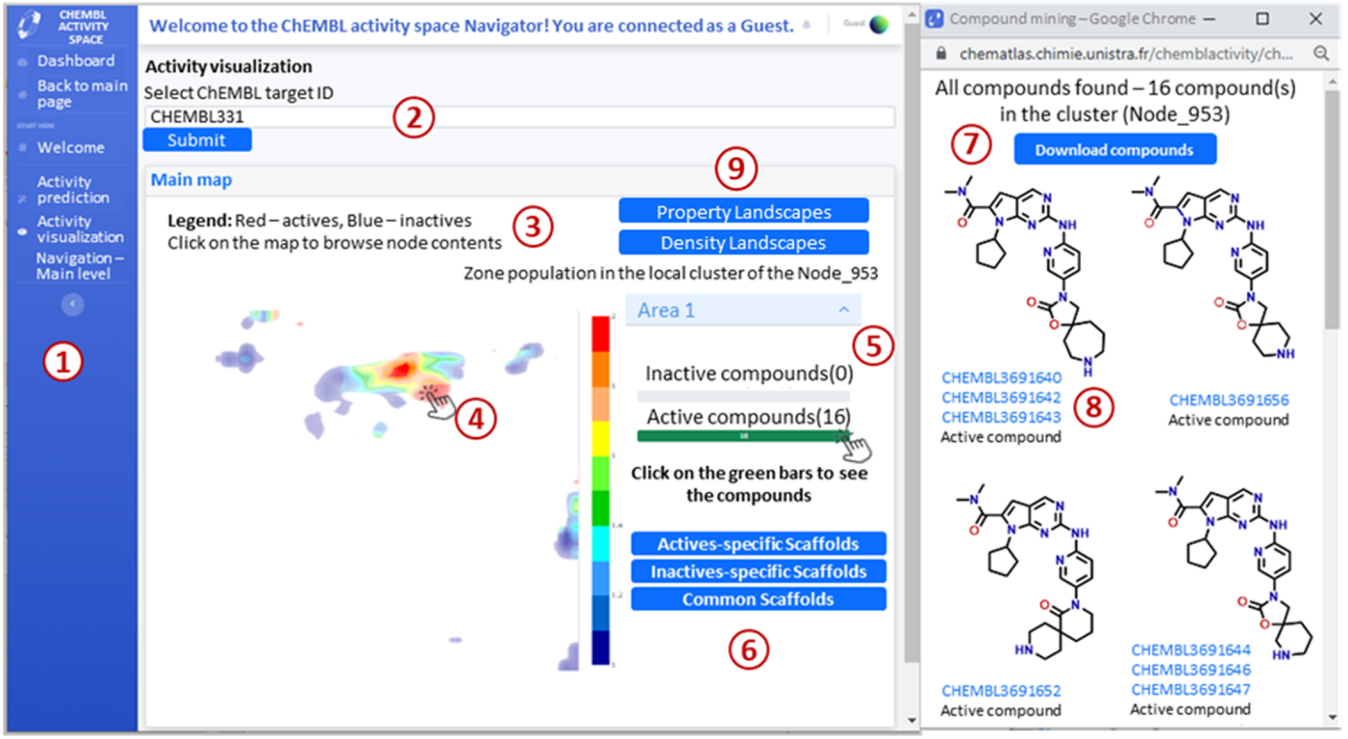


Figure 1. Activity visualization page of the ChEMBL activity space Navigator on the example of CDK4 (CHEMBL331). ① navigation bar; ② target selection menu; ③ legend of the map; ④ interactive activity landscape; ⑤ zone population information (if green, bars become clickable and corresponding compounds can be displayed); ⑥ buttons for displaying zone-typical scaffolds; ⑦ button for downloading compounds from the selected area of the chemical space; ⑧ compound identifiers/direct links to the source database (here, ChEMBL); ⑨ buttons to display property and density landscapes in the pop-up window.

hGTMs, new collections can be placed on these maps, leaving numerous possibilities for further ChemSpace Atlas extensions.

Functionality of ChemSpace Atlas. At the early stages of modern drug design, medicinal chemists are mostly working with large and ultralarge compound collections as the source of "hits" leading to future drug candidates. Understanding the chemotype composition and physicochemical property profiles of the starting library, as well as physicochemical properties of composing compounds, is prerequisite for a rational selection process. Distribution histograms featuring physicochemical properties and scaffold frequency plots provide a generalized picture for the entire library, but cannot be backtracked to underlying chemotypes. This is the strength of ChemSpace Atlas: hGTM separates—at varying levels of hierarchical resolutions—the chemotypes on the maps, all while associating particular properties of interest of local resident clusters, by means of property landscapes.

There are three main types of chemical space analysis performed in drug design:

- unbiased exploration;
- target-oriented (with respect to the biological target responsible for the particular activity); and
- compound-oriented (focused on compounds possessing particular structural features).

Table 2 summarizes the most popular chemical space-related questions that appear during each of the above-mentioned types of analysis and ChemSpace Atlas solutions that allow us to answer them. Tables S4–S6 in the SI provide a comparative analysis of our tool with previously existing web implementations in the context of the questions provided in Table 2. Moreover, in the SI, one can also find four case studies for each of the mentioned tasks with detailed instructions of ChemSpace Atlas interface usage and main insights derived with its help.

From the main page of ChemSpace Atlas (https://chematlas.chimie.unistra.fr/), one can select the section of the chemical space to explore. All Navigators support the same set of options. Almost 20 various physicochemical properties and almost 750 activity landscapes allow users to analyze libraries from different perspectives:

- physicochemical property visualization (18 calculated properties);
- activity visualization (749 ChEMBL activities);
- activity prediction (for a selected ChEMBL target);
- activity profiling (749 ChEMBL activities);
- tracking specific areas of the chemical space based on structural features ("ChemSpace Tracker");
- analogue search ("ChemSpace Tracker");
- structural analysis of selected regions of the chemical space with the help of MCSs and scaffolds;
- precomputed library comparison (ChEMBL vs ZINC, ZINC vs COCONUT, ChEMBL vs COCONUT, etc.).

In order to facilitate navigation, a small set of "tracking" compounds can be provided by the user. These molecules will be projected onto the GTMs, appearing as dots (at the barycenter of their responsibility "clouds") on the selected landscapes. These dots will help to choose the zones of the chemical space worth exploring in the context of users' needs. "ChemSpace Trackers" also allow users to search for analogues of a provided compound. Apart from simple navigation, ChemSpace Atlas can be used for efficient analysis of underlying libraries: chemotype distribution, physicochemical properties, (reported and/or predicted) biological activity, and commercial availability.

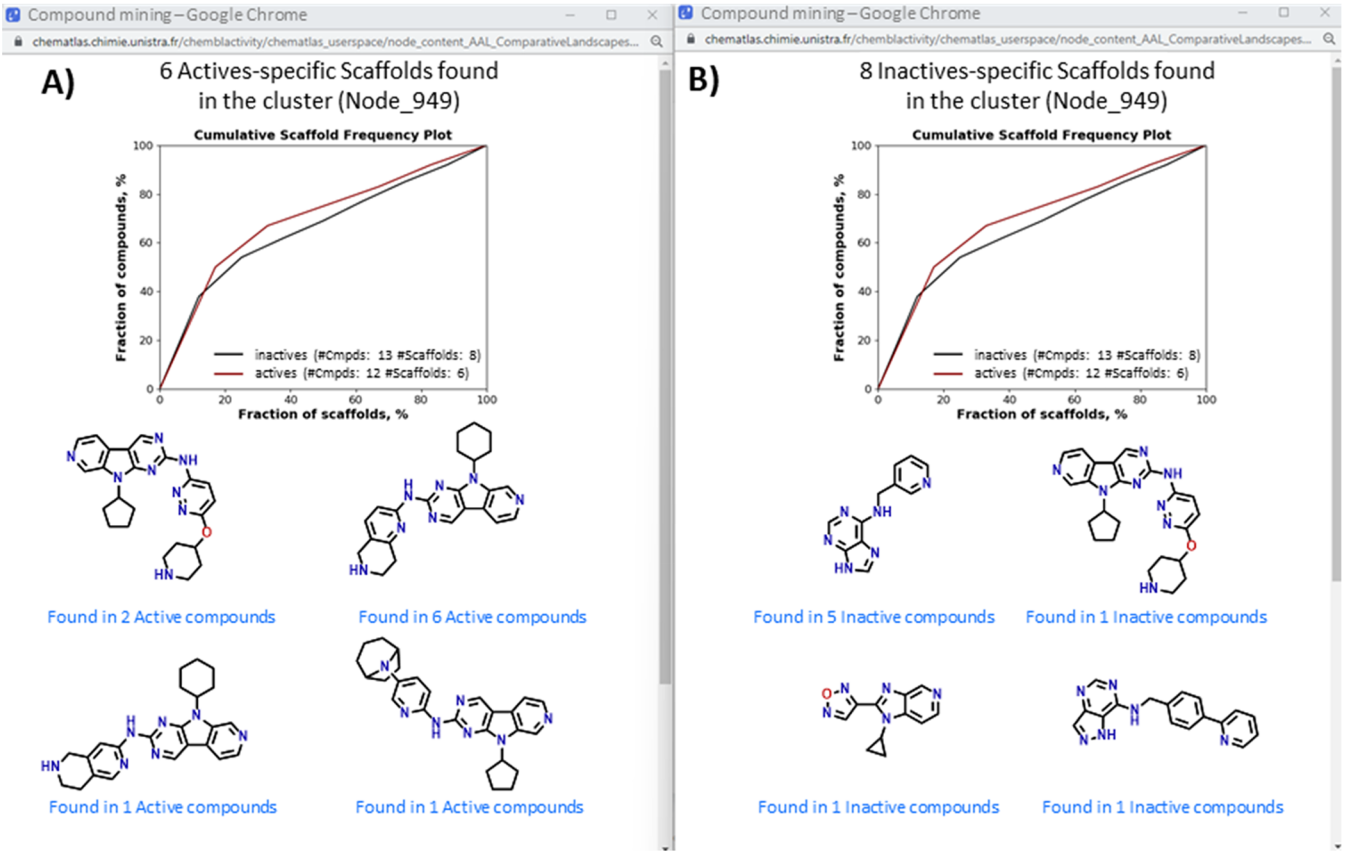


Figure 2. Scaffold analysis of ligands CDK4 (CHEMBL331) residing in a mixed zone around Node 949: (A) actives-specific scaffolds and (B) inactives-specific scaffolds. On the top of the page, the scaffold frequency plot providing information on the distribution of molecules over scaffolds is given.

Moreover, activity prediction based on the consensus model of landscapes based on seven universal maps is also available.

Interface of ChemSpace Atlas Exemplified on Cell Division Protein Kinase 4 Inhibitor Monitoring. *Target-Oriented Chemical Space Navigation.* From the main page of NP Navigator, one can access the ChEMBL activity space Navigator and activity Profiler section. On the Activity visualization page (Figure 1), the user can provide a ChEMBL identifier of the biological target of interest ("②" in Figure 1). As a case study, one of the important targets for breast cancer treatment was used: cell division protein kinase **4** (CDK4, CHEMBL331). The displayed landscape ("④" in Figure 1) relies on a compound ligand series from ChEMBL database, tested against CDK4 and "colored" by the relative prevalence of "actives" (defined as compounds with IC_{50} or K_i lower than a given threshold Th_{act}) versus inactives (compounds with IC_{50} or K_i higher than $10 \times Th_{act}$), where Th_{act} is chosen in such a way to ensure a relative balance of actives versus inactives in this "coloring" set. Compounds of intermediate activity ($Th_{act} < IC_{50}/K_i < 10 \times Th_{act}$) and molecules (rendered by a unique stereochemistry-depleted structure) featuring both some "active" and some "inactive" stereoisomers as above-defined are not included in the "coloring" set.

Each node of the map is colored according to the weighted class ratio of its residents: red regions contain exclusively active compounds, blue regions contain exclusively inactives, all colors between encode zones with members of both classes in different proportions; white areas are empty (see "③" in Figure 1). This map is interactive and by clicking on the selected area of the landscape, the detailed information about zone composition (3 × 3 nodes around the selected point) can be displayed ("⑤" in Figure 1). As mentioned above, zooming zones are overlapping, so the selected node on the map may belong to one or several (up to four) neighboring areas. Adjacent zones may contain very similar compounds, although some of the molecules might be present only in one of these. Therefore, for the detailed analysis, it is recommended to explore all zones proposed by ChemSpace Atlas. Two bars that illustrate the number of compounds residing in the area are clickable and allow one to display respective chemical structures in a pop-up window and download them ("⑦" in Figure 1). The source identifiers provided for each molecule are hyperlinked to the corresponding library's web interface, allowing direct access to the detailed compound's information ("⑧" in Figure 1). One mapped item may point to multiple compound identifiers if, in the source library, there were several stereoisomers.

It is also possible to perform scaffold analysis of the selected zone ("⑥" in Figure 1) and display Bemis-Murcko scaffolds specific to actives and inactives separately, as well as shared scaffolds if any were found (Figure 2). The number of the compounds corresponding to each scaffold is provided under each structure, and the text is hyperlinked to the window with corresponding compounds and their IDs. In addition, the cumulative scaffold frequency plot is showing the percentage of compounds that corresponds to the percentage of the most frequent scaffolds. This graph provides information on the distribution of molecules over scaffolds and can be used to estimate scaffold diversity of the analyzed library or compare

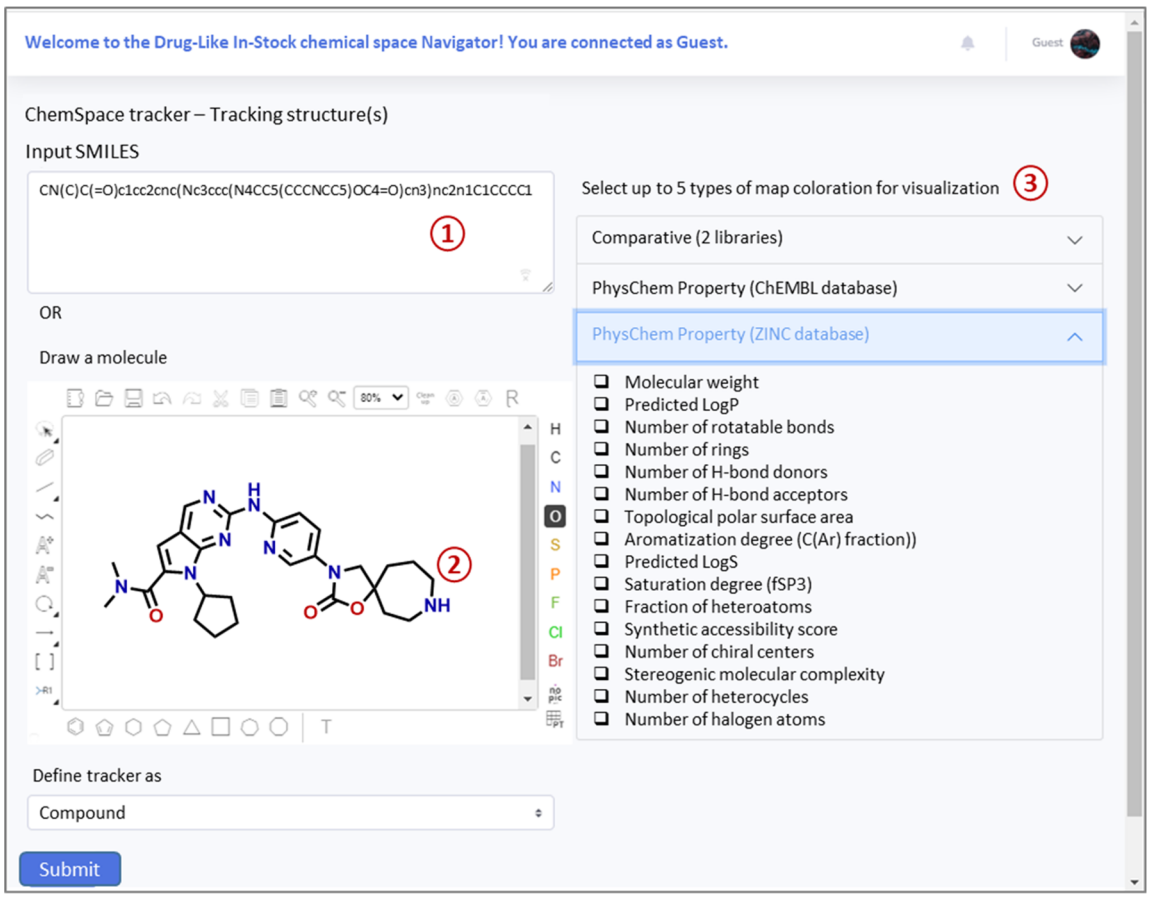


Figure 3. Input page on Chemspace Tracker: ① zone of text input (SMILES); ② structure sketcher; ③ selection of up to five landscape types.

different collections: the closer the curve is to the diagonal line, the more diverse the library is.[50] In Figure 2, actives and inactives of CDK4 residing in the local zone around Node 949 are compared.

To obtain the complementary view on the chemical space of analyzed compounds, one can display density or property landscapes ("⑨" in Figure 1). See the example of the detailed analysis of CDK4 ligands with the help of these landscapes in the SI (case study 2). The major chemotypes of biologically tested against CHEMBL331 compounds were detected using density landscape. Their structural features were discussed and difference between actives and inactives were highlighted using compound-by-compound comparison (see Figures S12, S13, and S15 in the SI) and scaffold analysis (see Figures S14 and S16 in the SI) in the low-populated local areas. ChemSpace Atlas functionality has also allowed an easy detection of an activity cleaf (Figures S13 and S14 in the SI). With the help of property landscapes (Figure S17 in the SI), it was concluded that the main difference between actives and inactives lies in the number of rings in general and heterocycles in particular and in the number of H-bond acceptors. At the same time number of H-bond donors was not really helpful in discriminating actives from inactives. In addition, it was shown that two main "islands" of active compounds have different values of molecular weight and the fraction of sp^3-hybridized carbon atoms. In addition, In case study 3 in the SI, the activity against CDK4 was predicted for the set of new compounds. The resulting hits were analyzed with the help of activity landscapes not only of CDK4, but also three other targets from the same kinase family—CDK1, CDK7, and CDK9—in order to assess the potential selectivity of those compounds.

Thus, the Activity visualization page allows users to analyze the activity space of the selected target, compare actives and inactives, detect areas of the chemical space enriched with active ligands, and characterize these areas with respect to the structural features of the underlying compounds. In addition, Activity Predictor (see Case Study 3 in the SI) and Activity Profiler (Case Study 4 in the SI) allow users to predict the potential biological activities of up to 50 compounds.

Compound-Oriented Chemical Space Navigation. For the further exploration of the chemical space around the actives of CDK4, the first ligand from Figure 1 (matching ChEMBL IDs CHEMBL3691640, CHEMBL3691642, and CHEMBL3691643) was selected as a compound of reference for the compound-oriented chemical space analysis of the Drug-Like chemical space of the ChEMBL and ZINC database. This compound is a potent inhibitor of CDK4 that has slightly varying IC_{50} values for its different stereoisomers, from 12 nM to 17 nM.

On the ChemSpace tracker page of the Drug-Like In-Stock chemical space Navigator, the user can provide a list of up to 50 SMILES ("①" in Figure 3) or draw a molecular structure in the sketcher window ("②" in Figure 3). These could, for example, be chosen to include potent CDK4 actives and hence play the role of chemical "trackers" that allow one to pinpoint the regions of the chemical space where known inhibitors of CDK4 reside.

Upon compounds submission, they will be standardized, fragmented to calculate ISIDA descriptor vectors, and projected onto the first universal map. On the right side of the page, the

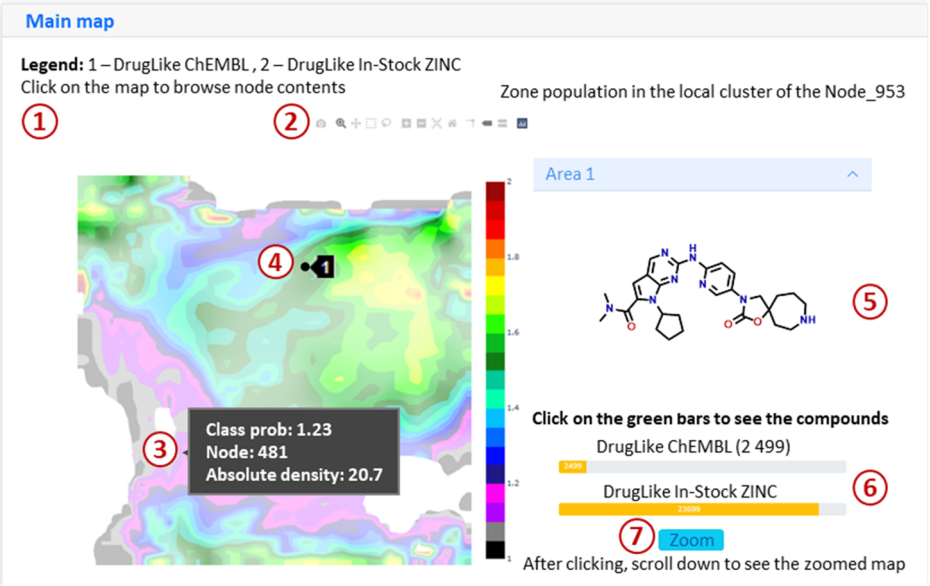


Figure 4. Main level of the landscape visualization: ① legend of the map; ② Plotly toolbar allowing different types of navigation through the plot; ③ hover-activated information about the node composition (absolute density corresponds approximately to the number of compounds residing in the node, and class probability indicates the proportion of ChEMBL(1) and ZINC(2) compounds); ④ black dots represent user-defined molecules—ChemSpace trackers—and hover-activated ChemSpace tracker information (index number of compound in the provided list); ⑤ selected tracking compound; ⑥ the number of closest neighbors of the selected compound on this level of hGTM (if green, bars become clickable and corresponding compounds can be displayed); ⑦ zoom button enabling display of the next level of navigation focusing on the selected zone of the chemical space.

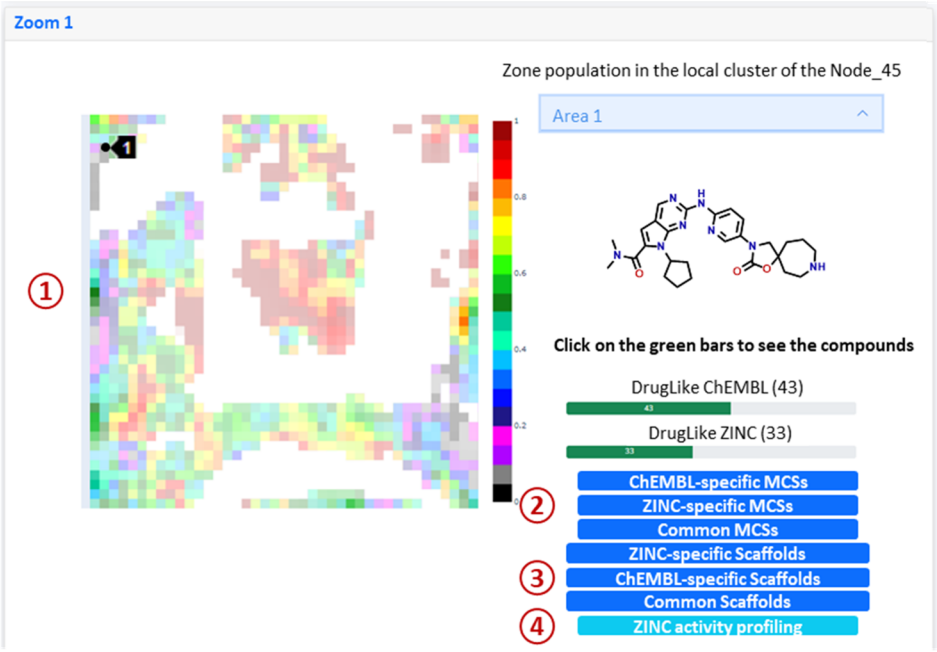


Figure 5. Zoomed level of the landscape visualization: ① zoomed map with one "tracking" compound projection; ② buttons to display the lists of common and library-specific MCSs; ③ buttons to display the lists of common and library-specific scaffolds; and ④ buttons to display results of the ZINC activity profiling with a consensus profiler.

drop-down menu enables the selection of up to five types of map coloration, i.e., landscapes ("③" in Figure 3) to be generated. Meanwhile, the Progress page reports the advancement of the procedure. In the case where provided compounds are out of the applicability domain of the map (i.e., situated too far from the GTM manifold in the initial highly dimensional descriptor space), an error message will be displayed.

After the projection, the user will be redirected to the main result page containing one of the selected landscapes (Figure 4). The colored background of the map corresponds to the library (libraries) that were selected as a basis for the landscape (in provided example - ZINC (red regions) and ChEMBL (black regions); all colors between correspond to the areas occupied by both libraries). User-defined compounds are displayed as black dots (see "④" in Figure 4).

After clicking on one of the dots, the respective compound will be shown on the right side of the map ("⑤" in Figure 4). Similar to the target-oriented chemical space navigation, several

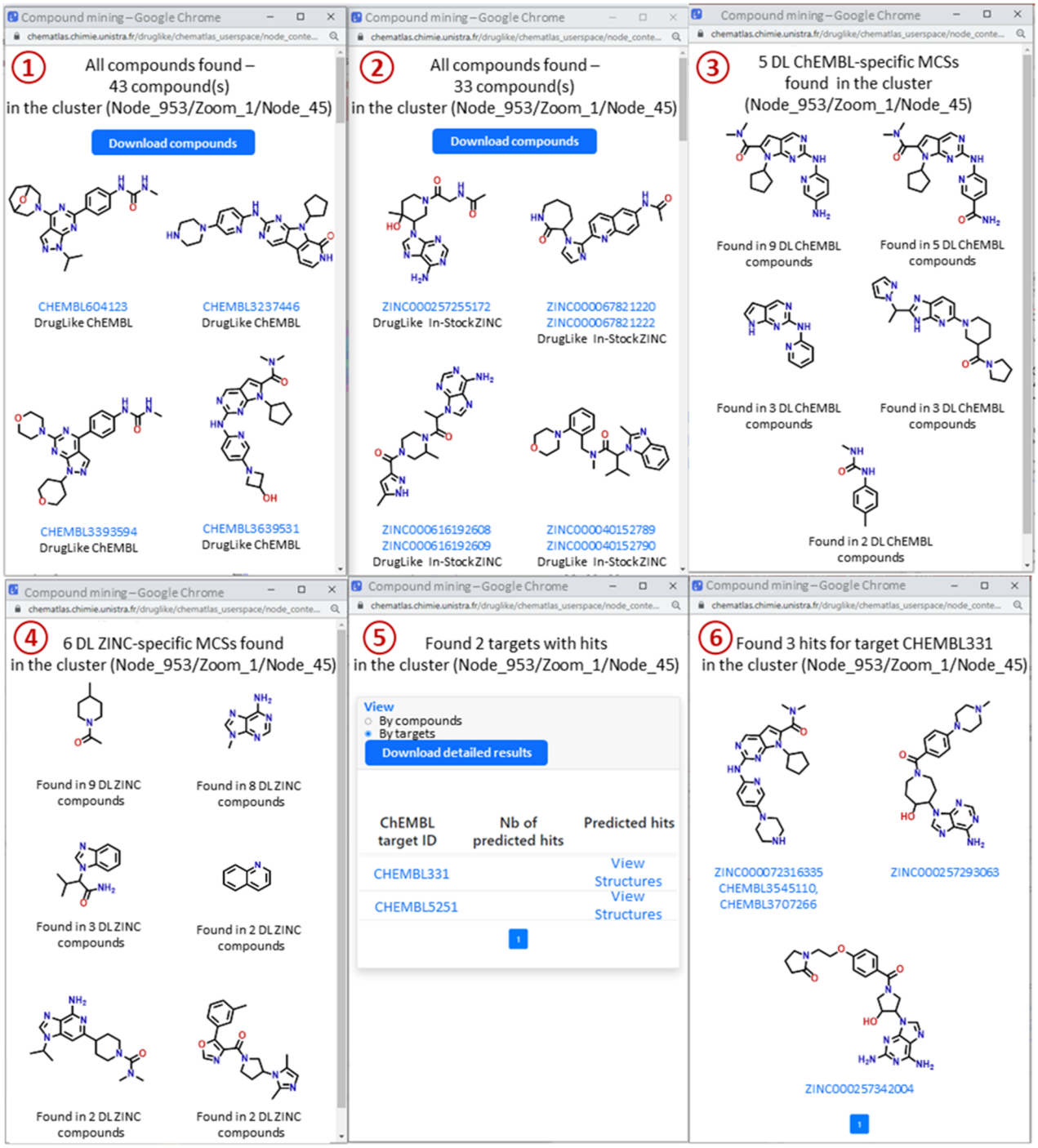


Figure 6. Pop-up windows demonstrating the content of the selected area: ① ChEMBL compounds; ② ZINC compounds; ③ ChEMBL-specific MCSs; ④ ZINC-specific MCSs; ⑤ results of the activity profiler of ZINC compounds; ⑥ predicted hits of CDK4 (the first compound is Ribociclib, recently approved by the FDA as a drug targeting CDK4).

zooming zones might be proposed to the user for investigation. Zooming zones are relevant for a tracking compound if its cumulated responsibility over the 9 nodes of the zone exceeds 0.85. Below the chemical structure, similarly to the activity visualization, two bars illustrate the proportion of drug-like ChEMBL and ZINC compounds found in the closest surrounding of the selected "tracker" ("⑥" in Figure 4). As long as the bars are yellow, corresponding compounds cannot be displayed, because there are too many of them in that map zone. In such a case, the "Zoom" button ("⑦" in Figure 4) appears, which allows one to visualize the zoomed map—the next level of navigation (Figure 5).

Once the bars become green, the closest neighbors of the selected tracking compound can be displayed for each featured library: 43 ChEMBL compounds ("①" in Figure 6) and 33 ZINC molecules ("②" in Figure 6). At the last level of zooming, MCSs ("②" in Figure 5) and scaffold analysis ("③" in Figure 5) are available. Users can retrieve library-specific and common MCSs and scaffolds characterizing selected zone. In a given

example, there are 5 ChEMBL-specific MCSs ("③" in Figure 6) and 6 ZINC-specific MCSs ("④" in Figure 6).

The consensus Activity Profiler may then predict the polypharmacological profile of the ZINC compounds residing in the selected area ("④" in Figure 5). Results of the profiling can be visualized as shown by "⑤" and "⑥" in Figure 6. In a given example, out of 33 closest ZINC neighbors of the reference compound CHEMBL3691640, 3 compounds were predicted as active against CDK4. Recall that this prediction is much more robust than the simple "hint" of CDK4 activity due to the closeness to CDK4-active trackers on the visualized map. It is a consensus prediction proving that given ZINC compounds are systematically found to reside in CDK4-active-enriched neighborhoods in a majority of the seven universal maps, each based on distinct descriptors highlighting complementary chemical information. One of those compounds is common to ZINC and ChEMBL (ZINC000072316335, CHEMBL3545110, CHEMBL370726). It is Ribociclib, which was approved by the FDA in 2017 for the treatment of breast cancer (IC_{50} against CDK4 is 10 nM).

This example demonstrates that, with the help of ChemSpace Atlas, it becomes easier to navigate chemical space in a search of potential drug candidates starting from structural analysis of the knowledge space (ChEMBL activity Navigator), and finishing with commercially available analogues search and activity profiling of compounds similar to known ligands.

Technical Details of Web Implementation. ChemSpace Atlas runs on a server version of Ubuntu 18.04[51] with Apache 2.4[52] as an open-source HTTP Web server. An Anaconda[53] installation with Python 3.6 is linked to the Apache server. All physicochemical properties, respective landscapes, and MCSs are precomputed, hierarchically organized, and stored on the dedicated server. The ChemSpace Atlas front-end is developed with jQuery,[54] which is a fast, lightweight, cross-browser, and feature-rich JavaScript library. The Bootstrap toolkit[55] is used to design the responsive interface. Chemical structures handling is done using two libraries: Epam sketcher,[56] as a web-based chemical structure editor, and OpenChemLib-js,[57] for compounds visualization in 2D in the results pages.

The ChemSpace Atlas back-end is developed using custom PHP and Python CGIs that process the data entered by the user (either list of SMILES or single compound drawn in Sketcher). Standardization is performed using ChemAxon[41] Standardizer and pK_a calculations plugins. Compounds projection followed by landscapes visualization is performed dynamically with custom Python scripts in the context of the ChemSpace tracker, using the Plotly library[58] (version 4.8).

CONCLUSIONS

Here, we report freely available ChemSpace Atlas, which is a highly polyfunctional web tool that allows navigating through the chemical space of unprecedently large libraries. Almost 40 000 hierarchically related GTMs enable intuitive navigation through hundreds of millions of compounds. The distinctive feature of the ChemSpace Atlas, compared to other online tools, is that it is not limited to a simple visualization of the similarity relationships in the chemical space but it also allows users to analyze physicochemical properties and biological activities, perform polypharmacological profiling, perform analogs search, and perform detailed structural analysis with the help of MCSs and scaffolds. ChemSpace Atlas is a "Big Data"-compatible tool: it provides at least a 10-fold increase in the size of the featured libraries, with respect to the existing tools (and is still growing), all while preserving high polyfunctionality. This is achieved by the different scales of chemical space analysis: from a global bird's eye view of the entire library to structural pattern detection in small clusters.

A user-defined compound set can be used to "track" the chemical space regions containing molecules with specific structural features. It also can be used for analogues searches. Almost 20 precomputed physicochemical properties and thousands of MCSs characterizing each zone enable a detailed analysis of featured libraries in a different context. Almost 750 biological activities from ChEMBL can also be visualized, and pharmacological profiling using a consensus of seven universal maps is available.

In the future, ChemSpace Atlas will not be limited to the navigators and libraries featured here. They are simply an initial core that can easily be updated in order to increase functionality, the scope of analyzed chemical space, or even the domain of its application. In our previous works, the basis for the DNA-encoded libraries (DELs) and Synthons Navigators were created but have not yet been implemented in the web interface. For the former, ∼2500 DELs were designed using commercially available building blocks (BBs) resulting in 2.5B DEL compounds that were compared to biologically relevant molecules from ChEMBL using the first universal map.[59] GTM-based coverage score allows one to compare each DEL to ChEMBL and choose several optimal DELs containing the maximum possible percentage of biologically relevant chemotypes. In this way, DEL Navigator will have slightly different functionality—apart from all existing features, it will also allow one to select the optimal DEL out of 2500 pregenerated libraries for the particular task, according to the coverage of the desired chemical space (e.g., all biologically relevant compounds, ligands of a selected target, etc.).

The second mentioned navigator featuring the chemical space of synthons will be based on the universal map of synthons—fragments of the organic BBs contributed to the final molecules upon chemical reaction.[36] They represent BB without the leaving groups with their position and reactive centers type (electrophilic, nucleophilic, radical, etc.) being encoded with special numeric marks on the "connecting" atoms. Synthons universal map combined with SynthI (now Synt-On) tool[60] for libraries design will allow one not only to search for already synthesized analogues of the provided compound, like it is already possible with currently implemented navigators, but also to generate libraries that can be synthesized using purchasable BBs.

Other directions of the future ChemSpace Atlas development can be the analysis and prediction of ADMETox properties and de novo compound generation, which was not considered herein. The autoencoder sequence-to-sequence neural network has already been combined with GTM in the work by Sattarov et al.[61] The incorporation of such methodology in ChemSpace Atlas will complement its usage by introducing the guided rational exploration of the novel regions of the chemical space.

DATA AND SOFTWARE AVAILABILITY

The ChemSpace Atlas tool is freely accessible at https://chematlas.chimie.unistra.fr.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00509.

Some methodological details of GTM construction and examples of the ChemSpace Atlas usage (PDF)

AUTHOR INFORMATION

Corresponding Author

Alexandre Varnek − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* orcid.org/0000-0003-1886-925X; Email: varnek@unistra.fr

Authors

Yuliana Zabolotna − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Fanny Bonachera − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Dragos Horvath − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* orcid.org/0000-0003-0173-5714

Arkadii Lin − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Gilles Marcou − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* orcid.org/0000-0003-1676-6708

Olga Klimchuk − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00509

Author Contributions

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

Funding

Yu.Z. thanks the Doctoral School of the University of Strasbourg for a Ph.D. fellowship.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* **2014**, *19* (7), 859−868.

(2) Liu, R.; Li, X.; Lam, K. S. Combinatorial chemistry in drug discovery. *Curr. Opin. Chem. Biol.* **2017**, *38*, 117−126.

(3) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 218−227.

(4) Kodadek, T. The rise, fall and reinvention of combinatorial chemistry. *Chem. Commun. (Cambridge)* **2011**, *47* (35), 9757−63.

(5) van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 644−651.

(6) Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55* (9), 1824−1835.

(7) Patel, H.; Ihlenfeldt, W. D.; Judson, P. N.; Moroz, Y. S.; Pevzner, Y.; Peach, M. L.; Delannee, V.; Tarasova, N. I.; Nicklaus, M. C. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* **2020**, *7* (1), 384.

(8) Humbeck, L.; Weigang, S.; Schafer, T.; Mutzel, P.; Koch, O. CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein-Protein Interaction Modulators. *ChemMedChem.* **2018**, *13* (6), 532−539.

(9) Shivanyuk, A.; Ryabukhin, S. V.; Bogolubsky, A. V.; Mykytenko, D. M.; Chupryna, A. A.; Heilman, W.; Kostyuk, A. N.; Tolmachev, A. A. Enamine REAL database: making chemical diversity real. *Chim. Oggi-Chem. Today* **2007**, *25*, 58−59.

(10) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681.

(11) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065−6073.

(12) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56* (7), 1253−1266.

(13) Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A. Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb. Sci.* **2012**, *14* (11), 579−589.

(14) Lu, J.; Carlson, H. A. ChemTreeMap: An interactive map of biochemical similarity in molecular datasets. *Bioinformatics* **2016**, *32* (23), 3584−3592.

(15) Villoutreix, B. O.; Lagorce, D.; Labbé, C. M.; Sperandio, O.; Miteva, M. A. One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discovery Today* **2013**, *18* (21), 1081−1089.

(16) Awale, M.; van Deursen, R.; Reymond, J.-L. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013**, *53* (2), 509−518.

(17) Gütlein, M.; Karwath, A.; Kramer, S. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. *J. Cheminform.* **2012**, *4* (1), 7.

(18) Janssen, A. P. A.; Grimm, S. H.; Wijdeven, R. H. M.; Lenselink, E. B.; Neefjes, J.; van Boeckel, C. A. A.; van Westen, G. J. P.; van der Stelt, M. Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome−Inhibitor Interaction Landscapes. *J. Chem. Inf. Model.* **2019**, *59* (3), 1221−1229.

(19) Carbó-Dorca, R. About the concept of Chemical Space: a concerned reflection on some trends of modern scientific thought within theoretical chemical lore. *J. Math. Chem.* **2013**, *51* (2), 413−419.

(20) González-Medina, M.; Medina-Franco, J. L. Platform for Unified Molecular Analysis: PUMA. *J. Chem. Inf. Model.* **2017**, *57* (8), 1735−1740.

(21) Borrel, A.; Kleinstreuer, N. C.; Fourches, D. Exploring drug space with ChemMaps.com. *Bioinformatics* **2018**, *34* (21), 3773−3775.

(22) Wenderski, T. A.; Stratton, C. F.; Bauer, R. A.; Kopp, F.; Tan, D. S. Principal component analysis as a tool for library design: A case study investigating natural products, brand-name drugs, natural product-like libraries, and drug-like libraries. *Methods Mol. Biol.* **2015**, *1263*, 225−42.

(23) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Machine Learning Res.* **2008**, *9*, 2579−2605.

(24) Probst, D.; Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform* **2020**, *12* (1), 12.

(25) Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2018**, *34* (8), 1433−1435.

(26) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10* (1), 215−234.

(27) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43* (1), 59−69.

(28) 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* (1), 564−572.

(29) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **2015**, *34* (6−7), 348−56.

(30) 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* (3−4), 301−12.
(31) 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* (3), 331−343.
(32) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided Mol. Des.* **2015**, *29* (12), 1087−108.
(33) Tino, P.; Nabney, I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE PAMI* **2002**, *24* (5), 639−656.
(34) Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. Diversifying chemical libraries with generative topographic mapping. *J. Comput. Aided Mol. Des.* **2019**, *34*, 805−815.
(35) Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: a New Look at the Natural Product Chemical Space. *Mol. Inform.* **2021**, *40* (9), 2100068.
(36) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath, D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek, A. A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. *J. Chem. Inf. Model.* **2021**, *62*, 2171−2185.
(37) Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A. Chemography: Searching for Hidden Treasures. *J. Chem. Inf. Model.* **2021**, *61* (1), 179−188.
(38) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 2012, *52* (7), 1757−1768.
(39) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47* (D1), D930−D940.
(40) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminform* **2021**, *13* (1), 2.
(41) *ChemAxon. JChem.*, Version 20.8.3; ChemAxon, Ltd.: Budapest, Hungary, 2020.
(42) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, *29* (12), 855−68.
(43) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324−37.
(44) Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform* **2020**, *12* (1), 20.
(45) Chen, C. Y. TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* 2011, *6* (1), e15939.
(46) Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M. T.; Piperno, A.; Rescifina, A. Putative Inhibitors of SARS-CoV-2 Main Protease from A Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Mar. Drugs* **2020**, *18* (4), 225.
(47) Zeng, X.; Zhang, P.; Wang, Y.; Qin, C.; Chen, S.; He, W.; Tao, L.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res.* **2019**, *47* (D1), D1118−D1127.
(48) Banerjee, P.; Erehman, J.; Gohlke, B. O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II–a database of natural products. *Nucleic Acids Res.* **2015**, *43* (Database issue), D935−D939.
(49) Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges* **2014**, *5* (2), 450−472.
(50) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* **2011**, *51* (9), 2174−85.
(51) Sobell, M. G. *A Practical Guide to Ubuntu Linux*; Pearson Education, 2015.
(52) Apache2 Software Distribution (2021) Apache2 Documentation. License from https://www.apache.org/licenses/LICENSE-2.0. Access date: May 2021.
(53) Anaconda Software Distribution. (2021). *Anaconda Documentation*. Retrieved from https://docs.anaconda.com/. Access date: May 2021.
(54) jQuery Software Distribution (2021). *jQuery Documentation*. Retrieved from https://jquery.com. Access date: May 2021.
(55) Bootstrap Software Distribution (2021). *Bootstrap Documentation*. Retrieved from https://getbootstrap.com. Access date: May 2021.
(56) LifeSciences Unit of EPAM Systems. *EPAM Documentation*. Retrieved from https://lifescience.opensource.epam.com/ketcher/. Accessed May 2021.
(57) JavaScript port of the OpenChemLib Java library. *OpenChemLib-js*. Retrieved from https://github.com/cheminfo/openchemlib-js. Accessed May 2021.
(58) Plotly Technologies, Inc., Montreal, QC: (2015). *Collaborative data science*. Retrieved from https://plot.ly. Accessed May 2021.
(59) Zabolotna, Y.; Pikalyova, R.; Volochnyuk, D. M.; Horvath, D.; Marcou, G.; Varnek, A. Exploration of the chemical space of DNA-encoded libraries. *ChemRxiv. Cambridge: Cambridge Open Engage* **2021**, DOI: 10.33774/chemrxiv-2021-dpbdx.
(60) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A. SynthI: A New Open-Source Tool for Synthon-Based Library Design. *J. Chem. Inf. Model.* **2021**, *62*, 2151−2163.
(61) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019**, *59* (3), 1182−1196.