DenovoProfiling: a webserver for de novo generated molecule library profiling

Zhihong Liu
Guangdong Institute of Microbiology https://orcid.org/0000-0001-8370-2419
Jiewen Du
Beijing Jingpai Technology Co., Ltd. https://orcid.org/0000-0003-4149-8629
Bingdong Liu
Guangdong Institute of Microbiology https://orcid.org/0000-0002-6007-7902
Zongbin Cui
Guangdong Institute of Microbiology https://orcid.org/0000-0002-3443-3168
Jiansong Fang
Guangzhou University of Chinese Medicine https://orcid.org/0000-0002-6998-5384
Liwei Xie (xielw@gdim.cn)
Guangdong Institute of Microbiology https://orcid.org/0000-0002-4747-1753

Software

Keywords: Generated Molecule, Webserver, DenovoProfiling

Posted Date: February 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-142605/v2

Abstract

With the advances of deep learning techniques, various architectures for molecular generation have been proposed for de novo drug design. Successful cases from academia and industrial demonstrated that the deep learning-based de novo molecular design could efficiently accelerate the drug discovery process. The flourish of the de novo molecular generation methods and applications created a great demand for the visualization and functional profiling for the de novo generated molecules. The rising of publicly available chemogenomic databases lays good foundations and creates good opportunities for comprehensive profiling of the de novo library. In this paper, we present DenovoProfiling, a webserver dedicated to de novo library visualization and functional profiling. Currently, DenovoProfiling contains six modules: (1) identification & visualization, (2) chemical space, (3) scaffold analysis, (4) molecular alignment, (5) drugs mapping, and (6) target & pathway. DenovoProfiling could provide structural identification, chemical space exploration, drug mapping, and target & pathway information. The comprehensive annotated information could give users a clear picture of their de novo library and could guide the further selection of candidates for synthesis and biological confirmation. DenovoProfiling is freely available at http://denovoprofiling.xielab.net.

Introduction

The main objective of drug discovery is to identify a molecule with desired biological properties[1]. Primarily, high throughput screening (HTS) techniques allow a large size of chemical library testing[2,3]. However, HTS is expensive and with low hit rates, only limited to large pharmaceutical companies. Computational based virtual screening methods are used to reduce the size of testing molecules. Various ligand-based[4,5] and structure-based[6] virtual screening methods have been proposed. However, the cost and time consuming for developing a new drug are still increasing[7].

De novo drug design is one of the most promising and scalable approaches to address this issue, and in particular, with the advances of deep learning techniques[8–10]. In the early stage, evolutionary algorithms were used for de novo molecular generation[11], which is commonly based on the combinations of molecular fragments derived from a drug-like library. Over the past years, artificial intelligence algorithms, like deep learning, reinforcement learning, and transfer learning are proposed in the molecule generation field, inspired by the wide applications of those methods to generate text, images, video, and music[12,13].

Recently, several architectures for molecular generation, such as recurrent neural networks (RNN)[1,14,15], variational autoencoders (VAE)[16], and generative adversarial networks (GANs)[17] have been developed and proven successful in generating target-focus molecule library. Furthermore, scaffold-constrained molecular generation methods[18,19] were developed for lead optimization. Yang et al. also developed linker constraints molecular generation methods using deep conditional transformer neural networks for fragment-based drug design (FBDD)[20]. Zhavoronkov et al. developed a deep generative model with reinforcement learning and discovered potent discoidin domain receptor 1 inhibitors in 21 days, and

further animal experimental results in one lead candidate demonstrated favorable pharmacokinetics[21]. This important case illustrated the utility of a deep generative model for the rapid design of compounds with synthetically feasibility, and bioactivity for the target of interest. Yang et al. developed a generative model using long short-term memory (LSTM) neural network and generated a focused library containing 672 valid molecules[22]. After filtering with various criteria, synthesis, and bioactivity testing, they identified a highly potent inhibitor against p300 with IC50 of 10 nM. These successful cases demonstrate that the deep learning-based de novo molecular design could accelerate the drug discovery process.

The flourish of the de novo molecular generation methods and applications created a great demand for the visualization and functional profiling for the generated molecules. Generally, the generative models could generate a large chemical library based on sampling criteria and could output with various formats. The following issues, particularly for medicinal chemists, are to visualize, analyze, and select the candidates among the generated molecules. Owing to the development of combinatorial chemistry and high-throughput screening technologies, chemical structures and bioactivity data have rapidly accumulated in the past years and are becoming available in public repositories[23,24]. There are various well-established cheminformatics and bioinformatics databases available for drug discovery, which provide comprehensive information for bioactive compounds, drugs, targets, pathways, and disease, like and PDB database[25], PubChem[26], DrugBank[27], ChEMBL[28], and BindingDB[29]. The rising of publicly available databases creates good opportunities for comprehensive profiling of the de novo library.

Dealing with chemical libraries is a common practice in drug discovery. Thus, various cheminformatics tools have been developed for chemical library

processing and data analysis. Well-known tools for dealing with chemical library are ChemicalToolbox[30], DataWarrior[31], WebMolCS[32], ChemMine[33], CART[34], MONA[35], and CSgator[36]. Those tools mainly focused on specific functionality, such as large library visualization, structure search, or clustering analysis. Even more, some tools are desktop applications, which limited the application. Web-based tools dedicated to de novo generated molecule profiling are rare.

In this work, we present the DenovoProfiling, a webserver for de novo generated molecule library profiling. We aim to provide a user-friendly public webserver to support the structure and chemical space visualization, scaffold analysis, molecular alignment, drug profiling, target & pathway profiling. We integrated cheminformatics tools and databases to provide comprehensive annotations for the de novo generated molecules. We believe that DenovoProfiling could be an efficient tool for the user to capture the knowledge of de novo generated molecules quickly. DenovoProfiling is freely available at http://denovoprofiling.xielab.net.

Methods

Framework

The framework of DenovoProfiling was outlined in **Fig. 1**. We integrated the well-known public database PubChem, ChEMBL, DrugBank, and employed open-source cheminformatics toolkits, and other tools to provide comprehensive information for user-submitted de novo chemical library. The profiling process is fully automatic, which user only submit its de novo library files with multiple formats are supported. DenovoProfiling contains 6 modules: identification & visualization, chemical space, scaffold analysis, molecular alignment, target & pathways, and drug mapping.

Supported formats

Four widely used chemical formats are supported in DenovoProfiling: SDF (structure-data file), SMILES (simplified molecular-input line-entry system), InChI (International Chemical Identifier), and CDX (ChemDraw Exchange). All those formats files can be uploaded or the file contents can be pasted and submitted to the web server, except for the binary CDX format, which cannot be pasted. The Open Babel[37] program was used for chemical file format conversion.

Modules

Currently, DenovoProfiling provides 6 profiling modules. Each module is functional individually and the user could select the module of interest. The implementations for each module are described as follows.

Identification & Visualization

Identification & Visualization module aims to check whether the de novo structures are already existing and visualize the de novo chemical structures. The submitted de novo molecules were converted into InChIKeys using Open Babel[37]. Subsequently, the InChIKeys were submitted to the PubChem using PubChemPy (https://pubchempy.readthedocs.io), a python package for interacting with PubChem. PubChem is the world's largest collection of freely accessible chemical information with over 109 million compounds[26]. The PubChem compound IDs (CID) were retrieved when de novo molecules were matched. ChemDoodle Web component, a light-weight JavaScript/HTML5 toolkit for chemical graphics, developed by iChemLabs was used for structure visualization[38]. For non-SDF format, Open Babel was used to generate 2D structures for structure visualization. Meanwhile, the drug-like descriptors including molecular weight (MW), ALogP, number of hydrogen bond acceptors (HBA), number of hydrogen bond donors (HBD), number of rotatable bonds (RotBonds), and topological polar surface area (TPSA) were calculated using PaDEL[39] and plotted using Radar Chart.

Chemical Space

Chemical space visualization is an efficient way to know the structural similarity or properties similarity of the corresponding molecules through the closeness of the points in this chemical space. Each molecule was defined by a set of numerical descriptors or fingerprints and a set of all molecules corresponded to the points in the same coordinate-based space. Three important approaches: similarity

maps, principal components analysis (PCA), and drug-like properties distribution were used in DenovoProfiling. The chemical similarity heatmap was generated and interactive, in which the user could move or click the cells of the similarity matrix, and the corresponding structures are visualized beside. The principal component analysis (PCA) was used to visualize the chemical space based on PubChem fingerprints. The frequency distribution histogram of drug-like descriptors mentioned in Identification & Visualization was also plotted.

Scaffold Analysis

Scaffold is an important concept for medicinal chemistry when measuring the novelty of a molecule. Generally, for medicinal chemists, a scaffold defines the core structure essential for pharmacological activity, which is data set dependent and could vary in the different target systems. Bemis and Murcko proposed the Bemis-Murcko (BM) scaffold framework[40], an objective, invariant, and data set independent scaffold representation method, which is widely used in cheminformatics studies. The BM scaffold method dissects molecules into ring systems, linkers, side-chain atoms, and the framework. Scaffold-based classification approach (SCA)[41], an atomic framework of BM scaffold[40] was used here and widely applied in cheminformatics studies[42] and drug discovery projects[43,44]. The scaffolds were generated for de novo molecules and the number of molecules for each scaffold was calculated and plotted.

Molecular Alignment

When a de novo molecular library was generated, a straightforward point is to align the focused library, in particular scaffold focused library, to compare the structures of molecules. The molecular alignment module is designed to satisfy this demand. Weighted Gaussian Algorithm (WEGA)[45], developed by the previous lab, was used here for molecular alignment. WEGA is an efficient and accurate way for molecular alignment and calculating shape similarity. The shape and pharmacophore combined approach in WEGA was used in DenovoProfiling. Both shape and pharmacophore features are considered in the alignment process. The first molecule was used as the template for alignment. After alignment, the user could select the molecules of interest to see or download the alignment results. The three-dimensional conformation alignment was rendered using 3Dmol.js[46].

Drugs Mapping

A similar chemical structure may have similar property or activity. Drugs Mapping module was designed to fast retrieve the drugs which are chemically similar to the de novo molecules. The structures and names of drugs were derived from DrugBank[27]. Inorganic molecules, salts, and duplicates were removed using Open Babel. The similarities between the de novo molecule against the drugs were calculated. 2D similarity calculations are based on the atom center fragment[47]. Tanimoto coefficient was used as a metric to quantify the similarity between two molecules. The similar drugs with similarity over 0.5 against de novo molecules were preserved.

Target & Pathway

The bioactivities of targets and corresponding ligands were derived from the ChEMBL database. Duplicates were removed and compounds with multiple binding affinity data, the most potent with minimal value were chosen. After data processing, ligand structures, target, bioactivity data, and corresponding references were obtained and saved in the MySQL database. The target proteins and their bioactivity data for submitted de novo molecules were queried by using the generated InChIKeys. The retrieved results were summarized in an interactive table and a compound target network using a dynamic, browser-based visualization library(vis.js). The targeted proteins were further functional enriched using python client of bioinformatics web service DAVID[48]. The UniProt IDs of the targets retrieved from ChEMBL were used as input for functional enrichment. The enriched KEGG pathways were provided and could be downloaded through an interactive table. For each pathway, the pathway term, gene count, percent, P-value, fold enrichment, Benjamini value, and false discovery rate (FDR) are provided.

Web Server Implementation

DenovoProfiling is a publicly accessible platform, which can be accessed through a web browser using the browser server framework. The D3 library of JavaScript (d3js.org/) is used to illustrate the scatterplots, radical plot, and heatmaps. Storage and management of the submitted job data are implemented by MySQL. The back-end server was developed by the Golang language. The tools used for constructing the DenovoProfiling are summarized in the online tutorial of the help page.

Results And Discussion

To test the functionality of DenovoProfiling, we collected or generated 3 datasets for different purposes. The first dataset (Drug Dataset) contains 60 drug molecules randomly selected from DrugBank to check the corrected information retrieved from different profiling database in DenovoProfiling. This dataset aims to verify the utility of identification & visualization, chemical space, drugs mapping, target & pathway. The second dataset (Random Dataset) contains 500 molecules randomly generated using REINVENT [14], an RNN architecture pre-trained on more than one million bioactive structures from ChEMBL. This dataset is mainly dedicated for verify the utility of the We have developed an interface of REINVENT as a de novo module in our DeepScreening[49]. This dataset aims to verify the utility of identification & visualization, chemical space, scaffold analysis, drugs mapping, target & pathway. The third dataset (Focused Dataset) contains 50 scaffold-focused de novo molecules based on a scaffold constrained molecular generation approach[18] for verifying the molecular alignment module.

Table 1 The datasets for testing the functionality of DenovoProfiling.

Index	Dataset	Molecules	Source
1	Drug Dataset	60	drug molecules randomly selected from DrugBank[27]
2	Random Dataset	500	de novo molecules randomly generated using REINVENT[14]
3	Focused Dataset	50	de novo molecules based on a scaffold constrained molecular generation[18]

Structure identification and visualization of de novo library

For a de novo generated library, the first real request is to visualize the chemical structures and know the structure novelty. The Identification & Visualization module was designed to satisfy this demand. Using the Random dataset as input, the snapshot of this module was shown in Fig. 2. The user could browse the structures with mapped PubChem Compound ID (CID). The properties including molecular weight, LogP, HBA, HBD, number of rotatable bonds, TPSA are given by clicking the upright plus button. The CID is provided at the bottom right and linked the PubChem which provides more detailed compound information.

Chemical space exploration of de novo library

Chemical structures data are sophisticated, in particular for the de novo generated molecular library, and expert knowledge is highly required[50]. In this module, similarity maps, principal components analysis (PCA), and drug-like properties distributions are provided in DenovoProfiling. Using Drug Dataset as input, the snapshot of similarity heatmap was shown in Fig. 3. The generated similarity heatmap is interactive, in which the user could move or click the mouse to the target cell, and the corresponding structures, name of molecules, and the similarity value are visualized. Meanwhile, the distribution of the similarities was also plotted. Using the Random Dataset (500 molecules) as input, the snapshots of PCA results were shown in Fig. 4. Each point represents a molecule, and the user could move the mouse to the point, and the corresponding structure and molecule name returned immediately. Meanwhile, the distribution of drug-like properties was also plotted, as shown in Fig. 5. The interactive chemical space exploration could help users capture the structure relations, descriptors landscapes of the de novo library conveniently.

Scaffold Analysis of de novo library

The scaffold is an important concept in drug discovery and medicinal chemistry. Medicinal chemists are seeking chemical with novel scaffolds for a specific biological target[42]. Using the Random Dataset as input, the snapshots of the scaffold analysis were shown in Fig. 6, the complexity and cyclicity of the scaffolds and statistics of each scaffold were interactive illustrated with scatter plot and histogram plot (Fig. 6). As shown in Fig. 7, the structures of scaffolds and their number of molecules were illustrated in

the grid table. The members of molecules for each scaffold could be browsed by clicking the upright plus button.

Molecular alignment of the scaffold-constrained library

Generally, medicinal chemist starts from drug target, and attempt to generate a target-focused library or a scaffold-constraint library for structural optimization. In this case, shape or pharmacophore features based molecular alignment is a good way to compare the difference of the target-focused or scaffold-constraint de novo library. The shape and pharmacophore combined approach in WEGA was used in DenovoProfiling for molecular alignment. Using Focused Dataset as input, as shown in Fig. 8, DenovoProfiling correctly aligns all the structures to the first structure of the library. The important features, carboxyl, hydroxyl, and benzene ring are corrected recognized, and overlayed. The user could browse the alignment results, and select the molecules of interest to see the alignment result. Stick and line render methods are supported. The user also cloud click the download button to download all the alignment results for local analysis.

Drugs mapping of de novo library

De novo generated libraries usually are randomly and cover a larger chemical space. Though, Identification & Visualization, as mentioned above, identify the structures which have been reported. The structural similar drugs against the de novo library are the interest of medicinal chemists. They could fast capture the novelty and pharmacological activities of the de novo compounds when compared with the drugs library. Firstly, we used the Drug Dataset as input, DenovoProfiling calculated the similarity between the submitted Drug Dataset and the drug library. The similar drugs for each submitted molecule were returned. As shown in Fig. 9A, the drugs were corrected recognized and identified 363 similar drugs with similarity values over 0.5. Then, using Random Dataset, DenovoProfiling identified 423 similar drugs with similarity values over 0.5 (Fig. 9B). The grid view (Fig. 9) and the table view (Fig. 10) are provided. For the randomly sampled 500 de novo compounds, as shown in Fig. 10, 3 compounds with a maximal drug similarity over 0.9, and their DrugBank ID are also provided and linked to the original database. Details of this drug information could be obtained directly.

Target and pathway profiling for de novo library.

The modules we described before are structural annotations for the de novo library. Functional profiling is another important part of user concerns for de novo library proofing. Firstly, we used Drug Dataset as input, DenovoProfiling retrieved 365 bioactivity data for the 60 drugs. As shown in Fig. 11, the bioactivity data such as Ki, Kd, IC50, and EC50, and corresponding references are extracted. All those results can be analyzed via a user-friendly table view (Fig. 11). Those results are also can be downloaded for local analysis. The compound target relations were further illustrated using a compound target network (Fig. 11). We further used the Random Dataset as input, DenovoProfiling retrieved 14 bioactivity data for the 500 de novo molecules(Fig. 12). The targets are further enriched to pathways and KEGG pathways are

summarized in table (Fig. 12). DenovoProfiling enriched 115 pathways and 5 pathways for the Drug Dataset and Random Dataset, respectively(Fig. 13).

Conclusion

De novo drug design is one of the most promising and scalable approaches to accelerate the drug discovery process. Deep learning-based de novo molecular generation has shown powerful performance in generating de novo target-focused or property-focused libraries. Fast profiling the de novo generated molecules becomes a practical issue in the de novo drug design. To address this issue, we developed DenovoProfiling, a web-based profiling server for de novo generated molecules. DenovoProfiling supports structure identification, structural visualization, chemical space exploration, scaffold analysis, molecular alignment, drugs profiling, and target & pathway profiling. These functional modules provide structural and functional annotations for de novo molecules generated from various methods. We believe this web-based tool could facilitate de novo drug design and accelerate drug discovery.

Declarations

Availability and data and materials

The web platform can be accessible at http://denovoprofiling.xielab.net. The source code for the webserver was deposited in GitHub at https://github.com/nanomolar/DenovoProfiling.

Competing interests

The authors do not have any possible conflicts of interest.

Funding

This work was funded by the National Natural Science Foundation of China (Grant No. 81703416, 81900797, 82072436), GDAS Project of Science and Technology Development (Grant No. 2019GDASYL-0103009, 2018GDASCX-0102, 2021GDASYL-20210102003), Guangdong Basic and Applied Basic Research Foundation (2020B1515020046).

Authors' contributions

LX and JF designed the study. ZL, JD, and BL implemented the web site. The manuscript was written through contributions of all authors and revised by LX and JF. All authors read and approved the final manuscript.

Acknowledgments

We also thank professor Jun Xu from Sun Yat-sen University for providing useful suggestions.

References

[1] Segler, M. H. S.; Kogej, T.; Tyrchan, C.; et al. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* 2018, *4* (1), 120–131

[2] Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* 2009, *9* (5), 580–588

[3] Wildey, M. J.; Haunso, A.; Tudor, M.; et al. High-Throughput Screening. In *Annual Reports in Medicinal Chemistry*; 2017; pp 149–195

[4] Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discov. Today* 2011, *16* (9–10), 372–376

[5] Zheng, M.; Liu, Z.; Yan, X.; et al. LBVS: An Online Platform for Ligand-Based Virtual Screening Using Publicly Accessible Databases. *Mol. Divers.* 2014, *18* (4), 829–840

[6] Slater, O.; Kontoyianni, M. The Compromise of Virtual Screening and Its Impact on Drug Discovery. *Expert Opin. Drug Discov.* 2019, *14* (7), 619–637

[7] Mullard, A. New Drugs Cost US$2.6 Billion to Develop. *Nat. Rev. Drug Discov.* 2014, *13* (12), 877–877

[8] Yang, X.; Wang, Y.; Byrne, R.; et al. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* 2019, *119* (18), 10520–10594

[9] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* 2015, *521* (7553), 436–444

[10] Chen, H.; Engkvist, O.; Wang, Y.; et al. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* 2018, *23* (6), 1241–1250

[11] Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for de Novo Drug Design – A Survey. *Appl. Soft Comput.* 2015, *27*, 543–552

[12] Schneider, G.; Clark, D. E. Automated De Novo Drug Design: Are We Nearly There Yet? *Angew. Chemie Int. Ed.* 2019, *58* (32), 10792–10803

[13] Bian, Y.; Xie, X.-Q. Generative Chemistry: Drug Discovery with Deep Learning Generative Models. 2020, *5276*, 1–29

[14] Olivecrona, M.; Blaschke, T.; Engkvist, O.; et al. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* 2017, *9* (1), 48

[15] Blaschke, T.; Arús-Pous, J.; Chen, H.; et al. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* 2020

[16] Blaschke, T.; Olivecrona, M.; Engkvist, O.; et al. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* 2018, *37* (1–2), 1700123

[17] Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; et al. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv* 2017

[18] Langevin, M.; Minoux, H.; Levesque, M.; et al. Scaffold-Constrained Molecular Generation. *J. Chem. Inf. Model.* 2020, acs.jcim.0c01015

[19] Li, Y.; Hu, J.; Wang, Y.; et al. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *J. Chem. Inf. Model.* 2020, *60* (1), 77–91

[20] Yang, Y.; Zheng, S.; Su, S.; et al. SyntaLinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chem. Sci.* 2020, *11* (31), 8312–8322

[21] Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; et al. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* 2019, *37* (9), 1038–1040

[22] Yang, Y.; Zhang, R.; Li, Z.; et al. Discovery of Highly Potent, Selective, and Orally Efficacious P300/CBP Histone Acetyltransferases Inhibitors. *J. Med. Chem.* 2020, *63* (3), 1337–1360

[23] Lipinski, C. A.; Litterman, N. K.; Southan, C.; et al. Parallel Worlds of Public and Commercial Bioactive Chemistry Data. *J. Med. Chem.* 2015, *58* (5), 2068–2076

[24] Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *J. Med. Chem.* 2012, *55* (16), 6987–7002

[25] Burley, S. K.; Berman, H. M.; Bhikadiya, C.; et al. RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy. *Nucleic Acids Res.* 2019, *47* (D1), D464–D474

[26] Kim, S.; Chen, J.; Cheng, T.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* 2019, *47* (D1), D1102–D1109

[27] Law, V.; Knox, C.; Djoumbou, Y.; et al. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* 2014, *42* (D1), D1091–D1097

[28] Mendez, D.; Gaulton, A.; Bento, A. P.; et al. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* 2019, *47* (D1), D930–D940

[29] Liu, T.; Lin, Y.; Wen, X.; et al. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* 2007, *35* (Database), D198–D201

[30] Bray, S. A.; Lucas, X.; Kumar, A.; et al. The ChemicalToolbox: Reproducible, User-Friendly Cheminformatics Analysis on the Galaxy Platform. *J. Cheminform.* 2020, *12* (1), 40

[31] Sander, T.; Freyss, J.; Von Korff, M.; et al. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* 2015, *55* (2), 460–473

[32] Awale, M.; Probst, D.; Reymond, J. L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. *J. Chem. Inf. Model.* 2017, *57* (4), 643–649

[33] Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules. *Nucleic Acids Res.* 2011, *39* (SUPPL. 2), 486–491

[34] Deghou, S.; Zeller, G.; Iskar, M.; et al. CART - A Chemical Annotation Retrieval Toolkit. *Bioinformatics* 2016, *32* (18), 2869–2871

[35] Hilbig, M.; Rarey, M. MONA 2: A Light Cheminformatics Platform for Interactive Compound Library Processing. *J. Chem. Inf. Model.* 2015, *55* (10), 2071–2078

[36] Park, S.; Kwon, Y.; Jung, H.; et al. CSgator: An Integrated Web Platform for Compound Set Analysis. *J. Cheminform.* 2019, *11* (1), 17

[37] O'Boyle, N. M.; Banck, M.; James, C. A.; et al. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* 2011, *3* (1), 33

[38] Burger, M. C. ChemDoodle Web Components: HTML5 Toolkit for Chemical Graphics, Interfaces, and Informatics. *J. Cheminform.* 2015, *7* (1), 35

[39] Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* 2011, *32* (7), 1466–1474

[40] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* 1996, *39* (15), 2887–2893

[41] Xu, J. A New Approach to Finding Natural Chemical Structure Classes. *J. Med. Chem.* 2002, *45* (24), 5311–5320

[42] Liu, Z.; Ding, P.; Yan, X.; et al. ASDB: A Resource for Probing Protein Functions with Small Molecules. *Bioinformatics* 2016, *32* (11), 1752–1754

[43] Zhao, C.; Huang, D.; Li, R.; et al. Identifying Novel Anti-Osteoporosis Leads with a Chemotype-Assembly Approach. *J. Med. Chem.* 2019, *62* (12), 5885–5900

[44] Guo, Q.; Zhang, H.; Deng, Y.; et al. Ligand- and Structural-Based Discovery of Potential Small Molecules That Target the Colchicine Site of Tubulin for Cancer Treatment. *Eur. J. Med. Chem.* 2020

[45] Yan, X.; Li, J.; Liu, Z.; et al. Enhancing Molecular Shape Comparison by Weighted Gaussian Functions. *J. Chem. Inf. Model.* 2013, *53* (8), 1967–1978

[46] Rego, N.; Koes, D. 3Dmol.Js: Molecular Visualization with WebGL. *Bioinformatics* 2015, *31* (8), 1322–1324

[47] Yan, X.; Gu, Q.; Lu, F.; et al. GSA: A GPU-Accelerated Structure Similarity Algorithm and Its Application in Progressive Virtual Screening. *Mol. Divers.* 2012, *16* (4), 759–769

[48] Huang, D. W.; Sherman, B. T.; Tan, Q.; et al. DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists. *Nucleic Acids Res.* 2007, *35* (suppl_2), W169–W175

[49] Liu, Z.; Du, J.; Fang, J.; et al. DeepScreening: A Deep Learning-Based Screening Web Server for Accelerating Drug Discovery. *Database* 2019, *2019*, 1–11

[50] Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; et al. Progress in Visual Representations of Chemical Space. *Expert Opin. Drug Discov.* 2015, *10* (9), 959–973

Figures

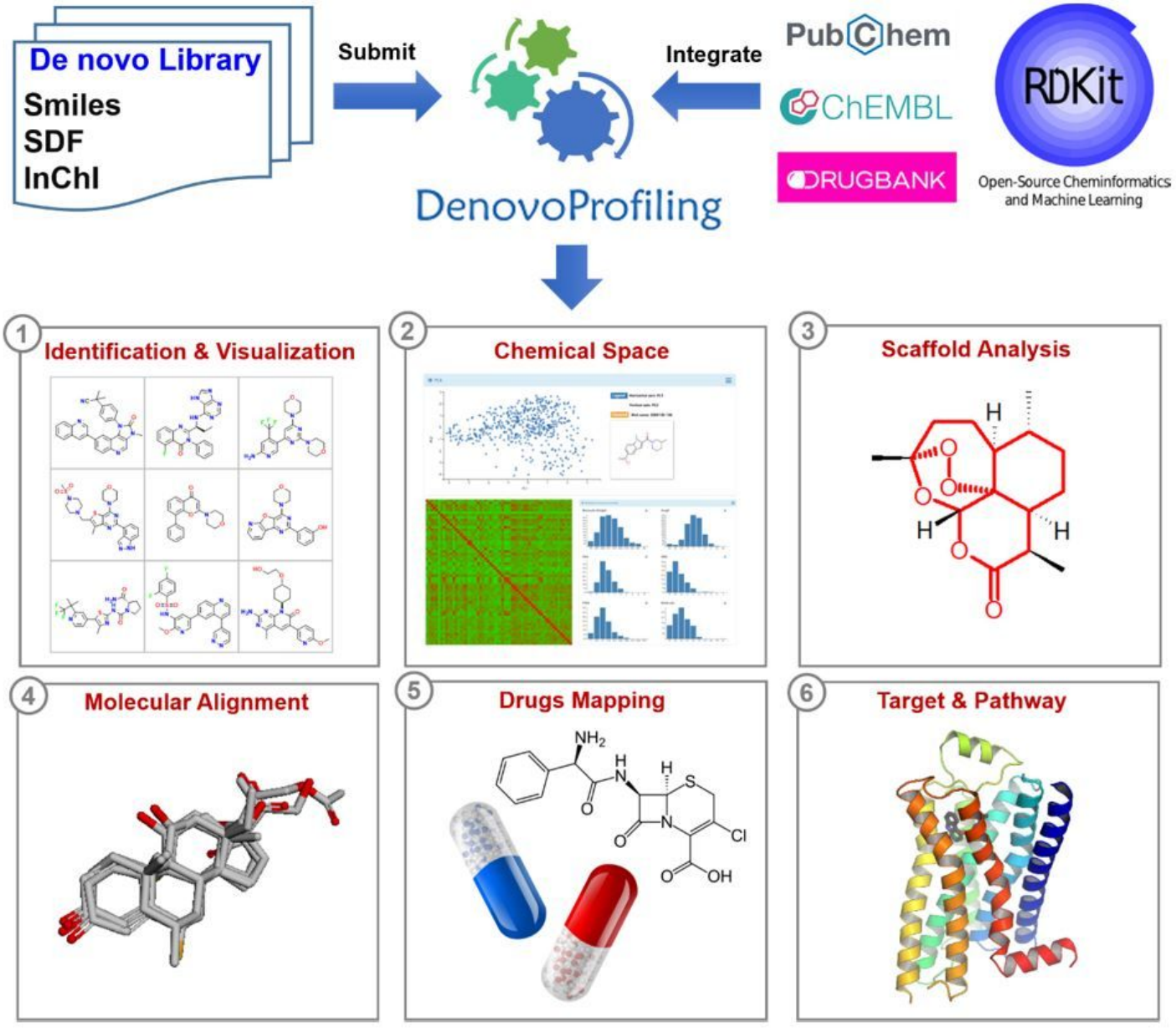


Figure 1

The framework of DenovoProfiling web platform.

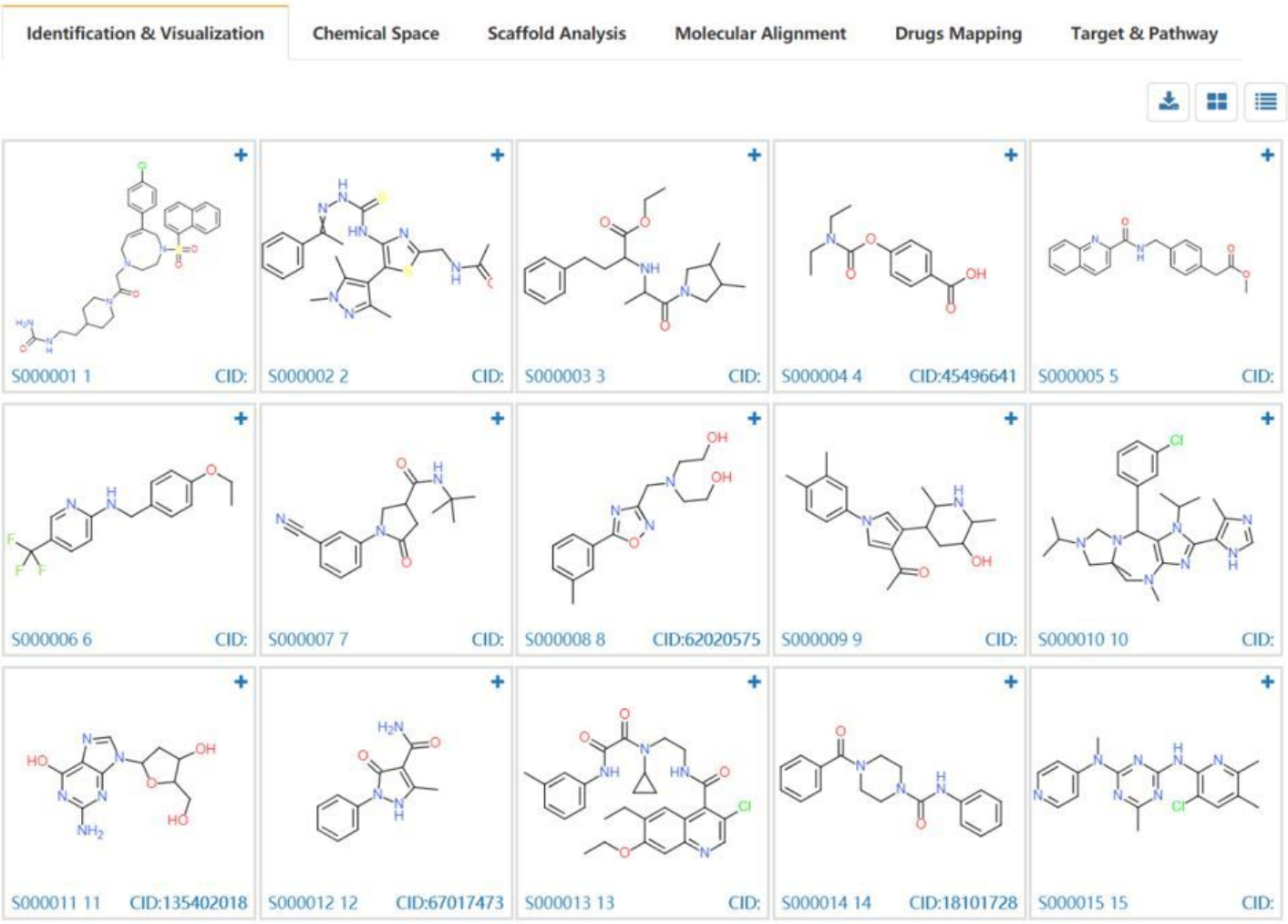


Figure 2

Structure identification and visualization of de novo library using Random Dataset.

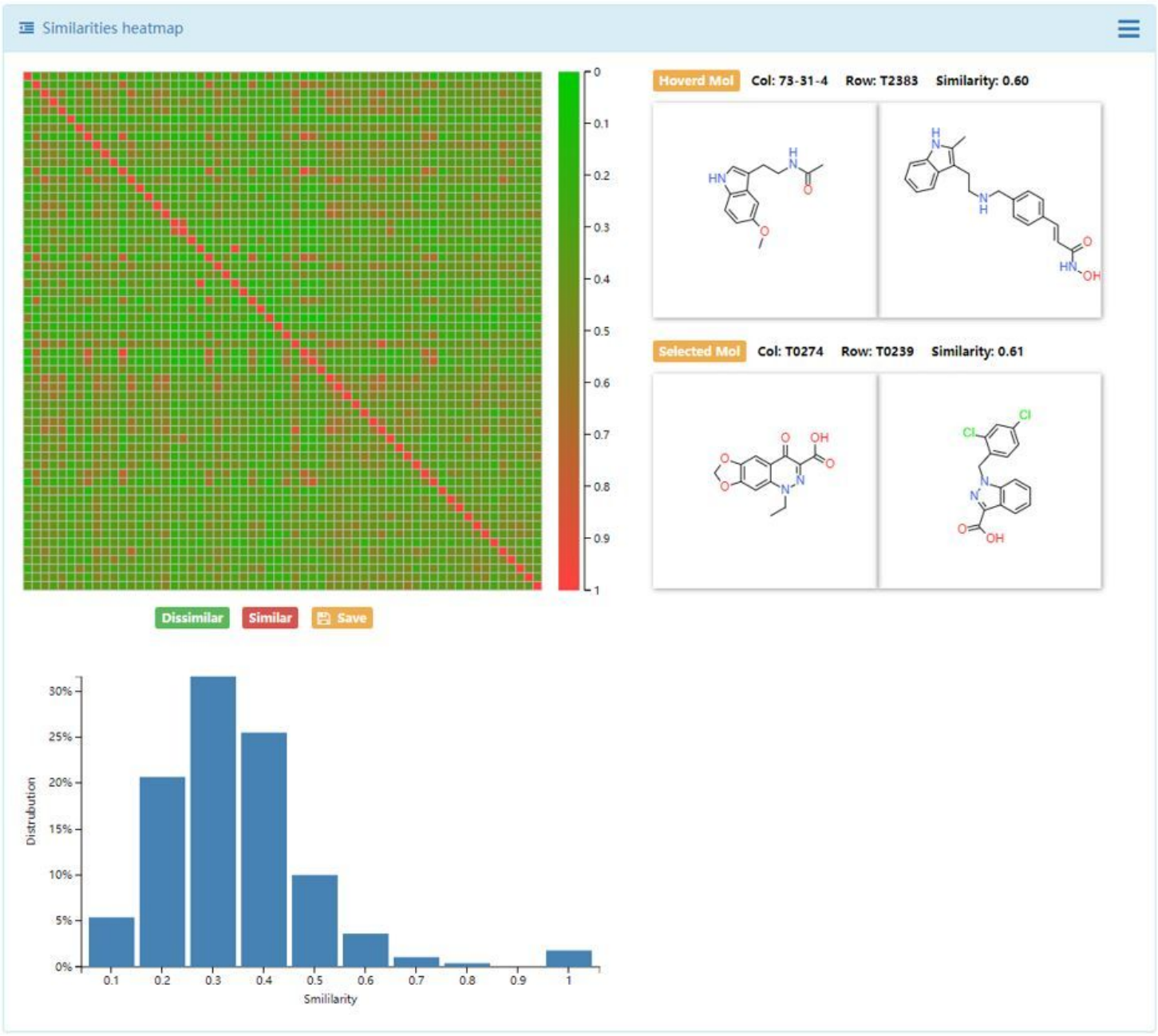


Figure 3

Chemical space illustration using similarity heatmap based on Drug Dataset.

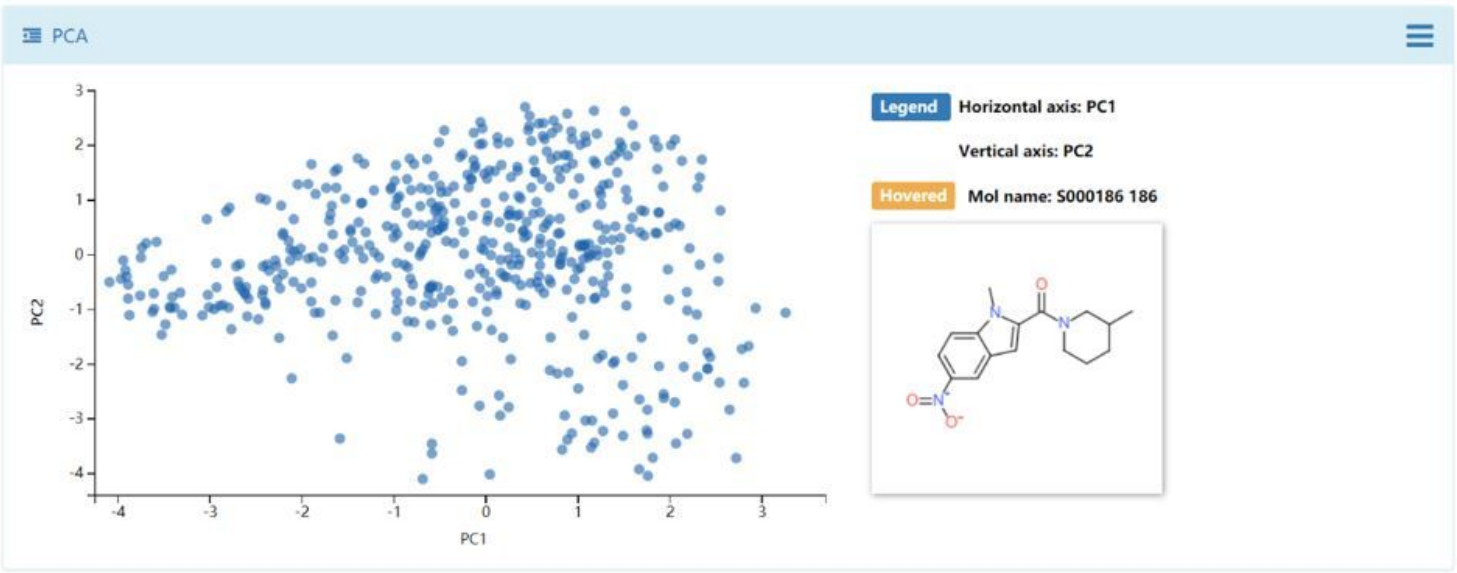


Figure 4

Chemical space illustration using principal component analysis (PCA) based on Random Dataset.

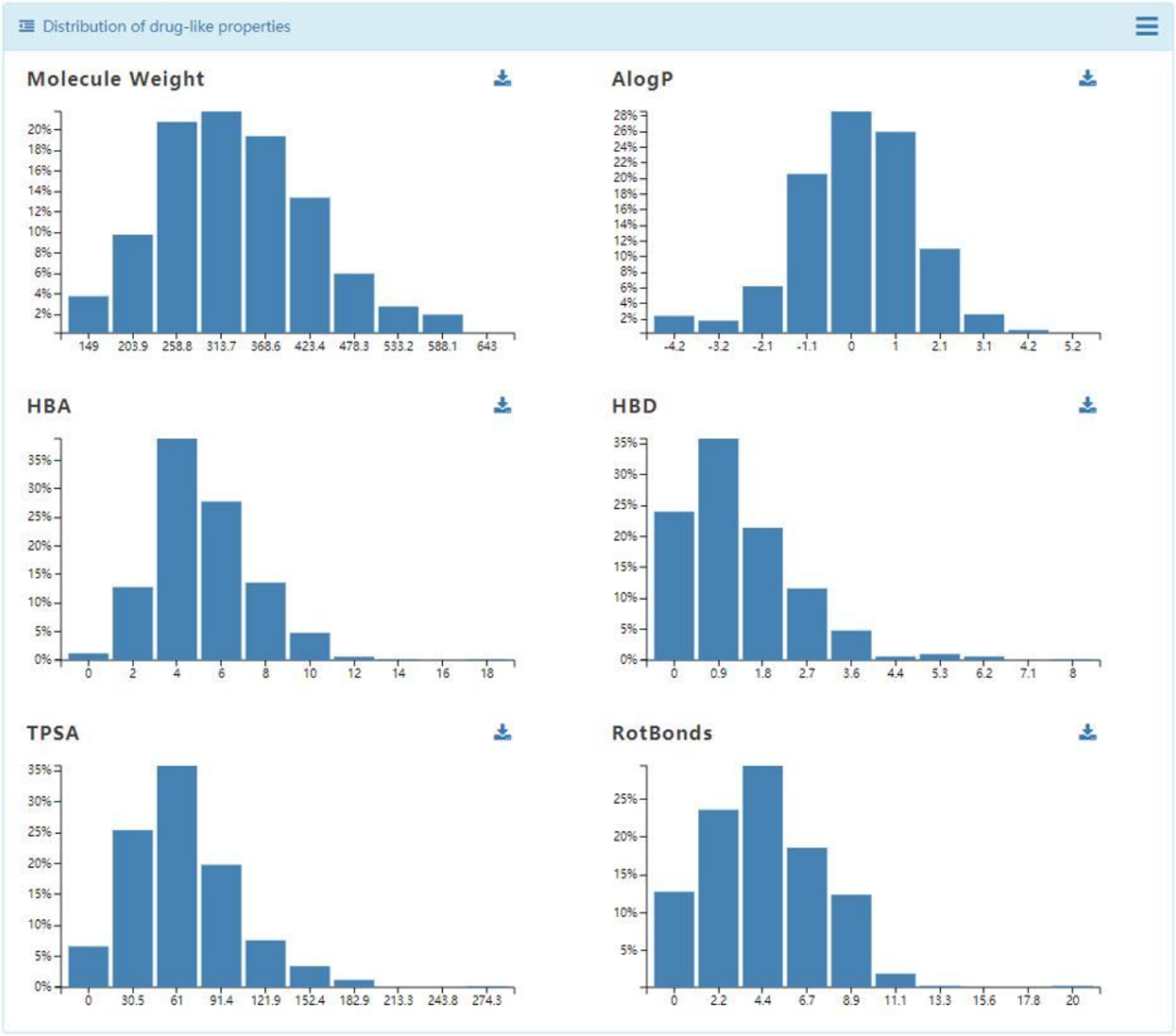


Figure 5

Distribution of drug-like properties based on Random Dataset.

Figure 6

Scaffold statistics of chemical scaffolds of de novo library based on Random Dataset.

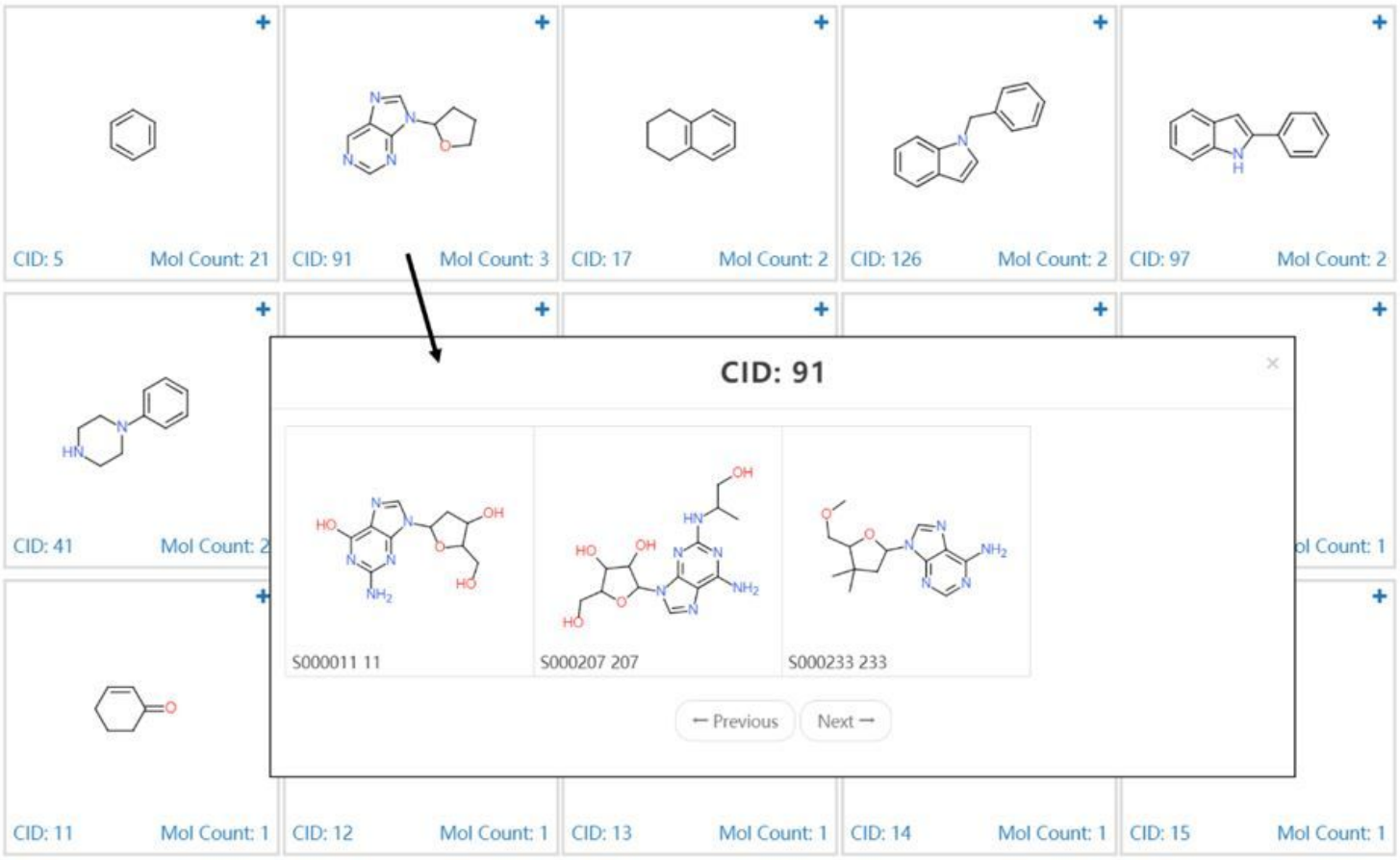


Figure 7

Grid view of the chemical scaffolds of the de novo library based on Random Dataset.

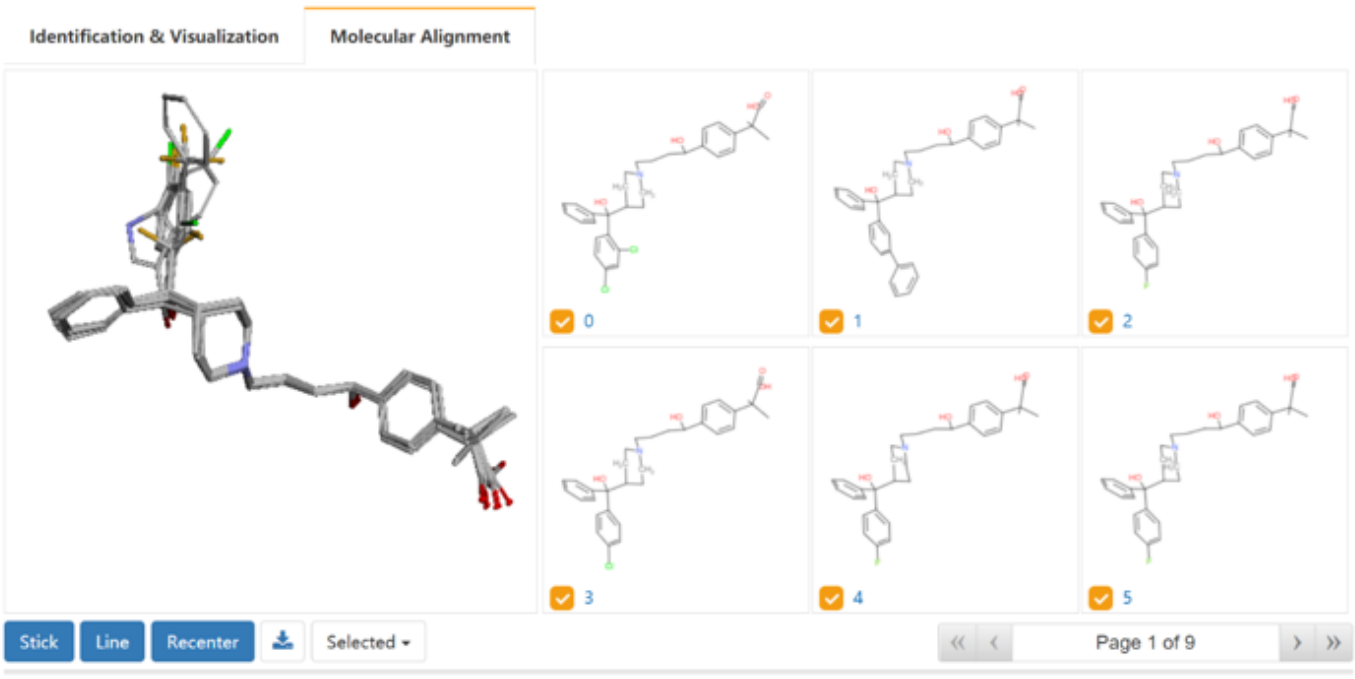


Figure 8

Molecular alignment of the scaffold-focused de novo library based on Focused Dataset.

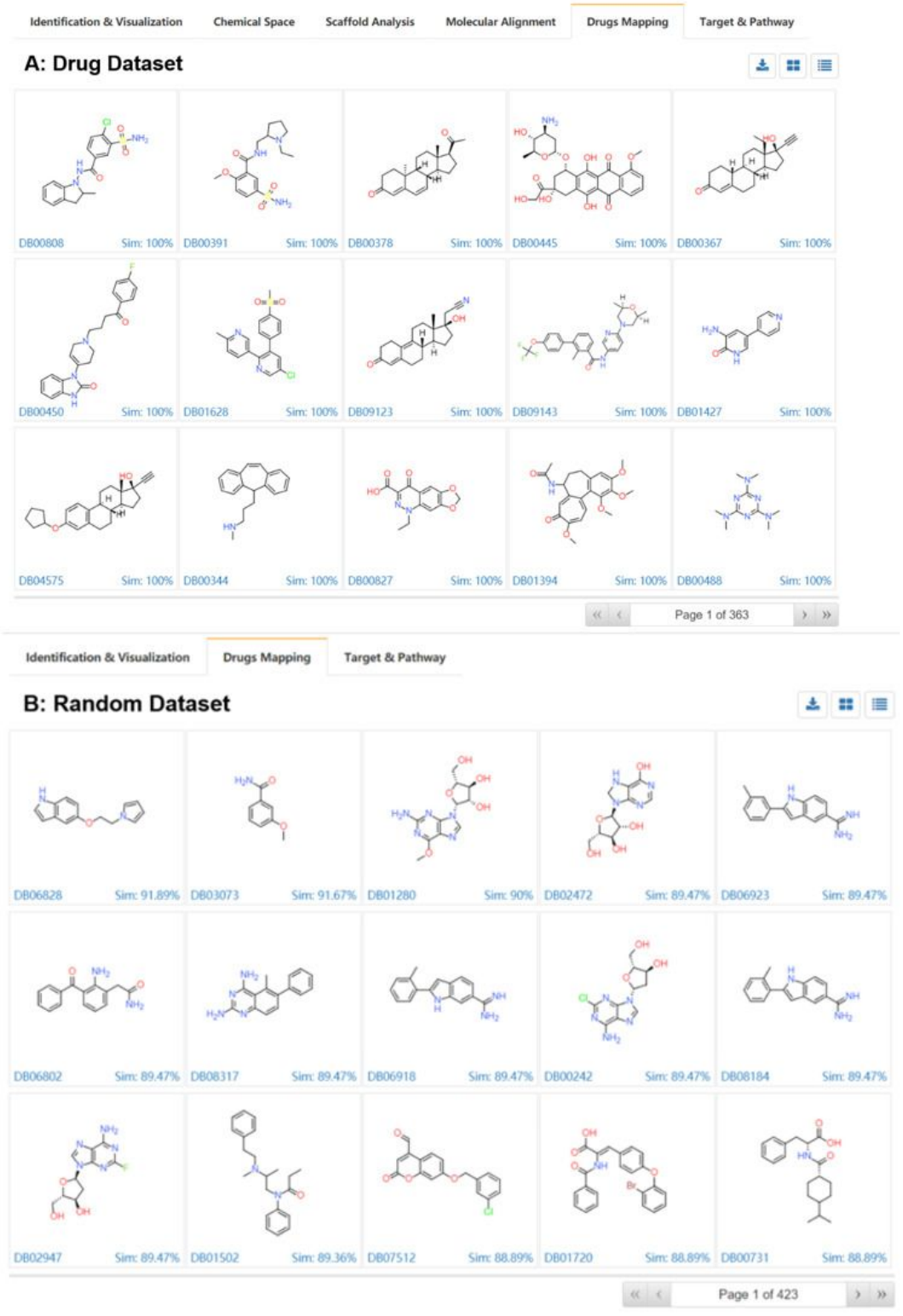


Figure 9

Grid view of the drugs mapping. A: Drug Dataset results; B: Random Dataset results.

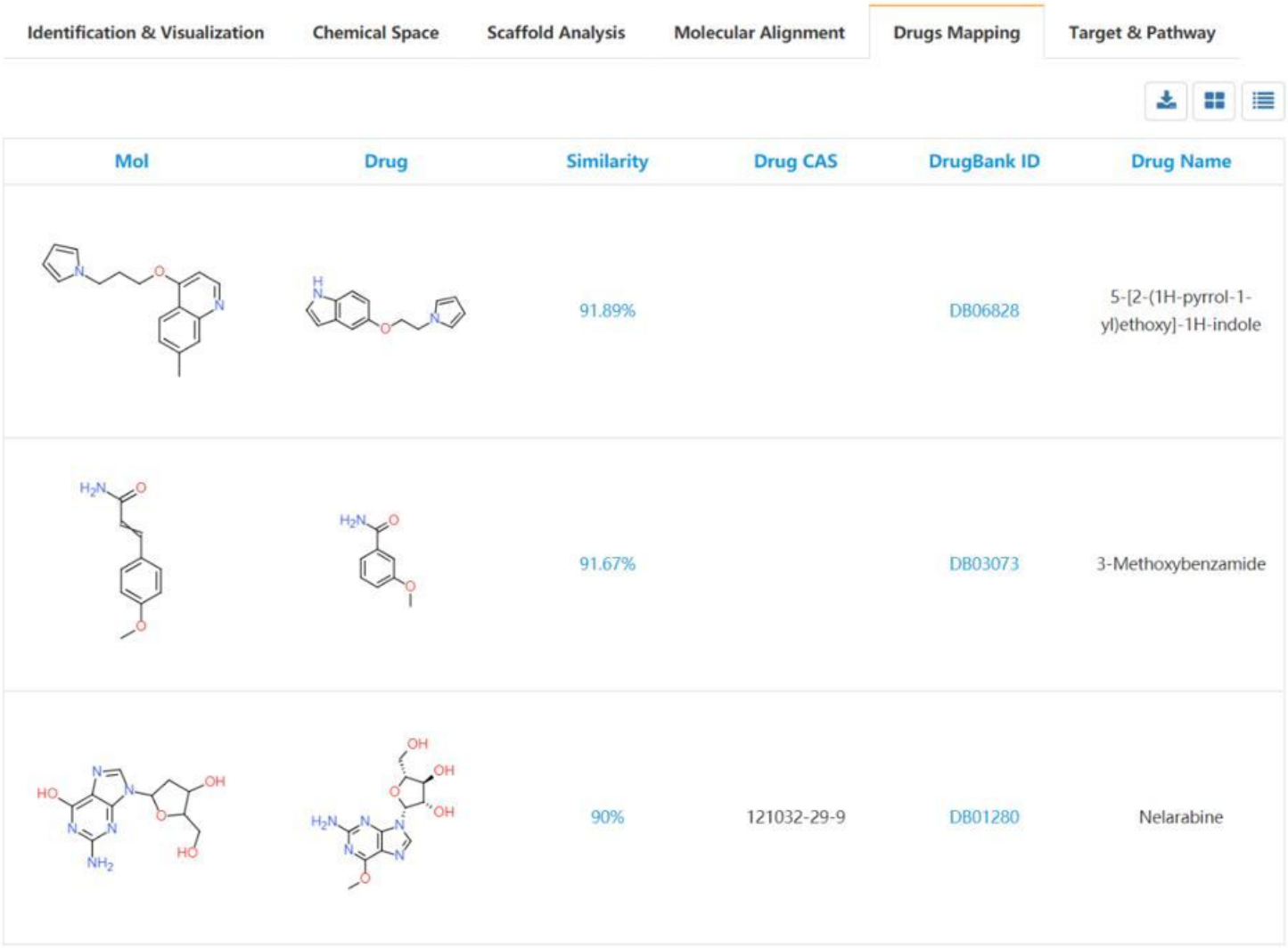


Figure 10

Table view of drugs mapping using Random Dataset.

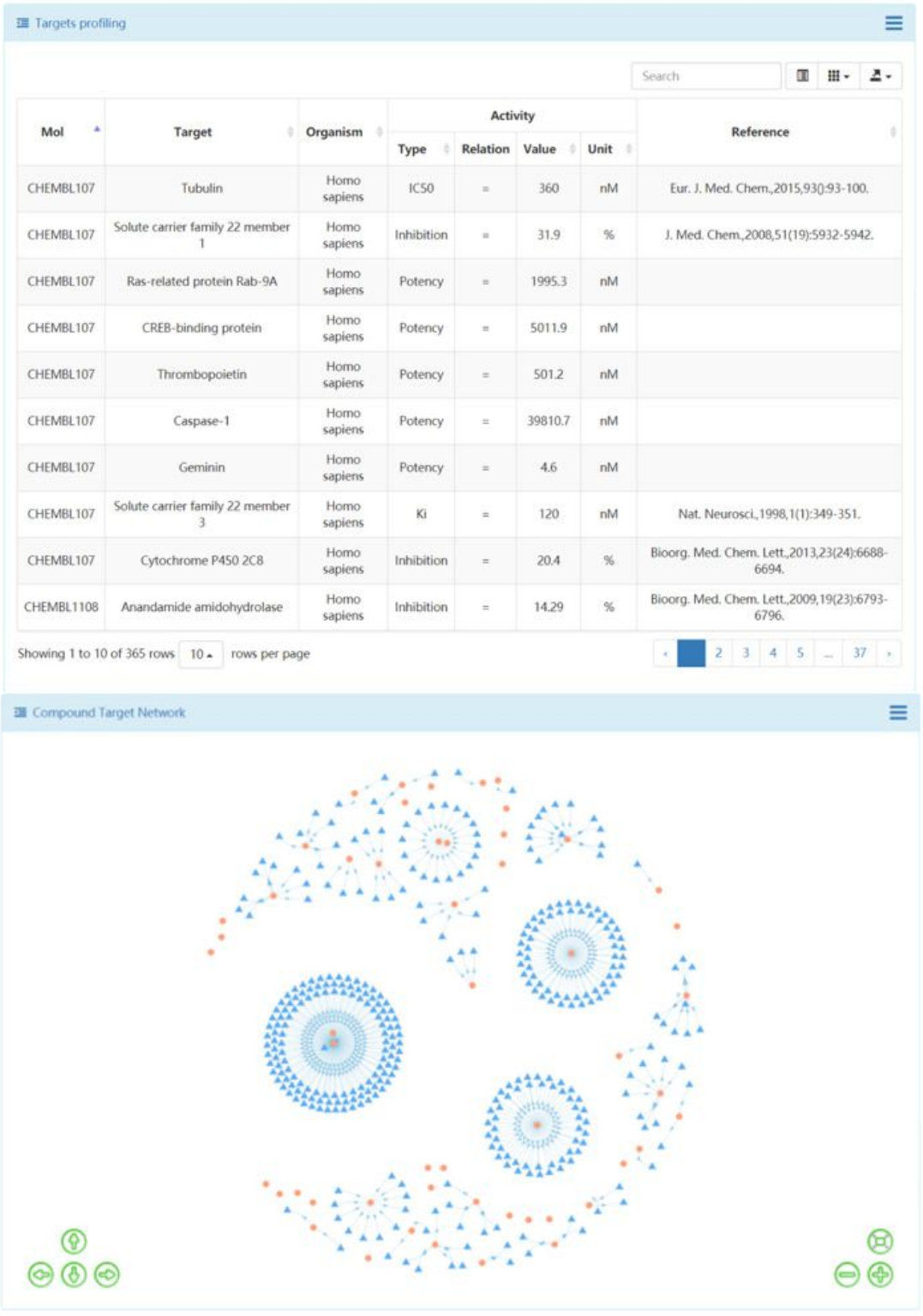


Figure 11

The identified targets in ChEMBL for Drug Dataset and the compound target network.

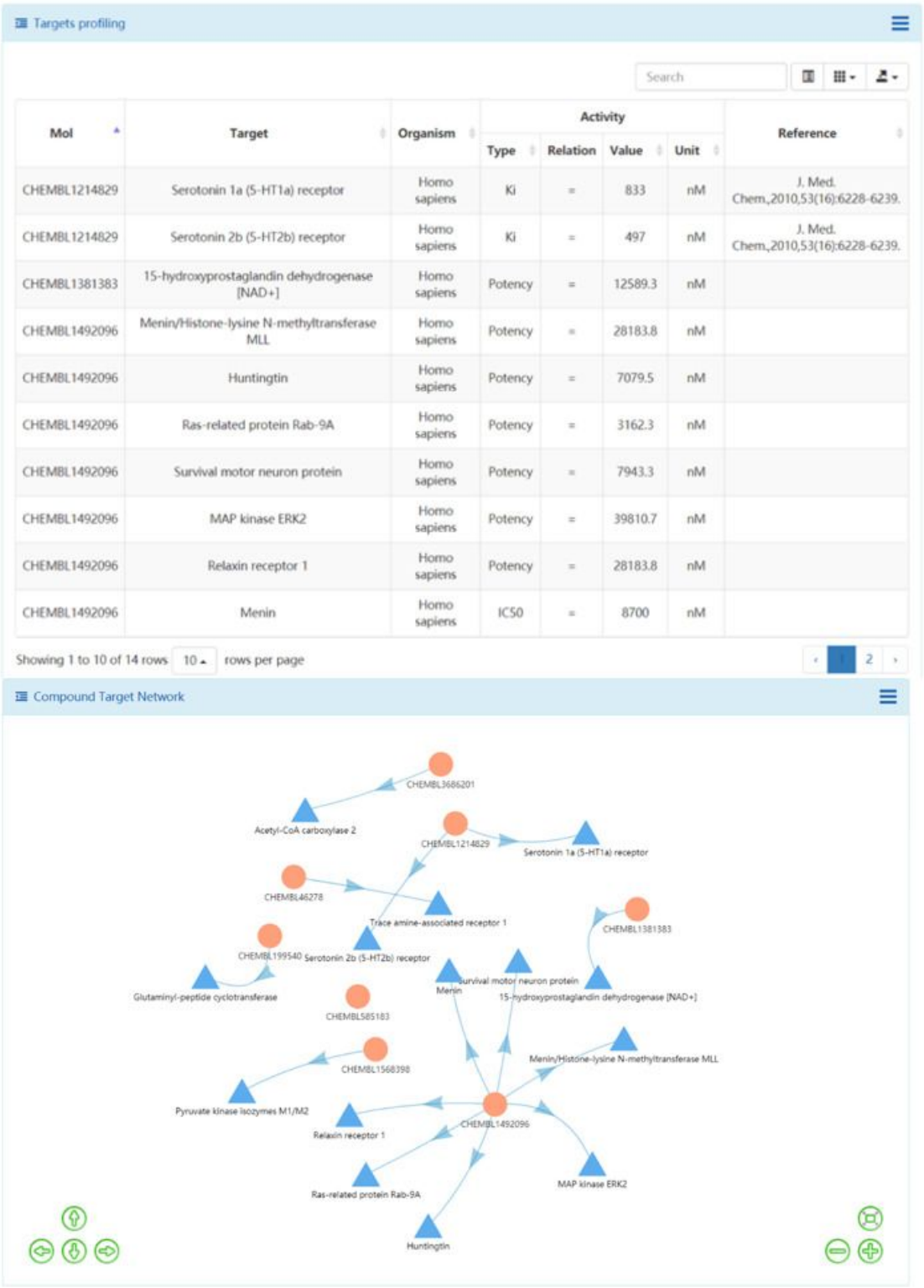
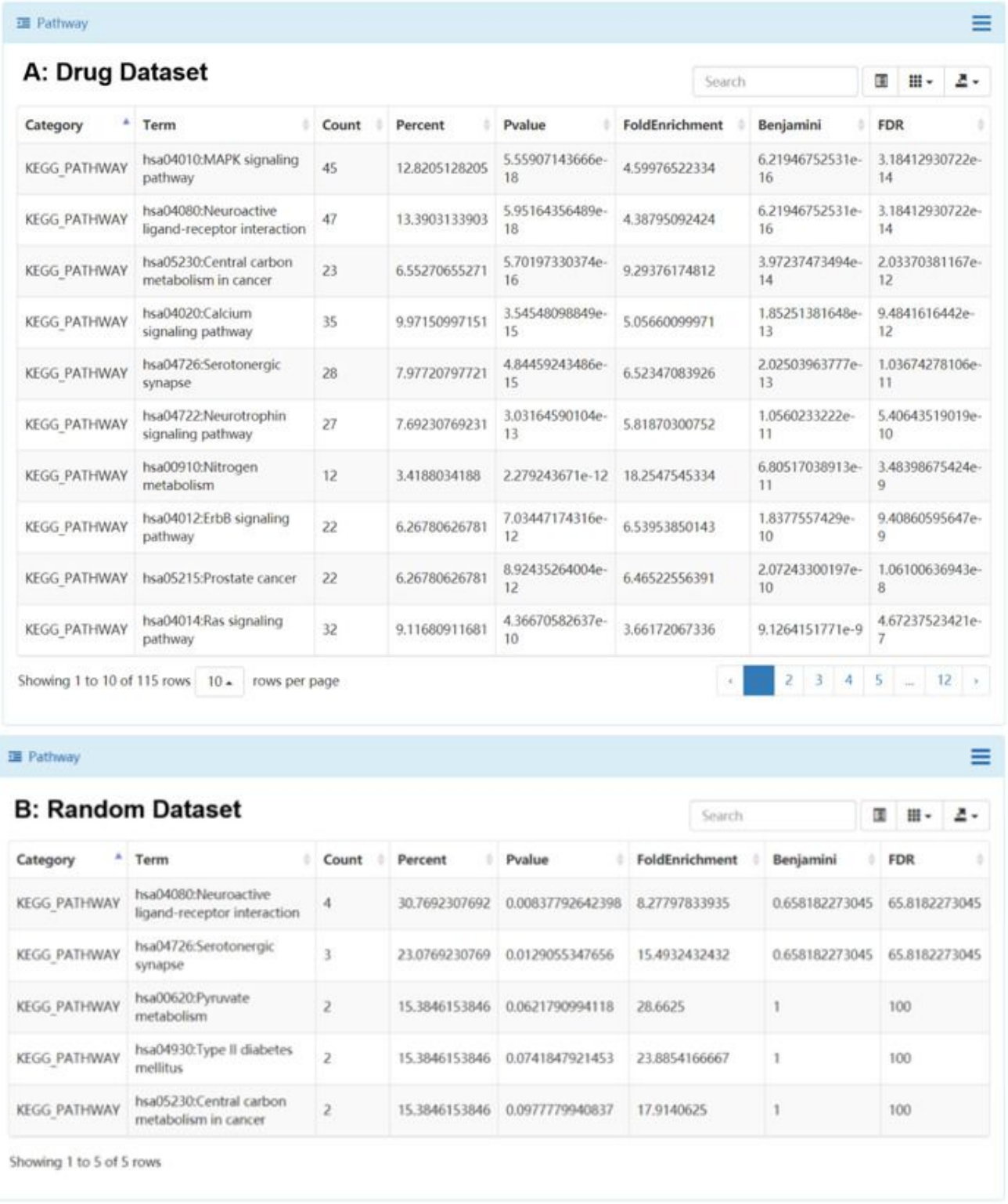


Figure 12

The identified targets in ChEMBL of de novo molecules and the compound target network.



Pathway

A: Drug Dataset

Search

Category	Term	Count	Percent	Pvalue	FoldEnrichment	Benjamini	FDR
KEGG_PATHWAY	hsa04010:MAPK signaling pathway	45	12.8205128205	5.55907143666e-18	4.59976522334	6.2194675253 1e-16	3.18412930722e-14
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	47	13.3903133903	5.95164356489e-18	4.38795092424	6.2194675253 1e-16	3.18412930722e-14
KEGG_PATHWAY	hsa05230:Central carbon metabolism in cancer	23	6.55270655271	5.70197330374e-16	9.29376174812	3.97237473494e-14	2.03370381167e-12
KEGG_PATHWAY	hsa04020:Calcium signaling pathway	35	9.97150997151	3.54548098849e-15	5.05660099971	1.85251381648e-13	9.4841616442e-12
KEGG_PATHWAY	hsa04726:Serotonergic synapse	28	7.97720797721	4.84459243486e-15	6.52347083926	2.02503963777e-13	1.03674278106e-11
KEGG_PATHWAY	hsa04722:Neurotrophin signaling pathway	27	7.69230769231	3.03164590104e-13	5.81870300752	1.0560233222e-11	5.40643519019e-10
KEGG_PATHWAY	hsa00910:Nitrogen metabolism	12	3.4188034188	2.279243671e-12	18.2547545334	6.80517038913e-11	3.48398675424e-9
KEGG_PATHWAY	hsa04012:ErbB signaling pathway	22	6.26780626781	7.03447174316e-12	6.53953850143	1.8377557429e-10	9.40860595647e-9
KEGG_PATHWAY	hsa05215:Prostate cancer	22	6.26780626781	8.92435264004e-12	6.46522556391	2.07243300197e-10	1.06100636943e-8
KEGG_PATHWAY	hsa04014:Ras signaling pathway	32	9.11680911681	4.36670582637e-10	3.66172067336	9.1264151771e-9	4.67237523421e-7

Showing 1 to 10 of 115 rows 10 rows per page

‹ 1 2 3 4 5 … 12 ›

Pathway

B: Random Dataset

Search

Category	Term	Count	Percent	Pvalue	FoldEnrichment	Benjamini	FDR
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	4	30.7692307692	0.00837792642398	8.27797833935	0.658182273045	65.8182273045
KEGG_PATHWAY	hsa04726:Serotonergic synapse	3	23.0769230769	0.0129055347656	15.4932432432	0.658182273045	65.8182273045
KEGG_PATHWAY	hsa00620:Pyruvate metabolism	2	15.3846153846	0.0621790994118	28.6625	1	100
KEGG_PATHWAY	hsa04930:Type II diabetes mellitus	2	15.3846153846	0.0741847921453	23.8854166667	1	100
KEGG_PATHWAY	hsa05230:Central carbon metabolism in cancer	2	15.3846153846	0.0977779940837	17.9140625	1	100

Showing 1 to 5 of 5 rows

Figure 13

The enriched KEGG pathways using the identified targets in ChEMBL. A: Drug Dataset results; B: Random Dataset results.