Chapter 15
A Multimodal Data Analysis Approach for Targeted Drug Discovery Involving Topological Data Analysis (TDA)

Muthuraman Alagappan, Dadi Jiang, Nicholas Denko, and Albert C. Koong

Abstract In silico drug discovery refers to a combination of computational techniques that augment our ability to discover drug compounds from compound libraries. Many such techniques exist, including virtual high-throughput screening (vHTS), high-throughput screening (HTS), and mechanisms for data storage and querying. However, presently these tools are often used independent of one another. In this chapter, we describe a new multimodal in silico technique for the hit identification and lead generation phases of traditional drug discovery. Our technique leverages the benefits of three independent methods—virtual high-throughput screening, high-throughput screening, and structural fingerprint analysis—by using a fourth technique called topological data analysis (TDA). We describe how a compound library can be independently tested with vHTS, HTS, and fingerprint analysis, and how the results can be transformed into a topological data analysis network to identify compounds from a diverse group of structural families. This process of using TDA or similar clustering methods to identify drug leads is advantageous because it provides a mechanism for choosing structurally diverse compounds while maintaining the unique advantages of already established techniques such as vHTS and HTS.

Keywords In silico • Topological data analysis • Virtual screening • High-throughput screening • Fingerprint • Computer aided drug discovery

M. Alagappan, B.S. • D. Jiang, Ph.D. • A.C. Koong, M.D., Ph.D. (✉)
Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305, USA
e-mail: muthuram@stanford.edu; jiangd@stanford.edu; akoong@stanford.edu

N. Denko, M.D., Ph.D.
Department of Radiation Oncology, Wexner Medical Center and Comprehensive Cancer Center, Ohio State University, Columbus, OH 43210, USA
e-mail: nicholas.denko@osumc.edu

C. Koumenis et al. (eds.), *Tumor Microenvironment*, Advances in Experimental Medicine and Biology 899, DOI 10.1007/978-3-319-26666-4_15

15.1 Introduction

In silico drug discovery refers to the use of computational technology to advance the process of drug discovery. More generally, in silico refers to a testing environment, in contrast to in vitro and in vivo, named after silicon-based computer chips. In silico drug discovery methods include basic processes such as database storage and querying and also technically advanced methods such as virtual ligand screening and topological data analysis [1]. These methods accelerate various steps in the drug discovery process, ranging from identification, 3D-reconstruction, and modeling of protein targets to identification of promising hits and lead drug compounds [2]. Empirically, in silico drug discovery, or as some refer to it, computer-aided drug design (CADD), has had numerous successes, such as aiding in the discovery of drugs against human immunodeficiency virus (HIV) including ritonavir and indinavir, as well as captopril, an angiotensin-converting enzyme (ACE) inhibitor for treating hypertension [3–5].

The rise in popularity of in silico drug discovery is correlated with the advent in computational power, the advancement of computational methods, and the failures of traditional drug discovery methods. Modern drug discovery on average requires 12.5 years and $1 billion for the launch of a technically successful drug, which is a highly prohibitive amount for most groups to pursue [6]. As a result, fewer entrants participate in drug discovery and those that do often necessitate exorbitant drug prices to recoup the costs of development. Fortunately, the improvement of computational speed and techniques has paved the way for in silico applications to lower the threshold and costs for participating in drug discovery and to accelerate the process of launching a successful drug. Modern processing power is constantly improving and current parallel processing systems can perform structure-based screening on more than 100,000 compounds per day [7]. We can only expect that as computational power and techniques continue to expand, the time and cost required to perform in silico drug discovery will be an even greater advantage when compared to traditional lead generation.

In this chapter, we describe a specific in silico approach to drug discovery that harnesses and combines the power of three popular techniques—virtual ligand screening, high-throughput screening (HTS), and fingerprint structural analysis—by using a fourth technique called topological data analysis (TDA), with the goal of identifying promising drug compounds from compound libraries, serving as a replacement to the hit identification stage and the subsequent lead generation stage of traditional drug discovery (Fig. 15.1). These methods will be employed once a

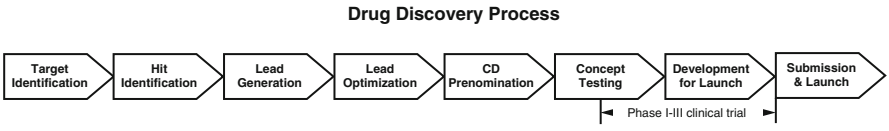


Fig.15.1 Steps in the traditional drug discovery pathway

biological target (receptor, enzyme, ion channel, etc.) has already been identified and validated.

The goal of the hit identification stage of drug discovery is to identify a subset of compounds from a much larger compound library that meet a success criterion against a certain target [8]. Hit identification is currently most often performed by high-throughput screening (HTS), a technique in which hundreds of thousands of compounds are assayed by robotics to determine which have the most significant activity against the target. The libraries that are screened include random libraries, thematic libraries (focused toward the biochemical function of the target), and knowledge-based libraries (focused toward compounds that possess characteristics known to be required of a lead) [8]. These HTS strategies have been extremely successful and adept at identifying potent leads and eventually successful drugs, and in fact, this process has initiated most approved drugs on the market. However, the current limitations of high-throughput screening for hit identification are cost, time required, and low hit rate. In a direct comparison of HTS to computer-aided drug discovery, researchers demonstrated a hit rate of 34.8 % (127 of 365 compounds) when using CADD versus only 0.021 % (81 of 400,000) when using HTS, representing a 1700-fold enrichment of the hit rate when using computational methods [9]. Our approach will offer a means of combining the unique advantages of HTS and virtual screening into a single workflow.

After hit identification, the next step in drug discovery is "hit-to-lead" (H2L) or lead generation. A lead is a hit that represents a compound series with the potential (signified by its potency, pharmacokinetic properties, selectivity, and low toxicity) to become a technically successful drug [8]. Lead generation takes into account practical aspects of the compound, such as its ability to be synthesized at large scales for low costs, availability of intellectual property, low cytotoxicity, favorable pharmacokinetics, and high selectivity amongst other characteristics.

This chapter focus on a novel approach to in silico drug discovery that aims to improve the two aforementioned stages: hit identification and lead generation, via the use of combined computational methods. The compounds that are identified must then undergo the later stages of drug discovery, which include lead optimization and concept testing (Fig. 15.1).

Our pathway for early stage lead generation will follow four stages: Stage 1 is virtual high throughput screening (vHTS), Stage 2 is traditional high throughput screening (HTS), Stage 3 is fingerprint analysis of small molecules, and Stage 4 will integrate the results with topological data analysis (TDA). Stage 1 and 2 are commonly used techniques for hit generation. In Stage 3, we will use PubChem's fingerprints to harvest complex, granular structural information about our compounds. Finally, in Stage 4 we will create a topological network of our compound library based on their structural similarities as derived from molecular fingerprints (Stage 3). The topological network will serve as a framework for more intelligent hit selection using the results from Stages 1 and 2 (Fig. 15.2).

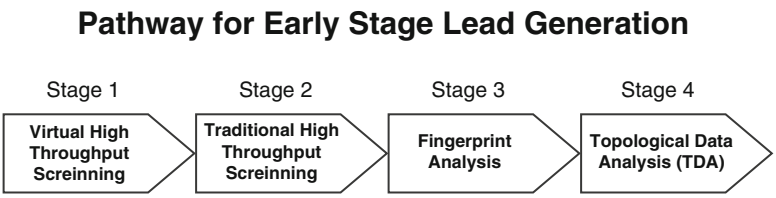


Fig. 15.2 A schematic view of our proposed multimodal technique for early stage lead generation

15.2 Drug Discovery Techniques

15.2.1 Virtual Ligand Screening

Virtual ligand screening or virtual high-throughput screening (vHTS) is an in silico approach to drug discovery which involves the computational simulation of compound-target binding to determine which compounds have the best binding potential.

15.2.2 High-throughput Screening (HTS)

High-throughput screening is a robotic approach for screening thousands of compounds in short time frames to obtain an initial sense of their viability.

15.2.3 Fingerprint Structural Analysis

Fingerprints are a collection of binary variables that describe structural features of a compound, which allow researchers to perform multivariable data analysis.

15.2.4 Topological Data Analysis (TDA)

Topological data analysis is a mathematical technique for studying relationships among data points in high dimensional datasets through the creation of similarity networks [10–13].

15.3 Choosing a Compound Library

The first step of this methodology requires the researcher to have access to a library of small molecule compounds, which will serve as the basis for the remainder of the techniques. The compound library can consist of any number of compounds,

ranging from thousands to millions of small molecules. Although even a generic, broad library of millions of compounds from a publicly available sources, such as PubChem's entire compound library of more than 50 million compounds, is suitable, it is preferable to choose selective libraries that are enhanced toward the functions of the intended target or that are already proven to be pharmacologically successful, such as The Library of Pharmacologically Active Compounds (Sigma-Aldrich, St. Louis, MO) or the FDA Approved Drug Library (Enzo Life Sciences, Farmingdale, NY). The advantage to smaller, selective compound libraries is multifold: (a) they allow for marginally faster computational processing times, (b) they are sold in sets that allow them to be easily tested in HTS assays, and (c) some have been empirically validated as potentially successful compounds, thus eliminating the need to iterate through inherently unstable compounds. This final advantage allows the researcher to be more successful in hit-to-lead advancement, as the hits are likely to be high quality, high potential drug compounds.

15.4 Stage I: Virtual High Throughput Screening (vHTS)

Structure-based drug design (SBDD) is a general term for drug discovery techniques that utilize the 3-dimensional structure of a target to predict and optimize the activity of drug compounds. A specific method of SBDD is virtual screening or virtual high throughput screening (vHTS), which is a technique for computationally simulating the drug-target interactions of compounds to quantify their effect potential [14]. Its objective is to quickly identify "hits" or promising compounds within a larger library, which makes it comparable to traditional high throughput screening. However, its comparative advantage is that all of the simulations can be done computationally rather than experimentally, resulting in large savings in time and cost.

The primary distinction to be made within vHTS is between a ligand-based and receptor-based approach [15]. In the ligand-based approach, a particular ligand of interest is used to identify similar ligands, using techniques such as similarity searching, fingerprint analysis, and quantitative structure–activity relationships (QSAR). In this method, the objective is to start with a potentially promising ligand and find others that share similar qualities. In contrast, the receptor-based approach starts with a 3D target protein of interest and attempts to identify potential ligands based on their ability to successfully interact with the target, through computational docking and scoring algorithms. We suggest the receptor-based approach of vHTS because it allows one to discover a broad diversity of ligands, as the identification methods are based on their relationship to the target rather than their similarity to each other.

The first step in vHTS is to select a docking program to run the virtual screen. Several such programs are publicly available including AutoDock (Scripps Research Institute, San Diego, CA), DOCK (UCSF, San Francisco, CA), LigandFit [16], and FlexX (Universitat Hamburg, Hamburg, Germany). We have personally employed AutoDock Vina as it offers comprehensive tools and is freely available for academic purposes.

Once a docking program is chosen and installed, the next process is to prepare the protein target, which involves two sub-steps: (a) resolving the 3D structure, and (b) choosing a binding site for docking. Owing to the advances in experimental methods such as nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and electron microscopy (EM), we now know the 3D structures of thousands of protein targets [17]. The vast majority of these structures are deposited in the Protein Data Bank (PDB), which contains structures ranging from protein fragments to large macromolecules. As of 2007, the PDB included over 53,000 protein structures with more than 8000 new structures being added each year [17]. However, if the 3D structure of a protein of interest has not yet been resolved, there are comparative homology modeling techniques, which involve sequence alignment and construction of missing coordinates, to predict its 3D structure [3]. Although not perfect, these methods are well-described and empirically successful [18]. Once the 3D structure is available, the researcher must determine the optimal binding site on the target for the program to use in its docking simulations. Binding sites are usually chosen based on either known biological information about the mechanism of the protein or on cocrystal structures that have a known ligand [3]. However, computational programs, including POCKET and SURFNEt also exist to discover novel binding sites within a target using techniques such as locating concave invaginations and evaluating protein dynamics [3].

Now that a docking program, compound library, and protein target are all chosen and prepared, the remaining steps involve running the vHTS on your library of compounds. When operating on the docking software, there are a few prerequisite steps. First, the docking site (active site) of the target protein needs to be defined by an accurate "grid box" that encompasses the entire surface of the binding site (Fig. 15.3). Second, the structure files of the target protein and screening compounds need to be formatted according to the requirements of the docking software. Third, requisite information for both the target protein and the virtual compound library need to be appropriately recorded in the configuration file of the docking software. With these steps completed, the virtual screening should proceed without complication.

The output of the vHTS should be a list of the compounds and their respective scores (often presented as binding energies) as it relates to their binding efficacy. The compounds with the best scores represent the "hits" from the vHTS, and any number of them can be tested experimentally to verify their binding activity. Although it is sufficient to proceed with these hits to the "hit-to-lead" stage of drug discovery, we will instead save these scores as Column 1 in a datasheet, which we will use later when we integrate the results of the three stages of analysis.

15.5 Stage II: Traditional High-Throughput Screening (HTS)

The introduction of automated high-throughput screening in the 1990s represented a significant milestone in the field of drug discovery. For perhaps the first time, large compound libraries, containing hundreds of thousands of compounds, could

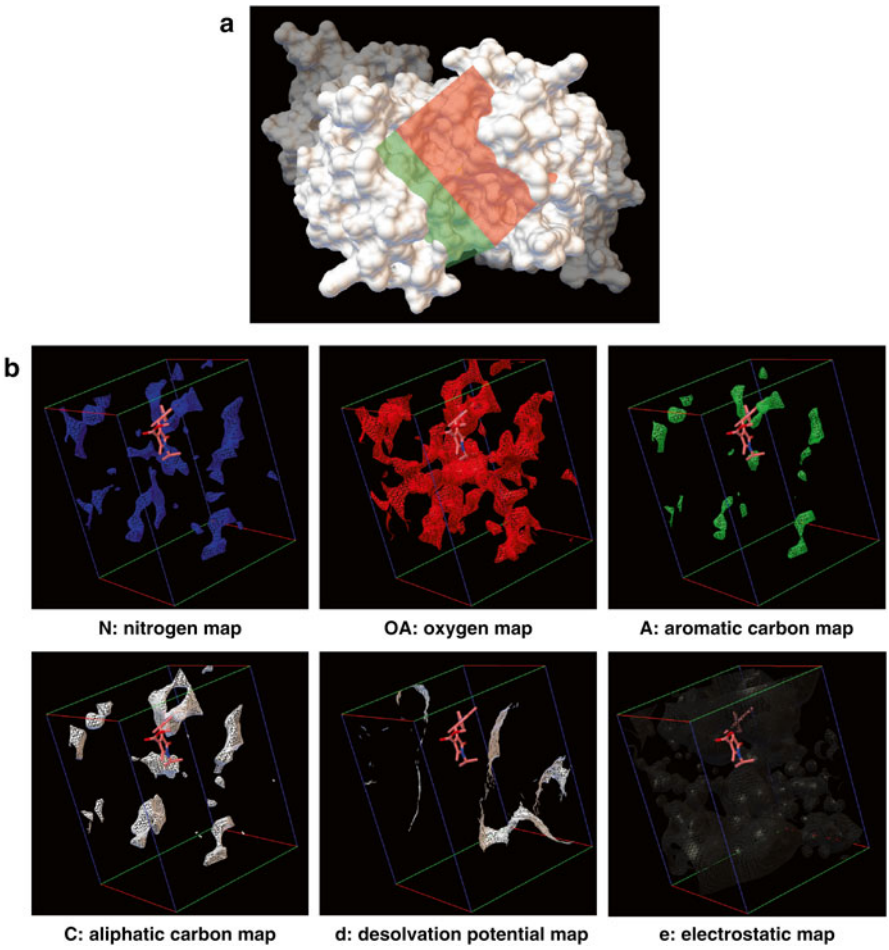


Fig. 15.3 A visual depiction of AutoDock's virtual screening software. (**a**) A grid box encompasses the ribonuclease domain of an IRE1α dimer. (**b**) Affinity grids are calculated for each atom of the compound

be quickly and efficiently screened by robots in the order of days to reveal a smaller subset of compounds with the highest potential for activity. By doing so, researchers could quickly identify a group of high-potential compounds to progress to the lead generation stage of discovery. This method further optimized the drug discovery process by eliminating the need for extensive "hit-to-lead" translation, as most of these compounds were already ready for large-scale pharmaceutical development [8]. These screenings were traditionally performed on large diversity-driven libraries, where the focus was on quantity of compounds rather than likelihood of activity. However, as discussed previously, researchers are currently trending toward using thematic libraries that are built with prior knowledge of the target using combinatorial chemistry.

High-throughput screening is the most empirically successful method of lead generation, as almost every new drug compound has emerged through a process involving HTS [8]. Examples are numerous and include Merck Research Laboratories' discovery of diketoacids, a novel HIV-suppressing class of compounds which block viral integration of HIV-1 integrase and Harvard University's discovery of monastrol, an anticancer agent that blocks mitosis by inhibiting kinesin Eg5, after screening the Diverset E (Chembridge Corporation, San Diego, CA) library of more than 16,000 compounds [8].

Although many researchers are now favoring virtual high-throughput screening to traditional high-throughput screening because of the improved speed, lower cost, and increased efficiency, there are still several unique comparative benefits to HTS, which help it remain as an important step in the drug discovery process. First, HTS allows you to identify compound hits without significant prior knowledge about the target, such as binding sites, compound affinities, or structural characteristics. This advantage allows researchers to identify hits for targets where the 3D structure is unresolved or for targets that are not mechanistically understood. Second, improvements in HTS efficiency and assay quality will improve the hit-finding ability of HTS compared to vHTS [19]. Newer assays are being developed to optimize sensitivity, thereby increasing the ability to detect weakly active compounds, which could become very successful future leads. Although favoring sensitivity will slightly increase the false positive rate, these falsely positive compounds can be discarded in future stages of the drug discovery process.

We briefly describe the steps of our high-throughput screens. We begin with a primary screen that tests our entire compound library. As there are thousands compounds to test at this stage, the screening is performed on the main reporter cell line at only one concentration per compound. Since the majority of compounds in the library will be negative hits, this step serves to quickly narrow down the library to only the most promising compounds based on their efficacy in the single tested concentration. Then a subset of compounds, now in the range of hundreds to thousands, will be tested in a secondary screen (Fig. 15.4). For this step, an approach called quantitative high throughput screening (qHTS) can be applied, in which each compound is tested at a series of concentrations on both the main reporter cell line and a counter-screen reporter cell line, usually used to control for undesired effects (e.g., cytotoxicity) [20]. The results of the secondary screen are a dose-activity curve for each compound at the various tested concentrations.

As before, these results can be independently used to advance certain compounds to the lead optimization stage. However, for our methodology, we will save the secondary screen assay results from one specific concentration as a "percent-inhibition" value into a new column (Column 2) in the same spreadsheet as before. These two columns will be utilized in the fourth stage of our drug discovery process, when we use topological data analysis to integrate the results.

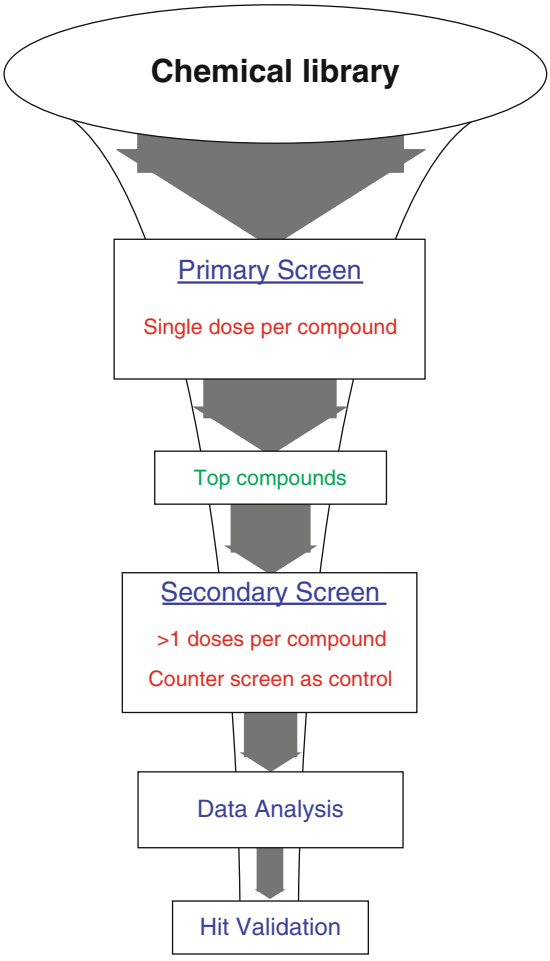


Fig. 15.4 The workflow of our high-throughput screening: a primary and secondary screen are performed on an initial compound library

15.6 Stage III: PubChem Fingerprint Analysis

In the final stage of this process, we use structural information about compounds to further enhance our ability to identify hits. The major advancement in the use of chemical and structural information to assist with drug discovery was quantitative structure–activity relationship (QSAR) models, which uses statistical models to predict how different physicochemical properties of chemical substances affect biological activity [21]. The earliest example of QSAR dates back to 1893, when Charles Richet described a novel relationship between the solubility of compounds and their toxicity [22]. From that starting point, QSAR has developed tremendously

as a means of rapidly evaluating compounds for their predicted activity, prioritizing compounds for synthesis, and serving as a cheaper alternative to costly biological assays [21]. As Verma et al. describe in their book, new evolutions in QSAR include expanded dimensionality such as, "3D-QSAR correlating activity with non-covalent interaction fields surrounding the molecules, 4D-QSAR additionally including ensemble of ligand configurations in 3D-QSAR, 5D-QSAR explicitly representing different induced-fit models in 4D-QSAR" [21].

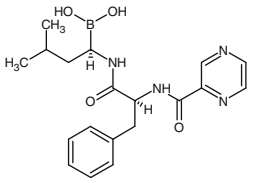
Virtual screening, the technique introduced in "Stage 1" of this chapter, is a natural, automated extension of QSAR, as it uses the structural properties of compounds to predict binding ability. However, in this stage, we want to introduce another technique, fingerprint analysis, which also incorporates structural information about compounds.

Fingerprints refer to a group of variables that are collectively used to convey information about the structural features of a compound. PubChem generates a list of 881 binary substructure fingerprints for every chemical compound in their database [23]. The fingerprints are distributed into broad structural categories, including element counts, atom pairs, ring characteristics, atom neighborhoods, and SMARTS patterns. PubChem offers download of fingerprint information for any compound in their database. Subsequent use of Base64 decoding can convert the fingerprint into a list of 881 binary variables for each compound (Fig. 15.5). We can then update our dataset to include a list of all of the compounds from the secondary screen (rows), which were each annotated with several variables: (1) the virtual binding score (from Stage 1) in Column 1, (2) the HTS "percent inhibition" score (from Stage 2) in Column 2, and (3) 881 binary fingerprint variables in Columns 3-883.

15.7 Stage IV: Topological Data Analysis

Using the results aggregated from the prior three stages, we now have a dataset that contains information about compounds with respect to their binding affinity to the target, their empirical HTS assay result, and their granular structural features. In this stage, we will employ topological data analysis (TDA), a mathematical technique for studying shapes and preserving high-dimensionality, to create a similarity network of our compounds. TDA will allow us to visualize our library of compounds in a two-dimensional network, whereby compounds (located in nodes) are connected to each other by a series of edges that reflects their level of shared similarity (Fig. 15.6). Therefore, two compounds that share several similar properties will appear closer together in the network, whereas two vastly different compounds will be farther apart. This network will make it simple to create subgroups or families of similar compounds, which can then be used to select the best compound leads.

We use Ayasdi Core (Ayasdi Inc., Menlo Park, CA) to create our TDA network of compounds. However, if access to Ayasdi Core is unavailable, alternative options include other TDA applications or employing weaker techniques such as K-means clustering to achieve somewhat similar results. We first discuss the methodology as



Bortezomib

156 bytes fingerprint:

AAADceJ7uAAAAAAAAAAAAAAAAAAAAAAAAA8QAAAAAAAAAABwAA
oHgAQCAAADSjBngQ+gJLIEACoAzV3VACCgCQ3EiAI2IG4dMgIYHrA0TG
UIIgglgDliMcYiICOAAAAAAAAAAAAAAAAAAAAAAAAAAAA==

Binary	ASCII	Binary	ASCII	Binary	ASCII	Binary	ASCII
000000	A	010000	Q	100000	g	110000	w
000001	B	010001	R	100001	h	110001	x
000010	C	010010	S	100010	i	110010	y
000011	D	010011	T	100011	j	110011	z
000100	E	010100	U	100100	k	110100	0
000101	F	010101	V	100101	l	110101	1
000110	G	010110	W	100110	m	110110	2
000111	H	010111	X	100111	n	110111	3
001000	I	011000	Y	101000	o	111000	4
001001	J	011001	Z	101001	p	111001	5
001010	K	011010	a	101010	q	111010	6
001011	L	011011	b	101011	r	111011	7
001100	M	011100	c	101100	s	111100	8
001101	N	011101	d	101101	t	111101	9
001110	O	011110	e	101110	u	111110	+
001111	P	011111	f	101111	v	111111	/

881 binary substructure fingerprint:

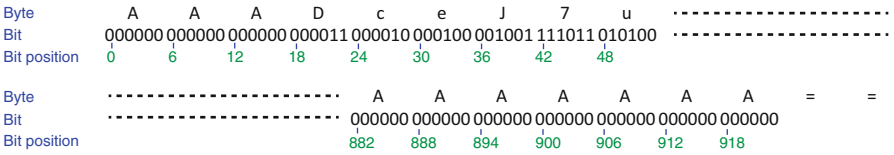


Fig. 15.5 Schematic diagram depicting the translation of a 156-byte PubChem fingerprint into an 881 variable binary output, which will be used to describe the structural features of each compound

it relates to TDA and then explain a parallel process that can be employed with clustering algorithms.

A topological analysis is created with two types of parameters: a metric and a lens. A metric is a mathematical function used to measure the similarity between two points in some space (usually between rows in the data). The choice of a metric

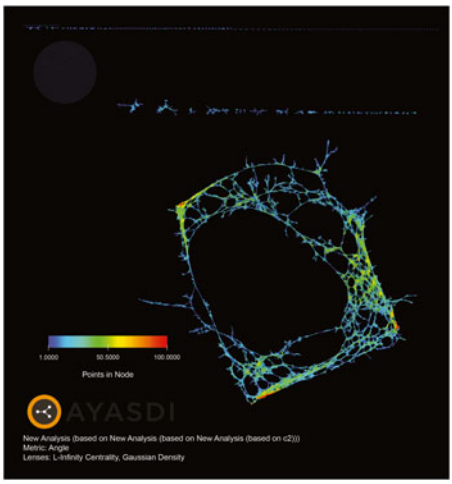


Fig. 15.6 An example network created by performing Topological Data Analysis (TDA) on sample biological data sets using the Ayasdi Core software platform (Ayasdi.com, Ayasdi Inc., Menlo Park, CA). Nodes in the network represent clusters of compounds and edges connect nodes that contain similar compounds. The color scheme in this graph is the average number of data points in each node

greatly influences how the similarity of data points is assessed. Lenses are real valued functions on the data points. Lenses are used to create overlapping bins in the data set, where the bins are preimages under the lens of an interval. Overlapping families of intervals are used to create overlapping bins in the data. Metrics are used with lenses to construct the Ayasdi Core output. Further, multiple lenses can be used in each analysis. There are two user-determined parameters that are important in defining the bins. The first is *resolution*, which determines the number of bins. Higher resolutions correspond to a greater number of bins and a fewer number of data points per bin. The second is *gain*, which determines the degree of overlap of the intervals. Increasing the gain increases the overlap between bins, such that they share more data points in common [10–13].

Once the bins are constructed, the software performs a clustering step within each bin. To do so, it uses single linkage clustering with a fixed heuristic for the choice of the scale parameter [11]. Finally, the network is formed by creating a node for each cluster and connecting nodes if the corresponding clusters contain at least one data point in common. TDA also allows the user to identify the important factors or variables that distinguish two groups of nodes. To do so, it employs a nonparametric statistical test (Kolmogorov–Smirnov) in combination with the p-value (t-test).

The metric we will select for our analysis is Angle, and the lenses are L-Infinity Centrality and Gaussian Density. The Angle metric computes the angle between two rows, by taking the inverse cosine of the dot product of the vectors. The L-infinity Centrality lens calculates for each point x the maximal distance from x to any other data point in the dataset. Therefore, this lens is strongly correlated with density in normally distributed variables, but not as much in bimodal distributions. The Gaussian Density lens applies a kernel Gaussian density estimator by treating each row in Euclidean data space to estimate the density of that point.

The next choice with TDA algorithms is which columns in the data set you want to use for creation of the similarity network. The columns you choose will dictate

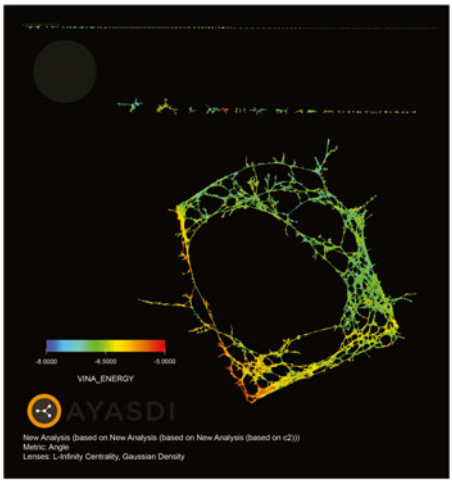


Fig. 15.7 The same network shown in Fig. 15.6 is now colored by virtual docking energy from Stage 1. For virtual docking energy ("VINA_ENERGY"), more negative values indicate higher binding affinity. Thus, nodes in red contain compounds with the best binding affinities. A range of −8 to −5 kcal/mol is displayed

the nature of the similarities. For our methods, we will use only the 881 structural fingerprint variables to create our network. The resulting graph will contain every compound in our library, connected to one another based on their structural fingerprint similarities (Fig. 15.6). Therefore, compounds in the same section of the network will inherently share a large set of common structural features. Ayasdi Core allows for further analysis by coloring the network by our variables of interest. In our technique, we will color our network first by the vHTS score and then by the HTS score. The resulting pictures of the network provide a visual demonstration of which areas of the network contain compounds with the best binding energies and HTS results. Since we know the network was created based on structural fingerprint similarities, we can infer that those areas of the network must inherently contain certain structural features that may be related to their success.

After creating the network, we will first examine it to understand its shape. Distinct "shapes" within the network represent families of compounds that share common properties, in this case structural features. These two-dimensional shapes instruct us of hidden multidimensional meaning about the structural features they share [11, 12]. We will physically create these groups by drawing in boundaries between distinct shapes. Each enclosed group or family will then contain a set of compounds that we hypothesize shares similarities.

Next, we will annotate the network with the findings from Stages 1 and 2 to begin the process of picking lead compounds. First, we will annotate the network by virtual binding score (Fig. 15.7). Annotating refers to overlaying the network with a heat map (also referred to as "coloring") that reflects each compound's value for a selected variable, in this case the virtual binding score. The annotations will depict which compounds within each family have the best virtual binding affinity. We can then select the 3–5 best compounds by virtual affinity score within each family as our leads. Second, we will annotate the network by the HTS assay result (Fig. 15.8). Once again, we will select the 3–5 best compounds in each family, making note of which compounds were also chosen because of their virtual affinity score.

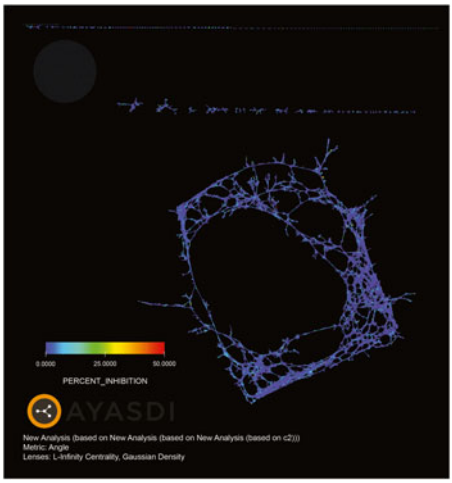


Fig. 15.8 The same network shown in Fig. 15.6 is now colored by "percent-inhibition" values by high-throughput screening from Stage 2. For HTS ("PERCENT_INHIBITION"), more positive values indicate a higher percent inhibition of our selected protein target. A range of 0–50 % inhibition is shown

This method of choosing lead compounds after stratifying by structural families is advantageous because it ensures that a diverse set of leads are chosen (as they come from a diverse collection of structural families) and it is also more forgiving to certain compounds that may not have great absolute virtual binding or HTS scores, but may still turn out to be excellent drug choices. If we had simply identified our hits by choosing the best 100 compounds by virtual binding score, we may unfairly bias ourselves towards one specific family of compounds that happen to be favored by that specific docking mechanism. As a result, we may overlook compounds with excellent in vivo potential that simply didn't have the features to make them successful in initial screens with virtual docking or in vitro HTS.

As a further refinement step, we can use the annotation results to determine which 1–2 families of compounds performed best in both the virtual screen and HTS. With this knowledge, we can recursively add more compounds from those families to our initial library, as we know those families may contain certain important features. This method of expanding compound libraries around the hit compounds is becoming more prevalent in the pharmaceutical industry [24].

If you do not have access to TDA, another process of lead identification can be accomplished using k-means clustering, whereby you separate your compounds into k clusters based on shared properties, in this case the 881 PubChem structural fingerprints. Once the clusters are chosen, you can rank the compounds within each cluster first by vHTS and next by HTS score. As was done above, select the 3–5 highest ranked compounds within each family for each scoring system. Instead of using a heat map or coloring mechanism to identify the best compounds, you will then numerically rank the compounds based on the value of their scores (vHTS and HTS).

To then accomplish the refinement step of adding new compounds, rank the entire compound library by these two scores to determine which 1–2 families demonstrate the most consistent success. Recursively add new compounds from these families to the initial compound library.

15.8 Conclusion

Our methodology described here combines virtual screening, high-throughput screening, structural fingerprints, and topological data analysis to more intelligently identify lead compounds from large compound libraries. Each of these individual techniques has several strengths but also shortcomings that can be mitigated by utilizing the techniques together. For example, virtual screening, although extremely fast and inexpensive, suffers from needing access to the 3D structure of the protein target and having some preexisting knowledge about its binding domains and tendencies. Conversely, HTS allows for comprehensive in vitro testing without any prior knowledge of the target, yet it often only tests the compound at one dose and in one type of assay, both of which are restrictions that could unfairly filter out otherwise excellent compounds. By combining both methods, we make use of their unique advantages without inheriting their shortcomings.

Researchers have already shown the benefit of combining multiple types of methods in their work. For example, prior research has shown that a machine-learning model trained on HTS and chemical fingerprints performed similar or better than a model trained on either subset alone [24]. But even such approaches suffer from the same problem of selectively favoring compounds that perform well in the tested assays. Although this seems like a logical choice, it misses the mark because the ability of HTS, virtual screening, or fingerprints to predict in vivo drug success is far from perfect. There are many compounds that demonstrate excellent in vitro success but fail to have an effect in vivo, and vice versa [25]. And most of these shortcomings are because of pharmacokinetic and metabolic properties that are potentially secondary to their chemical structures. In other words, there are many compounds that aren't structurally suited for vHTS or HTS success, which are eliminated at an early stage in the drug discovery processes, even though they may have been excellent drugs. Similarly, certain structural features may predispose a certain family of compounds to receive very high scores in vHTS and HTS, which would lead to a disproportionately large number of compounds from that family advancing through the drug discovery pipeline. Thus, by using topological data analysis to first create families of structural compounds, and then using our virtual screening and HTS results to choose the most promising compounds within each family, we ensure that a diversity of compounds are chosen for the next stage of the drug development cycle.

References

1. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol. 2007;152(1):9–20. doi:10.1038/sj.bjp.0707305.
2. Drews J. Drug discovery: a historical perspective. Science. 2000;287(5460):1960–4.
3. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334–95. doi:10.1124/pr.112.007336.

4. Van Drie JH. Computer-aided drug design: the next 20 years. J Comput Aided Mol Des. 2007;21(10-11):591–601. doi:10.1007/s10822-007-9142-y.
5. Talele TT, Khedkar SA, Rigby AC. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr Top Med Chem. 2010;10(1):127–41.
6. Bains W. Failure rates in drug discovery and development: will we ever get any better? 2004.
7. Agarwal AK, Fishwick CW. Structure-based design of anti-infectives. Ann N Y Acad Sci. 2010;1213:20–45. doi:10.1111/j.1749-6632.2010.05859.x.
8. Golebiowski A, Klopfenstein SR, Portlock DE. Lead compounds discovered from libraries. Curr Opin Chem Biol. 2001;5(3):273–84.
9. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem. 2002;45(11):2213–21.
10. Carlsson G. Topology and Data. Bull Amer Math Soc. 2009;46:255–308.
11. Singh G, Memoli F, Carlsson G. Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurograph Symp Point Based Graph. 2007.
12. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. Sci Rep. 2013;3:1236. doi:10.1038/srep01236.
13. Sarikonda G, Pettus J, Phatak S, Sachithanantham S, Miller JF, Wesley JD, et al. CD8 T-cell reactivity to islet antigens is unique to type 1 while CD4 T-cell reactivity exists in both type 1 and type 2 diabetes. J Autoimmun. 2014;50:77–82. doi:10.1016/j.jaut.2013.12.003.
14. Jain AN. Virtual screening in lead discovery and optimization. Curr Opin Drug Discov Devel. 2004;7(4):396–403.
15. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol. 2006;10(3):194–202. doi:10.1016/j.cbpa.2006.04.002.
16. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):213–21. doi:10.1107/S0907444909052925.
17. Dutta S, Burkhardt K, Swaminathan GJ, Kosada T, Henrick K, Nakamura H, et al. Data deposition and annotation at the worldwide protein data bank. Methods Mol Biol. 2008;426:81–101. doi:10.1007/978-1-60327-058-8_5.
18. Evers A, Gohlke H, Klebe G. Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. J Mol Biol. 2003;334(2):327–45.
19. Mayr LM, Bojanic D. Novel trends in high-throughput screening. Curr Opin Pharmacol. 2009;9(5):580–8. doi:10.1016/j.coph.2009.08.004.
20. Inglese J, Auld DS, Jadhav A, Johnson RL, Simeonov A, Yasgar A, et al. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc Natl Acad Sci U S A. 2006;103(31):11473–8. doi:10.1073/pnas.0604348103.
21. Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design--a review. Curr Top Med Chem. 2010;10(1):95–115.
22. Kubinyi H. 3D QSAR in drug design. In: Theory methods and applications, vol 1. New York: Springer; 1993.
23. Bolton E, Wang Y, Thiessen P, Bryant S. PubChem: Integrated platform of small molecules and biological activities. Annual Reports in Computational Chemistry. 2008;4:217–241.
24. Riniker S, Wang Y, Jenkins JL, Landrum GA. Using information from historical high-throughput screens to predict active compounds. J Chem Inf Model. 2014;54(7):1880–91. doi:10.1021/ci500190p.
25. Lin JH, Lu AY. Role of pharmacokinetics and metabolism in drug discovery and development. Pharmacol Rev. 1997;49(4):403–49.