METHODOLOGY

Hilbert-Curve Assisted Structure Embedding Method

Gergely Zahoránszky-Kőhalmi*, Kanny K. Wan and Alexander G. Godfrey

*Correspondence:
gzahoranszky@gmail.com
National Center for Advancing Translational Sciences (NCATS/NIH), Medical Center Dr., Rockville, USA
Full list of author information is available at the end of the article

Abstract

Motivation: Chemical space embedding methods are widely utilized in various research settings where dimensional reduction, clustering or effective visualization is required. Still, it remains unsolved to date to embed molecules into a chemical space in which chemotypes are organized along clear principles and which can be intuitively interpreted by medicinal chemists.

Results: In this study we present the Hilbert-Curve Assisted Space Embedding (HCASE) method which was designed to provide intuitive space embedding results. The method achieves this objective by mapping a set of reference scaffolds and, subsequently compounds, to a pseudo-Hilbert-Curve (PHC) with the help of the known Scaffold-Key algorithm. The PHC can be embedded into a higher dimensional space readily according to established algorithm. Through a series of experiments, we successfully demonstrate the unique and novel propertise of the HCASE method in this proof-of-concept study. Experiments involved a large and a natural product-derived set of $63,783$ and 546 scaffolds, respectively, both from the ChEMBL database. Chemical space embedding was performed on the DrugBank and CANVASS libraries. Comparative analysis demonstrated that the performance of the HCASE method is not only on par, to say the least, with prior art method, but it excelled in providing intuitive chemical space embedding.

Availability: https://github.com/ncats/hcase

Keywords: chemical space embedding; clustering; Hilbert-Curve; Scaffold-Keys; HCASE; dimension reduction

1 Introduction

Embedding molecular structures into a chemical space is a versatile technique that is central to a wide range of data analysis scenarios in cheminformatics. A number of methods, like principal component analysis (PCA) [1], multi-dimensional scaling (MDS) [2], t-Stochastic Neighbor Embedding (t-SNE) [3], Uniform Manifold Approximation and Projection (UMAP) [4] and the self-organizing maps (SOM) method [5], help reduce the dimensionality of data to facilitate subsequent cluster analyses or to provide insightful visualizations. While most of these methods can be performed in a relatively straightforward manner from an operational point of view, this somewhat deceiving simplicity comes at the cost of somelimitations to applicability and interpretability.

For instance, PCA [1] can only analyze linear relations present in the data at hand. This limitation is overcome by non-linear approaches, such as the related non-metric multi-dimensional scaling (MDS) [2] and manifold-supported methods [6], such as

t-Stochastic Neighbor Embedding (*t*-SNE) [3] and the recent Uniform Manifold Approximation and Projection (UMAP) [4].

All of these non-linear methods, with the exception of MDS are challenged with the means of computing the distance between the embedded datapoints. Interpretation of the underlying organizing principle of the embedded structures is convoluted for all known space-embedding methods. Also, the chemical space created by both linear and non-linear methods is influenced by the dataset at hand. This affects the interpretation of results, and makes the comparison of individually embedded datasets quite difficult. While this can be addressed to some extent by merging the datasets before the embedding process, but this solution is not rubust against the incorporation of additional data.

Background

The aim of performing a chemical space embedding analysis is to create a "map" of compounds. A coumpound's position in this map ideally should reflect structural and/or other profperties of interest (e.g. physichochemical properties), and as a result, the relative position of compounds within the map should be reflective of their similarities in these properties. A chemical space map can help medicinal chemists make quick, intuitive analyses about the structure and properties of compounds in a project based on their location in the map. For example, one would expect that compounds within the same chemotype in a structure-activity-relationship (SAR) series will be placed closely on the map, whereas dissimilar chemotypes farther apart.

While creating such maps is entirely possible with existing methods, e.g. with *t*-SNE, medicinal chemists and data analysts are challenged with the interpretations of the results. For demonstration purposes, a map (embedding) of approved drugs has been generated using the *t*-SNE algorithm.

In order to demonstrate various aspects of the chemical space embedding process, five drug molecules were selected randomly, as well as the five nearest neighbors (NNs), i.e. structurally most similar five compounds, of each (see: *Fig. 1*).

As shown on *Fig. 2a*, the resultant map shows a great clustering and separation of similar and dissimilar molecules, respectively, as one would expect. However, from a medicinal chemist's standpoint some important aspects of the data analysis remain hidden.

For instance, a chemist might want to know if certain regions of this map encode certain type of chemotype, e.g. based on size, complexity and so on. Unfortunately, maps generated with existing embedding methods provide little, if any help to chemists in this regard. Furthermore, generating a map often requires setting certain non-intuitive parameters, like the *perplexity* in the case of *t*-SNE, which many chemists may not be familiar with. This parameter influences which compounds should be closely or farther apart the resulant chemical space map [7]. The choice of the parameter can affect the layout of the map, and often in an unpredictable manner, as it is demonstrated on *Fig.2b*.

Finally, the layout of the map generated by the same space embedding method can be greatly altered when one adds or removes molecules when repeating the embedding process, as demonstrated on *Fig. 1b-1c*. This makes it challenging to compare

the embedding of a library that is changing over time. The only difference between the two maps is that the *Fig.2c* was generated using the 90% of the molecules of the embedding used in *Fig.2b* and the same highlighted molecules. The two maps show little resemblance despite the relatively small change in input.

Further information regarding the embedding process of drug molecules with the *t*-SNE algorithm is provided in Section *"Embedding of Drug Molecules with t-SNE Algorithm"* and *Fig. S1-S2* in *Additional File 1*.

In this study we introduce a novel space embedding method that addresses the above detailed challenges of existing space embedding methods in creating an intuitive chemical space.

Related Methods

Besides the general space embedding methods, chemistry specific space embedding methods exist [8]. The PCA-based "ChemGPS" [9] and Molecular Quantum Number [10] methods address the issue of creating embedding via a mechanism that is not influenced by the dataset at hand[11]. The SOM-related "generative topographic mapping GTM" method by *Lin et al* [12], and the "constellation plots" [13] take advantage of scaffold-compound relations to enhance the embedding. Furthermore, the GTM method defines a grid with the help of "landscape structures" that guides the subsequent embedding of compounds. While the GTM and constellation plot methods indeed address many challenges, the organizing principle of the compounds, or landscape structures of both methods is based on heuristics, and is not intuitive from a medicinal chemistry standpoint. A recent method (TMAP) [14] uses a combination of nearest neighbor and minimal-spanning trees and force-based network layout to generate embedding, but the organizing principle of the method is still based on heuristics. Thus, it cannot guarantee that regions in the resultant map can be intuitively interpreted.

The above methods intended to solve known challenges related to chemical space embedding, but none of them has solved all the aforementioned challenges to a degree that would result in intuitive chemical space maps for medicinal chemists. Nevertheless, these methods gave rise to many important concepts and aspects that are utilized in this study.

In this proof-of-concept study, we set forth criteria for a chemical space embedding method that provides intuitive results and easy interpretation from a medicinal chemistry point of view and devised a new method that produces results reflective of such characteristics. In the following section, the new method is introduced in details and its applicability is demonstrated via a set of experiments.

2 Computational Methods and Datasets

In this section, we detail the development of a novel chemical space embedding method. The description of other analytical methods and dataset involved in this study is also provided below.

2.1 Development of the Intuitive Structure Embedding Method

2.1.1 Rationale

Here, we define a set of criterions underpinning a method that is capable of providing a chemical space embedding so that the outcome of the analysis can be interpreted intuitively from a medicinal chemistry point of view:

- coordinates of structures generated by space embedding process is not influenced by the structural features of other compounds in the compound set to be embedded
- domains of a generated chemical space should convey well defined structural meaning
- mapping of structures to coordinates is deterministic
- the organizing principle should be simple to understand and should follow a well defined ordering of structures
- outcomes of space embeddings performed independently should be directly comparable both numerically and visually
- method must not be limited to capturing only linear relations
- ability to process reasonably large datasets (consisting of thousands of structures)
- ability to quantify distance between structures in the embedded space.

Existing chemical-space embedding methods, to our knowledge, don't meet all of the above criteria. However, most of these methods could be turned into one that meets almost all of these criteria following a two-step procedure. First, a pre-embedding is generated with the help of a pre-defined set of "landscape" structures, e.g. Bemis-Murcko scaffolds (BMSs) [15]. Next, the most similar landscape structure is identified for each compound in the data set at hand. Then, each compound would assume the coordinates of the landscape structure identified as the most similar to a given compound. However one of the most important criteria from the interpretation point of view is not met when using the above embedding strategy with existing methods in that the organizing principle of pre-embedding of landscape structures remains mostly hidden for the researcher. Moreover, the organizing principle is practically the result of certain optimization processes that largely depend on the input data at hand.

In this study, we aimed at constructing an embedding method that addresses this limitation so that it provides a simple, yet practical, embedding that can be interpreted intuitively by chemists and data analysts.

2.1.2 Method Design

In the light of the above collected criteria, we devised a novel chemical space embedding algorithm. The devised method was built on incorporating critical concepts introduced by prior art methods: use of landscape objects organized on a grid, use of embedding mechanism that is not influenced by the compound set to embedded, and the ability to change resolution of the embedding [5, 9, 12, 10, 11].

The foundation of the novel method is provided by a family of so-called space filling curves, namely by Hilbert-Curves [16, 17, 18]. Provided that an ordering between data points, here BMSs, exists, with the help of Hilbert-Curve it is possible to embed the data points into a space of higher dimension, such as $2D$, following an exact mathematical process. This embedding is a limit of embeddings resulted by utilizing so-called pseudo-Hilbert-Curves (PHCs) of increasing order. The peculiar characteristics of PHCs is that increasing the order of the PHC the position of a given data point will converge to a limit in the higher dimension, i.e. in the embedded space the positions of data points are stabilized utilizing PHCs of increasing order. Considering that implementation exists for embedding PHCs, the question remained: How can one obtain a well-defined and ordering of BMSs? Luckily, the Scaffold-Key (SK) algorithm addresses this exact question by providing a solution for the "intuitive" ordering of BMSs that was motivated by the analytical thinking of medicinal chemists [19]. For more information on the SK algorithm please refer to Section 2.3.

In the following section we provide the details of the structure embedding method that was designed with all the considerations detailed above.

2.1.3 Hilbert-Curve Assisted Structure Embedding Method

In order to define the chemical space of the Hilbert-Curve Assisted Structure Embedding (HCASE) method, a set of reference BMSs needs to be collected. The choice of reference BMS set depends on the context of scientific investigation. However, using a diverse set of BMSs or a collection of BMSs derived from compounds of a large bioactivity data set represent choices that can be adopted in a wide range of research settings. Note that compound structures that cannot be associated with a valid BMS structure are eliminated from the input set when generating the reference BMS set. Next, the SKs of reference BMSs is generated, and the BMSs are ordered according to their SK using alphanumeric ordering. In case of a tie, the InChI-Keys of BMSs are used to determine priority. In the arguably rare case when the InChI-Keys would be identical, then the "first" of such BMSs will gain priority. Of note, depending on the implementations of sorting algorithm, the choice of "first" BMS in a tie can be nondeterministic. Still, considering the low probability of such events, we consider the SK and InChI-Key based ordering practically deterministic.

Next, the reference BMS set is mapped on a line based on the rank of each BMS emerged from the SK-based ordering process. This line can be thought of a PHC which can be mapped to a $2D$ space, or even higher dimensions following a well-known algorithm [16]. The embedding of compounds with the help of such a line happens in a few steps.

First, the BMS of the compound at hand is extracted and the corresponding SK is generated. With the help of the SKs, the closest reference BMSs to the compound is identified. Next, the compound will assume the position of the closest reference BMS on the PHC. Finally, the PHC is mapped to a higher dimension space.

The process of mapping a PHC to higher dimension requires only two parameters as input: the order of the PHC and the number of dimensions. The latter was always set to $2D$ in this study, while the former was varied. Given the nature of PHCs, increasing the order of the PHC will lead to the stabilization of coordinates in the embedded space and to a more fine-grained embedding.

Reducing the algorithm to practice required us to take into consideration two observations. First, the number of potential coordinates in the embedded space is a function of the order of PHC and the number of dimensions in the available implementation of PHC algorithm [20, 21].

In $2D$, the PHC can be mapped on a $N \times N$ grid, where the value of N is given by *Eq. 1*, where z denotes the order of the PHC. Accordingly, the x and y coordinates can take on values between 0 and $z-1$, inclusive. Of, note we use the PHC-z notation in the text to distinguish PHCs of different order. Second, the PHC emerged from the reference BMS set contains a finite set of data points, i.e. BMSs. In the light of these limitations it was necessary to introduce a binning-mechanism in order to mimic the behavior of PHCs.

$$N = 2^z \tag{1}$$

The binning-mechanism treats the number of potential coordinates ($|D|$) in the embedded space as the number of bins (see: *Eq 2-3*). Then, the bin-size l is determined based on the ratio of the size of the reference BMS set ($|S|$) and the number of bins minus one (see: *Eq. 4*). Note, that the correction term is necessary as the Hilbert-curve implementation uses zero-indexing, hence the minus one term. Given a compound i and its closest reference BMS S_i, the bin index b_i of the compound is computed by first dividing the SK-based rank of S_i by the bin-size, then rounding the resultant number to the nearest integer (see: *Eq. 5*). Of note, when setting the parameters of the algorithm, it should be taken into account that limit of the resolution of the HCASE method is defined by the parameter combination where the number of potential coordinates exceeds the size of the reference BMS set.

$$D = \{(x, y)\} \quad | \quad \forall x : x \in [0, N-1], \quad \forall y : y \in [0, N-1] \tag{2}$$

$$|D| = N^2 \tag{3}$$

$$l = \frac{|S|}{|D| - 1} \tag{4}$$

$$b_i = \left\lfloor \frac{rank(S_i)}{l} \right\rceil \tag{5}$$

Computing the bin indices of each compound gives rise to a mapping on a PHC which can be mapped to $2D$ by defining the z parameter [16, 20]. The main steps of the HCASE algorithm are visualized on *Fig. 3*.

2.2 Pseudocode of the HCASE Method

The pseudocode of the HCASE method is provided below. Note that most of the functions highlighted with bold fonts represent well-known methods, therefore their pseudocode is not included. Such functions are: *generatePseudoHilbertCurve()*, *getHCCoordinates()*, *getScaffoldKey()* and *getBemisMurckoScaffold()*. The *binScaffolds()* and *getSKDistance()* functions are computed according to *Eq. 1-5* and *Eq. 6*, respectively.

Note that the lists in the pseudocode are zero-indexed. Furthermore, the elements of lists and tuples are also referenced according to array notation. Accordingly, the $D[0][0]$ in the pseudo code reads: in the first item of list D (which is a tuple), the value of the first variable.

Algorithm 1 HCASE Method

```
Input: int z (order of PHC)
Input: int n (number of dimensions)
Input: set of molecules M
Input: set of reference Bemis-Murcko scaffolds S

Variable: molecule mol
Variable: scaffold-key sk
Variable: list of (S, sk)-tuples S_sk
Variable: int b (bin index)
Variable: list of (sk, b)-tuples S_bin
Variable: Bemis-Murcko scaffold bms
Variable: (x ∈ ℕ, y ∈ ℕ)-tuple P
Variable: pseudo-Hilbert-Curve PHC
Variable: list of (mol, P)-tuples E

PHC := generatePseudoHilbertCurve (z, n)
for all S_i in S do
   sk := getScaffoldKey (S_i)
   S_sk.add (S_i, sk)
end for

S_sk := sort S_sk alphanumerically by SK in increasing order
S_sk := deduplicate S_sk by sk, keep first instance of identical tuples
S_bin := binScaffolds (S_sk, z, n)

for all m_i in M do
   bms := getBemisMurckoScaffold (m_i)
   sk := getScaffoldKey (bms)
   b := getClosestRefererenceBMSBinIndex (sk, S_bin)

   P := getHCCoordinates (b, PHC)
   E.add (m_i, P)

end for
return (E)

int function getClosestRefererenceBMSBinIndex(sk, S_bin)
Variable: int b_min (bin index of closest reference scaffold)
Variable: numeric d_sk
Variable: list of (b, d_sk)-tuples D

for all s_i in S_bin do
   d_sk := getSKDistance (sk, s_i[0])
   D.add (s_i[1], d_sk)
end for
D := sort D by d_sk and b, both in increasing order

b_min = D[0][0]

return (b_min)
```

2.3 Scaffold-Key Algorithm

As mentioned above, the general idea behind the SK algorithm was to provide an ordering of BMSs to imitate the thinking process of a medicinal chemist in analyzing BMSs based on their size, complexity and chemical composition. Furthermore, the SK algorithm aimed to provide a distance measure that surpasses fingerprint-based distance measure between scaffolds, due to known limitations [19]. To this end, 32 so-called "Scaffold-Keys" were defined that each capture unique structural aspects of a given BMS. The definition behind these 32 keys define the ruleset of the algorithm that is publicly disclosed in the original publication by *Ertl* [19]. The SK algorithm generates a 32-key SK for a given BMS which can be used to sort the BMSs or to define a distance measure between BMSs. Distance ($d_{SK}(i,j)$) between a pair of SKs of respective BMSs i and j can be quantified with the help of their SK according to *Eq. 6* as defined by *Ertl*. $SK_i(n)$ and $SK_j(n)$ denote the value of the n^{th} key in the SK of BMS i and j, respectively.

$$d_{SK}(i,j) = \sum_{n=1}^{32} \frac{\sqrt{|SK_i(n) - SK_j(n)|^3}}{n} \tag{6}$$

Since the SK algorithm does not have a publicly available implementation it was necessary to create an in-house implementation based on the published ruleset. The implementation follows the ruleset as truthfully as possible, with the only exception that optionally, it is possible to generate the InChI-Key [22] of BMS as an extra (last) key on the top of the original 32 keys. Moreover, a few of the original rules were defined in a slightly vague manner, therefore we could only attempt to match those as closely as possible in light of insufficient information. Nevertheless, clarification of rules, where it was necessary, is provided in "Appendix". Implementation of the SK algorithm is publicly available as a source-code repository at: https://github.com/ncats/hcase [23].

SKs were generated with the in-house implementation of the SK algorithm, as well as the d_{SK} distances between BMSs.

2.4 General Cheminformatics Operations

Structures of substances were subject to the same standardization scheme unless otherwise stated. Standardization comprised of keeping only the largest compound of each substance and was performed in KNIME [24] with the help of CDK nodes [25, 26, 27, 28]. Bemis-Murcko scaffolds (BMSs) [15] were generated for molecules using RDKit [29] cheminformatics suite and RDKit KNIME nodes [30]. Molecule structures were depicted with RDKit and ChemAxon's Marvin Sketch [31].

2.5 K-Nearest-Neighbor Analysis

Using the RDKit implementation of Morgan algorithm [32, 29], Morgan-fingerprint was generated for compounds with parameters of radius = 3 and fingerprint length = 2,048. The k-Nearest-Neighbors (KNNs) were identified for query compounds with the help of computing the Tanimoto-similarity coefficient of pairs of compounds. In this study the value of k was set to 5.

2.6 Distance Measure in Embedded $2D$ Space

The distance of compounds i, j mapped to a PHC can be quantified as the difference of the respective bin indices b_i and b_j. This distance can be referred-to-as *rank distance*, i.e. d_r (see: *Eq. 7*).

$$d_r(i,j) = |b_i - b_j| \tag{7}$$

However, the idea of an intuitive embedding into $2D$ suggests that structural proximity of compounds should be reflected in proximity of $2D$ coordinates. Therefore, given the nature of the HCASE method, it is possible to define a perceived distance measure of the compounds in the embedded space as detailed below.

Compounds embedded in $2D$ using the HCASE method are mapped to a latent grid. Each point of the grid represent a specific BMS or a group of BMSs, depending on the size of the reference BMS set and the z parameter. Therefore, the distance of two embedded compounds i, j "stretched" on this grid can be perceived as their Chebyshev-distance [33] (see: *Eq 8*). Of note, the Chebyshev-distance is a metric. However, since it is applied as a perceived distance measure, in this study we will refer to the Chebyshev-distance metric as Chebyshev-distance measure.

$$d_C(i,j) = \max_n |i_n - j_n| \tag{8}$$

2.7 Quantifying Space Overlap Similarity of Different Embeddings

Given an embedding generated by the HCASE method, one can compute the number of compounds associated with a reference BMS. More precisely, one need to count the number of compounds mapped to the bin the respective BMS was assigned to. In the function of z the number of bins is provided by N (see: *Eq. 1*). This information can be condensed into an N-dimensional *embedding-vector*. In such vector, the value of each dimension reflects the number of compounds associated with a specific bin, which bin is actually a point in the latent grid behind the embedding.

Quantifying the similarity two embedding-vectors $\mathbf{A}$ and $\mathbf{B}$ can be performed in analogous manner to computing the similarity of two molecular count-fingerprints [34] with the help of a modified Tanimoto-similarity coefficient (see: *Eq. 9*) [35, 36, 37, 38].

$$\theta_{A,B} = \frac{\sum_{i=1}^{N} \mathbf{A}_i \mathbf{B}_i}{\sum_{i=1}^{N} \mathbf{A}_i^2 + \sum_{i=1}^{N} \mathbf{B}_i^2 - \sum_{i=1}^{N} \mathbf{A}_i \mathbf{B}_i} \tag{9}$$

2.8 Scaffold t-SNE Method

For the sake of comparison with the HCASE method, we implemented a variation of the t-SNE method. The modification involves the use of a reference scaffold set to serve as landscape objects for the embedding of molecules. This modification intends to convert the t-SNE method in a way that better represents a medicinal chemistry inspired embedding and that also enables a consistent embedding mechanism of molecules regardless of the input molecule set at hand.

In the first stage of this method, the embedding of a reference scaffold set is computed, and the Scaffold-Keys are computed for them. In the second stage, the

scaffold-keys of compounds to be embedded are computed. The d_{SK} is computed between each compound and reference scaffold. Compounds assume the embedded coordinates of the closest reference scaffold according to d_{SK}.

The modified t-SNE method will be referred-to-as *Scaffold t-SNE* method throughout the text. The pseudocode of the Scaffold t-SNE method is provided in "Appendix", adopting the same notions discussed in Section 2.2.

The Scaffold t-SNE analysis was performed at various perplexity-values of $\{5, 10, 20, 30, 40, 50\}$ which are considered as optimal [39, 7]. The rest of the parameter settings were left to default as defined by "SciKit" (Python library) implementation of the t-SNE algorithm [40, 41]. Note, that the default values of learning rate and the number of iterations are: 200 and $1,000$, respectively.

2.9 Input Data

Compound Libraries Compound libraries were collected from two sources: approved drugs of DrugBank database (version: 2.0.9) [42], and the CANVASS library [43]. These libraries are comprised of $2,073$ and 344 compounds, respectively.

ChEMBL Scaffolds A set of unique BMSs of size $63,783$ has been extracted from ChEMBL database (version: 24.1) [44] using the same procedure and KNIME workflow [45] that was used to derive the knowledge base of SmartGraph platform [46]. This set was derived from the set of all unique BMSs included in ChEMBL database based on the number of compounds they are associated with. That is, only BMSs were selected if they are connected to less than 100 and at least 5 unique compounds. Out of $63,783$ scaffolds, after processing by RDKit and deduplication by SKs, we identified $55,961$ unique BMSs.

Natural Products Scaffolds A set of natural products were extracted from the ChEMBL database (version: 23) consisting of $1,921$ compounds [43]. BMSs of these compounds were identified and their SKs were generated. Subsequently, the BMSs were deduplicate on the basis of the SKs, which resulted in a set of 546 scaffolds (NatProd scaffolds).

For the sake of reproducibility of the experiments, all source code and data used to perform the experiments are publicly available the source-code repository: https://github.com/ncats/hcase [23].

3 Results and Discussion

3.1 Clustering of Scaffolds Mapped on a Hilbert-Curve

First, the ChEMBL reference BMSs were ordered according to their SKs. Next, we sought to monitor the position of certain scaffolds as a result of the embedding process. To this end, we cherry-picked a set of BMSs in a way so that their ranks are separated by larger and smaller intervals (see: *Tab. 1* and *Fig. S3*). Additionally, the immediate 50 neighbors (in both directions) were also marked, with the respective colors. The maximal order of PHC to be used was determined by the size of the

ChEMBL reference scaffold set. A PHC of order $z = 8$ gives rise to a space that is defined by a latent grid of $65,536$ points, i.e. coordinates. The size of ChEMBL scaffold set $(55,961)$ is less than this value, but is larger than the number of potential coordinates in a space defined by a PHC of $z = 7$. Taken these in consideration, the order of PHCs employed in this investigation was varied in the range of $z = [2, 8]$.

As it was described in Section 2.1.3, the reference scaffolds are assigned to bins in the function of z. Consequently, low values of z give rise to a low-resolution latent grid, where many of the marked scaffolds are assigned only to a few grid points, as expected (see: *Fig. 4a-4c*). Increasing the value of z, i.e. the resolution of embedding, it can be seen that the marked BMSs start to separate well, giving rise to clusters, i.e. groups of closely-binned BMSs (see: *Fig. 4d-4g*).

Based on the results of the embedding, it can be seen that the HCASE method is able to produce clusters of varying granularity in the function of parameter z. This feature therefore provide opportunity to control the resolution of the embedding depending on the use case at hand. Furthermore, the position of clusters is the function of the bin index or the BMSs, and relative position of scaffolds is determined, due to the nature of PHCs. The stabilization property of PHCs is also demonstrated by the results. This is a consequence of PHC algorithm implementation and the fact that BMSs don't represent a continuum where the absolute value of scaffolds is known.

These findings support, that using the HCASE method, it is possible to develop an intuition for identifying the type of scaffolds, or group of scaffolds encoded by various segments of the embedded space. Therefore, we concluded that the properties of latent grid generated by HCASE method are adequate to serve as the basis for compound embedding.

3.2 Embedding of KNNs

Building on the promising results described in the previous section, we sought to analyze the embedding of a compound library with the help of ChEMBL reference scaffold set and the HCASE method. To this end, the embedding of the DrugBank data set was performed. The range of z values were identical to the range utilized in the previous section, considering that we used the same reference scaffold set, i.e. ChEMBL. To better understand the embedding process, we selected 5 molecules randomly from the DrugBank dataset and the $k = 5$ nearest neighbors of each was determined as described in Section 2.5.

We checked that the set of nearest neighbors (NNs) of 25 compounds and the 5 randomly selected query compounds constitute a distinct set which we found ideal for carrying out the analysis at hand. The list of query compounds, their NNs and the values of Tanimoto-similarity coefficients is provided in *Tab. 2* and *Fig. 1* in decreasing order of similarity.

Considering all data points, it can be seen in *Fig. 5* that the position of individual datapoints is stabilized with increasing order of the underlying PHC. Also, increasing values of z give rise to a finer-grained clustering of data points.

Regarding the KNNs, most of them are clustered closely to the query molecules, as expected, but some of them are place further away. For instance, at $z = 8$ we can make the following observations. In the case of query molecule DB04837, i.e. "blue" series, two of the NNs ("X", "V") are positioned farther from DB04837, which is explained by the more complex BMS present in those two NNs as compared to the rest of the series. Interestingly, the fifth NN ("Y") in the same series is co-positioned with the query compound DB01362 (color: aqua), but it can't be seen due to overlap of markers. The reason for this is that "Y" and DB01362 share the same BMS, i.e. the benzene ring. Consequently, they were mapped to the same reference scaffold hence positioned to the same coordinate in the embedded space.

Similar trends can be observed in the other NN series as well. Typically, when the BMSs of NNs differ in exocyclic groups, then they are embedded still relatively closely. However, when the BMSs differ by extra rings, then they will be place further away. This phenomenon can be explained by the ordering of scaffolds based on their SKs. These observations argue that the embedding results in clustering that matches closely the mindset of a medicinal chemists when analyzing chemotypes. For example, in the case of the "purple" series (query molecule: DB00977) most of the NNs in the series share the same or very similar BMS, except compound "L", whose BMS is more complex than that of other NNs, hence it is positioned further away from other members of the series. The peculiarity of this fact is more obvious when one considers the Taniomoto-similarity of the NNs to the query molecule in the "purple" series; compound "L" is the second NN of the query compound, still it is positioned the furthest from other compounds of the series. Separation of compound "L" from the rest of the series members would be considered correct from a medicinal chemist's view, as compound "L" has the most dissimilar BMS in that series compared to the other BMSs.

3.3 Embedding of Randomly Selected Compounds

In order to contrast the above findings, we selected 25 random molecules from the DrugBank dataset (see: *Fig. S4*) and compared their embedding with that of the NN series. In *Fig. S5*, the embedding of these 25 compounds is shown besides the embedding of the 5 query molecules of the previous experiment. As it can be seen, the embedding of the random set exhibit a much more reduced level of clustering as compared to the case of the NN series. While come clustering is present in this set, mainly contributed to the presence of benzene ring as the BMS in several compounds, the overall picture resembles a random distribution of the embedded coordinates.

In summary, the above findings demonstrate that it is possible with the HCASE method to embed compounds in a chemical space that is able to differentiate molecules based on chemotypes, and to provide a logical and intuitive arrangement of these chemotypes. Therefore, it can be argued that clustering emerging in the embedded space will be reflective of a medicinal chemist's analytical thinking.

3.4 Comparison of the Results of Different Embedding Outcomes

After concluding the HCASE method is able to generate intuitive embedding of a chemical library we intended to analyze how we can compare the outcome of different embeddings. This first required to investigate the effect of utilizing different scaffold reference sets, then to quantify how well different embedding results are aligned with each other.

To this end, we performed separately the embedding of the DrugBank and CANVASS libraries utilizing two different reference scaffold sets: ChEMBL and NatProd. As explained in Section 3.1 the upper limit of z depends on the size of the reference scaffold set at hand. We determined that this upper limit is $z = 8$ in case of the ChEMBL set. The NatProd scaffold reference set is comprised of 546 BMSs, hence the upper limit of z is 5.

3.4.1 Qualitative Comparison

First, let us consider the embeddings in the NatProd chemical space as shown in *Fig. S6*. The positions of compounds of both libraries are also distributed across all possible 16 coordinates at $z = 2$. At $z = 3$ the CANVASS compounds are assigned to only 55 coordinates, whereas in the case of DrugBank library all 64 potential coordinates are assigned to compounds. At higher z values, neither of the libraries are assigned to all coordinates, and they start to separate in the same chemical space (see: *Fig. 6a*).

In the case of the ChEMBL chemical space (see: *Fig. S7*) at $z = 2$, the coordinates associated with the embedded compounds of both libraries are distributed across all potential 16 coordinates. At $z = 3$, in the case of the CANVASS library, the compounds are only assigned to 43 different coordinates. However, the compounds of DrugBank dataset are assigned to all potential coordinates. At higher values of z, the overlap of the respective pairs of embeddings becomes less and less pronounced, i.e. the two dataset start to separate, as in the previous case (see: *Fig. 6b*).

Based on the qualitative comparison, it can be observed that the DrugBank dataset occupies larger portion of the embedded space. This is not surprising considering that CANVASS is a smaller library, and a less diverse one according to the results of the embeddings. Nevertheless, the overlap of the two libraries seems to be larger in the NatProd space. As seen, at $z = 4$ the CANVASS library is more spread-out in this space. Since this space is defined by scaffolds extracted from the natural products subset of ChEMBL, the CANVASS library indeed seems as a good representative of the natural product space. However, the drug molecules represent structures with BMSs that even better represent the underlying NatProd reference scaffold set. Considering that many drug molecules are natural product derivatives, and the presence of larger diversity in the DrugBank vs. the CANVASS library, the fair amount overlap in this space of the two libraries can be considered reasonable.

In the ChEMBL chemical space both libraries show clustering which becomes prominent at $z > 5$ values, although the clustering is more obvious in the case of CANVASS library. Drug molecules represent this chemical space also to a reasonable

degree, whereas the CANVASS molecules form "islands". These islands are mostly overlapping with members of the DrugBank library. Further, in this chemical space the unoccupied area is visible to a larger extent as compared to the NatProd space.

Based on the above findings, we concluded that the choice of the reference scaffold set influences the embedding in two major manner. First, the reference scaffold set serves as a perspective which the structural similarities are analyzed from. Accordingly, the embedding of CANVASS and DrugBank libraries paint a more similar picture in the NatProd space than in ChEMBL space. Second, the separation of structures can be promoted by the choice of the reference scaffold set.

3.4.2 Quantitative Comparison

In the previous section we investigated how the embeddings of two chemical libraries can be compared qualitatively. However, there can be cases when one might want to quantify the overlap (similarity) of two embeddings.

To this end, one of the natural solutions is provided by aggregating the number of compounds associated with each given coordinate in the embedded space. This information can be condensed to a heatmap, in which cells correspond to specific coordinates in the embedded space. The color of each cell is the function of the number of molecules assigned to the respective coordinate. This solution is shown in *Fig. 7*, which reflect the aggregated results of embedding the DrugBank and CANVASS libraries in the NatProd chemical space with the HCASE method at $z = 5$. The heatmap provide an intuitive way to quickly see which regions of the same chemical space are better covered by either of the libraries. Of note, the aggregated molecule counts were log_{10}-transformed to provide better visualization.

Beyond the graphical solution, it also possible to quantify the overlap of the embedding of two libraries by using a measure (θ) analogous to the Tanimoto-similarity coefficient of count-fingerprints, as described in Section 2.6. The results of quantifying the overlap of two libraries based on θ is provided in *Tab. 5*. The results confirm the qualitative observations that the overlap of the two datasets decreases with increasing values of z, i.e. by increasing the resolution of the embedding. At the highest resolution, the overlap is greater in the NatProd space than in the ChEMBL space, just as it was observed in the qualitative analysis. While the values of θ are quite small in most cases, still, it can be used to quantify the extent of overlap.

3.5 Perceived Distance in the Embedded $2D$ Space

The promise of utilizing a PHC for chemical space embedding is that the objects mapped to close proximity on the curve will also be embedded in the higher dimension space in close proximity. Therefore, we sought to explore whether those distance values translate in the embedded $2D$ space in a way that can be perceived as distance measure.

Considering that the reference scaffolds create a latent grid behind the embedded space, it seemed natural to investigate the relation between the rank-distances (d_r) of compounds and the Chebyshev-distances (d_C) of embedded coordinates (see: Section 2.6).

To this end, we first investigated the correlation of the two different types of distance measures with the help of the DrugBank and CANVASS compound libraries. First, the correlation was determined by taking into account all compounds per dataset. Results are shown in *Tab. 4*. It can be seen that there is a reasonable level of correlation between d_r and d_C in the case of both datasets. Furthermore, higher values of parameter z tended to results in slightly higher correlation of the two distance measures as compared to lower z values. The highest correlation was found to be 0.73 and 0.72 for the DrugBank and CANVASS datasets, respectively, when using the ChEMBL reference scaffold set. The highest correlation was found to be 0.65 in the case of DrugBank dataset and NatProd reference scaffold set combination. Interestingly, in the case of CANVASS dataset and NatProd reference scaffold set the the correlation values were lower as compared to other data series, resulting in a maximum value of 0.59 at $z = 2$. This might be an indication that the underlying latent grid has limited capacity to distinguish between chemotypes.

To further support these finding, we generated non-overlapping sets of randomly selected compounds from the DrugBunk dataset. Each set was comprised of 100 compounds. The mean and standard deviation of the correlation between the two distance measures is provided in *Tab. 3*. Similarly to the previous findings, the correlation tended to increase slightly with increasing values of z. Moreover, the maximal value of mean correlation was equal to the maximal correlation observed in utilizing the entire dataset, in both cases of using the ChEMBL and NatProd reference scaffold sets.

In summary, there is a reasonably good correlation between the two distance measures d_r and d_C. Also, with increasing values of z the correlation tends to increases slightly, which is not surprising in the light of the converging nature of embedding PHCs of higher and higher order. Therefore, we propose that Chebyshev-distance measure can be considered as a perceived distance measure to quantify distances in the embedded $2D$ space generated by the HCASE method.

3.6 Comparison of HCASE Method with Prior Art

As the final experiment, we set forth to compare the HCASE method with prior art. Considering that the primary feature of the t-SNE algorithm is to preserve neighborhood information of objects in the embedded space, we decided to use this method for comparison. However, in order to obtain a meaningful comparison with the HCASE method it was necessary to modify the original t-SNE algorithm to some extent as described in Section 2.8 based on considerations discussed in Section 2.1.1.

The first question we sought to answer was how the embedding of a reference scaffold set created by the Scaffold t-SNE method compares to that of generated by the HCASE method. To this end, the embedding of ChEMBL reference scaffold set using the t-SNE algorithm was generated. We highlighted the same set of cherry-picked scaffolds that was described in Section 3.1, preserving the coloring scheme. Of note, the t-SNE embedding operates on the Morgan-fingerprints of the ChEMBL reference set. As shown on *Fig. 8a*, the position of the clusters belonging

to the original clusters, is far more scattered as compared to HCASE embedding, although certain level of clustering can be observed. Further, the logic regarding the relative positioning of scaffolds is not transparent, therefore it is difficult to intuitively interpret the resultant chemical space produced by t-SNE embedding. As shown on *Fig. S8* this pehonomenon was observed across a range of perplexity values that were suggested as optimal (see: Section 2.8).

Next, we sought to investigate Scaffold t-SNE embedding of the same set of 5 randomly selected molecule and their $k = 5$ nearest neighbors that were described in Section 3.2, using the embedding of the reference scaffolds in the previous step. Interestingly, a high level of clustering can be observed in all KNN-series that is comparable to that produced by the HCASE method (see: *Fig. 8b*). The reason for this is that Scaffold t-SNE method takes advantage of the predefined chemical space of reference scaffolds, and using scaffold-keys the closest reference point of each compound is identified. Therefore, the embedding will reflect the differences and similarities of chemotypes to a great degree. This was observation was true in the case of all the applied perplexity values (see: *Fig. S9*). The embeddings produced by these two methods differ in three major standpoints.

First, relative placement of KNN series, even just focusing on the query compounds, cannot be explained with a simple organizing principle. Indeed, they follow the placement of the reference clusters determined by the t-SNE algorithm. However, the t-SNE algorithm does not guarantee a well-defined layout of embedded objects as opposed to the HCASE algorithm.

Second, the (relative) position of the coordinates of the embedded molecules produced by the Scaffold t-SNE method does not seem to converge, i.e. to stabilize, by varying the value of perplexity parameter. This feature of the Scaffold t-SNE method does not promote the intuitive interpretation of the results, and is in great contrast with the converging property of embedded coordinates produced by the HCASE method.

Finally, given a reference scaffold set, changing the resolution of the Scaffold t-SNE embedding is challenging, at best. If one wants to use a smaller subset of the reference scaffold set, then the Scaffold t-SNE method would require the creation of a new t-SNE space based on the subset of the scaffolds. However, there is no guarantee that the new space will resemble the original t-SNE space.

Of note, in this study only the perplexity parameters were varied when using the t-SNE algorithm. While more refined values might be obtained by thorough hyper-parameter search, the result will not influence the three major differences discussed above. Furthermore, these observations would translate to any other prior art method. Even, if they would be modified in an analogous manner as the Scaffold t-SNE method was derived from the original t-SNE method.

It can be concluded from the comparison of the two methods that existing space-embedding methods can be modified successfully to produce embeddings with reasonable clustering properties for chemotypes. While the clustering properties of such methods can be on par with that of the HCASE method, the HCASE method provides a clear advantage for interpretability.

4 Conclusions

In this proof-of-concept study we present a HCASE space embedding method that stands out from existing methods by its unique ability to produce an embedding that can be easily interpreted by medicinal chemists and data analysts. The novelty of the method is to create a well-defined latent grid of reference scaffolds, where the scaffolds are organized by increasing structural complexity. This is achieved by mapping the reference scaffolds based on their scaffold keys to a pseudo-Hilbert-Curve that can be readily embedded into higher dimensional space according to a well-established algorithm. Compounds are subsequently embedded into this grid based on their proximity to reference scaffolds measured by sScaffold-Key distances.

With the help of a series of experiments, we demonstrated that the HCASE method indeed meets all the criteria we set forth for an intuitive space embedding method. Namely, the embedding is able to cluster related chemotypes, and to lay out the chemotypes in a logical order in the embedded space. The ability to use a reference scaffold set to define a chemical space assures that independent compound libraries can be embedded into the same space in a consistent manner. This allows for direct comparison of the embeddings of different datasets visually, qualitatively and quantitatively, as long as the underlying reference scaffold set was the same. Furthermore, the HCASE method is able to generate a series of embeddings with increasing resolutions. In these series the positions of compounds converge as the resolution increases, which is not a property that has been accomplished by theother methods. We have also demonstrated that it is possible to quantify the distances between the embedded points in the HCASE space by computing the pairwise Chebyshev-distance values.

The chemotype-clustering ability of HCASE method was characterized with the help of two reference scaffold sets (ChEMBL: 63,783 scaffolds, NatProd: 546 scaffolds) and two compound libraries (DrugBank: 2,073 compounds, CANVASS: 344 compounds). The analysis of embedding KNN series has shown that HCASE method is able to cluster closely related structures in the embedded space. As expected, the degree of clustering was higher in the KNN series as compared to a series of randomly selected molecules. Also, we compared the overlap of the HCASE ebbending of the two compound libraries in two different reference scaffold set spaces. The results demonstrated that reference scaffold sets can be used to define a perspective for embedded space comparison, e.g. to compare embeddings in a natural product space. Furthermore, we provided the means to compare HCASE embeddings quantitatively.

Finally, we compared the properties of space embeddings generated by HCASE method and a prior art method, which was modified for the sake of meaningful comparison. We found that the clustering performance of the modified prior art method was nearly as good as that of the HCASE method. However, the results of the HCASE method can be easily interpreted from a medicinal chemistry point of view, unlike the results of the other method.

In conclusion, the presented HCASE method is attributed with novel and unique characteristics that can render it as a desirable data reduction and clustering method in any research setting where medicinal chemistry perspective is essential.

5 Outlook

In light of the structurally interpretable property of the HCASE method, it would be a natural extension to create interactive visualization of results. That is, when selecting a region of interest on the embedding plot, the underlying scaffold(s) could be visualized in an application to provide more structural context for the position of embedded compounds. Furthermore, inspired by SOM and GTM method, it might be helpful to quantify how well the chemotype of an embedded compound matches that of the reference scaffolds associated with that position. This property might be the mean of distances computed between a given compound and the reference scaffolds associated with its position.

Appendix

The following remarks needed to be made regarding the in-house implementation of Scaffold-Key algorithm [19, 23]. When a specific key is referred in the remarks then the key ID reflects the order of the keys as published by *Ertl* [19].

- Rings were determined as "smallest set of simple rings for a molecule" (SSSR) as implemented by RDKit [29].
- The following atoms were considered as heteroatoms in implementing the Scaffol-Key algorithm: Li, Be, B, N, O, F, Mg, Al, Si, P, S, Cl, Zn, As, Se, Br, Te, I, Pt, Hg, Mn, Fe, Co, Ni, Cu, Ga, Ge, Rh, Pd, Ag, Cd, Sn. Relevance: many keys related to heteroatoms.
- Definition of "multiple linker" was not provided by the Scaffold-Key algorithm. Hence, we quantified this property as the number of bonds associated with the branched linker atom. Relevance: key 19.
- In fully conjugated rings, the number of bonds was determined as follows. In RDKit, the conjugated system is associated with an ID. If all members of a ring are assigned to the same ID then the ring was considered as fully conjugated, otherwise as not fully conjugated. Relevance: key 7, 8.
- Number of multiple bonds in not fully conjugated rings: double, triple and aromatic bonds were all counted. Relevance: key 8.
- Exocyclic atoms: atoms connected to rings with double bond. Relevance: key 31.
- Exolinker atoms: atoms connected to linker substructure with double bond (exolinker is not part of the linking substructure). Relevance: key 31.
- Heteroatoms associated with more than two bonds: total number of non-hydrogen bonds for heteroatom. Relevance: key 32.

Pseudocode of the Scaffold t-SNE Method

Algorithm 2 Scaffold t-SNE Method

Input: set of molecules M
Input: set of reference Bemis-Murcko scaffolds S

Variable: molecule mol
Variable: scaffold-key sk
Variable: list of (S, sk)-tuples S_{sk}
Variable: list of (sk, b)-tuples S_{bin}
Variable: Bemis-Murcko scaffold bms
Variable: ($x \in \mathbb{N}$, $y \in \mathbb{N}$)-tuple P
Variable: list of P items T
Variable: list of (sk, P)-tuples $E_{scaffold}$
Variable: fingerprint fp
Variable: list of fingerprints F
Variable: list of (mol, T_{coord}) E

for all S_i in S **do**
  sk := **getScaffoldKey** (S_i)
  S_{sk}.add (S_i, sk)
  fp := generateFP(S_i)
  F.add(fp)
end for

S_{sk} := sort S_{sk} alphanumerically by SK in increasing order
S_{sk} := deduplicate S_{sk} by sk, keep first instance of identical tuples
T := tSNE (F)

$E_{scaffold}$:= (S_{sk}, T)

for all m_i in M **do**
  bms := **getBemisMurckoScaffold** (m_i)
  sk := **getScaffoldKey** (bms)
  P := **getClosestRefererenceBMSCoordinate** (sk, $E_{scaffold}$)

  E.add (m_i, P)

end for
return (E)

int *function* **getClosestRefererenceBMSCoordinate**(sk, $E_{scaffold}$)
($x \in \mathbb{N}$, $y \in \mathbb{N}$)-tuple P_mol
Variable: numeric d_{sk}
Variable: list of (P, d_{sk})-tuples D

for all e_i in $E_{scaffold}$ **do**
  d_{sk} := **getSKDistance** (sk, e_i[0])
  D.add (e_i[1], d_{sk})
end for
D := sort D by d_{sk} and b, both in increasing order

$P_{mol} = D$[0][0]

return (P_{mol})

Declarations

Competing interests

The authors declare that they have no competing interests.

Author's contributions

The idea of Hilbert-Curve Assisted Space Embedding (HCASE) method was conceived by Gergely Zahoránszky-Kőhalmi, PhD (GZK). GZK also designed and performed the experiments, wrote all the source code and the manuscript. Alexander G. Godfrey, PhD, lead of the "A Specialized Platform for Innovative Research Exploration (ASPIRE)" program at NCATS/NIH, and Kanny Wan, PhD provided inspiration for this study and contributed to finalizing the manuscript. The authors read and approved the final manuscript.

Acknowledgements

The authors are thankful to Sam Michael, PhD, Qian Zhu, PhD, Matt Hall, PhD and Min Shen, PhD for fruitful discussions.

Funding

This research was supported by the Intramural research program of the NCATS, NIH.

Document Template

The manuscript has been formatted with the help of the "BioMed Central TeX template - Version 0.6 (April 15th 2015)" [47, 48, 49, 50].

References

1. Hostelling, H.: Analysis of a complex statistical variables into principal components. Journal of Educational Psychology **24**(6), 417–441 (1933)
2. Quist, M., Yona, G.: Distributional scaling: An algorithm for structure-preserving embedding of metric and nonmetric spaces. J. Mach. Learn. Res. **5**, 399–420 (2004)
3. van der Maaten, L.J.P.: Learning a parametric embedding by preserving local structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP **5**, 384–391 (2009)
4. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. http://arxiv.org/abs/1802.03426v2
5. Kohonen, T.: SELF-ORGANIZING MAPS: OPHMIZATION APPROACHES. In: Artificial Neural Networks, pp. 981–990. Elsevier, ??? (1991). doi:10.1016/b978-0-444-89178-5.50003-8
6. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000). doi:10.1126/science.290.5500.2319. https://science.sciencemag.org/content/290/5500/2319.full.pdf
7. Distill: How to Use t-SNE Effectively. https://distill.pub/2016/misread-tsne/
8. Osolodkin, D.I., Radchenko, E.V., Orlov, A.A., Voronkov, A.E., Palyulin, V.A., Zefirov, N.S.: Progress in visual representations of chemical space. Expert Opinion on Drug Discovery **10**(9), 959–973 (2015). doi:10.1517/17460441.2015.1060216
9. Oprea, T.I., Gottfries, J.: Chemography: The art of navigating in chemical space. J. Comb. Chem. **3**(2), 157–166 (2001). doi:10.1021/cc0000388
10. Nguyen, K., Blum, L., van Deursen, R., Reymond, J.-L.: Classification of organic molecules by molecular quantum numbers. ChemMedChem **4**(11), 1803–1805 (2009). doi:10.1002/cmdc.200900317
11. Velkoborsky, J.: Hierarchical visualization of the chemical space. Master's thesis, Charles University, Department of Software Engineering, Prague, Czech Republic (2016)
12. Lin, A., Horvath, D., Afonina, V., Marcou, G., Reymond, J.-L., Varnek, A.: Mapping of the available chemical space versus the chemical universe of lead-like compounds. ChemMedChem **13**(6), 540–554 (2018). doi:10.1002/cmdc.201700561. https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.201700561
13. Naveja, J.J., Medina-Franco, J.L.: Finding constellations in chemical space through core analysis. Frontiers in Chemistry **7**, 510 (2019). doi:10.3389/fchem.2019.00510
14. Probst, D., Reymond, J.-L.: Visualization of very large high-dimensional data sets as minimum spanning trees. Journal of Cheminformatics **12**(1) (2020). doi:10.1186/s13321-020-0416-x
15. Bemis, G.W., Murcko, M.A.: The properties of known drugs. 1. molecular frameworks. J. Med. Chem. **39**(15), 2887–2893 (1996). doi:10.1021/jm9602928
16. Hilbert, D.: Über die stetige abbildung einer linie auf ein flächenstück. Mathematische Annalen **38**, 459–460 (1891)
17. Sanderson, G.: Hilbert's Curve: Is infinite math useful? https://www.youtube.com/watch?v=3s7h2MHQtxc&t=798s
18. Moon, B., Jagadish, H.V., Faloutsos, C., Saltz, J.H.: Analysis of the clustering properties of the hilbert space-filling curve. IEEE Transactions on Knowledge and Data Engineering **13**(1), 124–141 (2001). doi:10.1109/69.908985
19. Ertl, P.: Intuitive ordering of scaffolds and scaffold similarity searching using scaffold keys. J. Chem. Inf. Model. **54**(6), 1617–1622 (2014). doi:10.1021/ci5001983
20. Python library: Hilbert-Curve. https://pypi.org/project/hilbertcurve/

21. Hilbert-Curve Implementation Details. https://stackoverflow.com/questions/499166/mapping-n-dimensional-value-to-a-point-on-hilbert-curve
22. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: Inchi, the iupac international chemical identifier. Journal of Cheminformatics **7**(1), 23 (2015)
23. Hilbert-Curve Assisted Space Embedding (HCASE) Method Source Code Repository. https://github.com/ncats/hcase
24. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer, ??? (2007)
25. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.: The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. Journal of Chemical Information and Computer Sciences **43**(2), 493–500 (2003). doi:10.1021/ci025584y
26. Willighagen, E.L., Mayfield, J.W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C.T., Guha, R., Steinbeck, C.: The chemistry development kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics **9**(1) (2017). doi:10.1186/s13321-017-0220-4
27. The Chemistry Development Kit (CDK). https://github.com/cdk/cdk
28. CDK Nodes for KNIME. https://www.knime.com/community/cdk
29. Rdkit: Open-source cheminformatics software. http://www.rdkit.org
30. RDKit Nodes for KNIME. https://www.knime.com/rdkit
31. Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions, Marvin 16.1.25, 2016, ChemAxon. http://www.chemaxon.com
32. Morgan, H.L.: The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of Chemical Documentation **5**(2), 107–113 (1965). doi:10.1021/c160017a018
33. Cantrell, C.D.: Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press. Cambridge University Press, ??? (2000)
34. Heritage, J.R.H.W.: Molecular Hologram QSAR. US5751605A, 1996
35. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société vaudoise des sciences naturelles **37**, 547–579 (1901)
36. Tanimoto, T.T.: Tech. rep., ibm internal report. Technical report (I957)
37. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. Journal of Chemical Information and Computer Sciences **38**(6), 983–996 (1998). doi:10.1021/ci9800211
38. Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics **7**(1) (2015). doi:10.1186/s13321-015-0069-3
39. van der Maaten, L.: t-SNE. https://lvdmaaten.github.io/tsne/
40. Python Core Team. Python: A dynamic, open source programming language. Python Software Foundation. (2015). https://www.python.org/
41. SciKit-Learn Python Library. https://scikit-learn.org/stable/
42. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S.: DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research **42**(D1), 1091–1097 (2013). doi:10.1093/nar/gkt1068
43. Kearney, S.E., Zahoránszky-Kőhalmi, G., Brimacombe, K.R., Henderson, M.J., Lynch, C., Zhao, T., Wan, K.K., Itkin, Z., Dillon, C., Shen, M., Cheff, D.M., Lee, T.D., Bougie, D., Cheng, K., Coussens, N.P., Dorjsuren, D., Eastman, R.T., Huang, R., Iannotti, M.J., Karavadhi, S., Klumpp-Thomas, C., Roth, J.S., Sakamuru, S., Sun, W., Titus, S.A., Yasgar, A., Zhang, Y.-Q., Zhao, J., Andrade, R.B., Brown, M.K., Burns, N.Z., Cha, J.K., Mevers, E.E., Clardy, J., Clement, J.A., Crooks, P.A., Cuny, G.D., Ganor, J., Moreno, J., Morrill, L.A., Picazo, E., Susick, R.B., Garg, N.K., Goess, B.C., Grossman, R.B., Hughes, C.C., Johnston, J.N., Joullie, M.M., Kinghorn, A.D., Kingston, D.G.I., Krische, M.J., Kwon, O., Maimone, T.J., Majumdar, S., Maloney, K.N., Mohamed, E., Murphy, B.T., Nagorny, P., Olson, D.E., Overman, L.E., Brown, L.E., Snyder, J.K., Porco, J.A., Rivas, F., Ross, S.A., Sarpong, R., Sharma, I., Shaw, J.T., Xu, Z., Shen, B., Shi, W., Stephenson, C.R.J., Verano, A.L., Tan, D.S., Tang, Y., Taylor, R.E., Thomson, R.J., Vosburg, D.A., Wu, J., Wuest, W.M., Zakarian, A., Zhang, Y., Ren, T., Zuo, Z., Inglese, J., Michael, S., Simeonov, A., Zheng, W., Shinn, P., Jadhav, A., Boxer, M.B., Hall, M.D., Xia, M., Guha, R., Rohde, J.M.: Canvass: A crowd-sourced, natural-product screening library for exploring biological space. ACS Central Science **4**(12), 1727–1741 (2018). doi:10.1021/acscentsci.8b00747
44. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J.P.: The ChEMBL bioactivity database: an update. Nucleic Acids Research **42**(D1), 1083–1090 (2013). doi:10.1093/nar/gkt1031
45. SmartGraph Backend Source Code Repository. https://github.com/ncats/smartgraph_backend
46. Zahoránszky-Kőhalmi, G., Sheils, T., Oprea, T.I.: SmartGraph: a network pharmacology investigation platform. Journal of Cheminformatics **12**(1) (2020). doi:10.1186/s13321-020-0409-9
47. Lamport, L.: LATEX : a Document Preparation System. Addison-Wesley Pub. Co., ??? (1986)
48. BioMed Central TeX template - Version 0.6 (April 15th 2015). http://media.biomedcentral.com/content/production/bmc_article-tex.zip
49. BioMed Central TeX template - Version 0.6 (April 15th 2015) - article package. http://www.ctan.org/pkg/article
50. BioMed Central TeX template - Version 0.6 (April 15th 2015) - amsart package. http://www.ctan.org/pkg/amsart

Table 1: Cherry-picked BMSs of the ChEMBL Reference Scaffold Set.

Cherry-Picked Reference Scaffold Rank	Color
$5,000$	blue
$15,000$	orange
$16,000$	green
$25,000$	red
$26,000$	purple
$35,000$	brown
$44,000$	pink
$45,000$	gray
$55,000$	yellow-green

Table 2: $K = 5$ nearest neighbors of 5 randomly selected drug molecules. Fingerprint: Morgan (radius=3, lenght=2048).

C_{query}	C_{NN}	rank	sim
DB00006	DB04931	1	0.44
DB00006	DB01284	2	0.41
DB00006	DB00050	3	0.39
DB00006	DB09067	4	0.38
DB00006	DB06825	5	0.37
DB00849	DB01174	1	0.49
DB00849	DB00794	2	0.44
DB00849	DB05246	3	0.32
DB00849	DB01437	4	0.32
DB00849	DB00252	5	0.28
DB00977	DB01357	1	0.70
DB00977	DB04575	2	0.63
DB00977	DB00655	3	0.51
DB00977	DB00783	4	0.51
DB00977	DB04573	5	0.50
DB01362	DB01249	1	0.88
DB01362	DB09135	2	0.86
DB01362	DB09134	3	0.62
DB01362	DB09313	4	0.27
DB01362	DB01578	5	0.22
DB04837	DB11609	1	0.35
DB04837	DB00257	2	0.31
DB04837	DB00333	3	0.27
DB04837	DB01231	4	0.26
DB04837	DB08944	5	0.26

Table 3: Correlation of Chebyshev-distances and SK-rank distances in embedded subsets of DrugBank dataset.

Reference Scaffold Set	z	Correlation - Mean	Correlation - Std
ChEMBL	2	0.70	0.03
ChEMBL	3	0.72	0.03
ChEMBL	4	0.72	0.03
ChEMBL	5	0.73	0.03
ChEMBL	6	0.73	0.03
ChEMBL	7	0.73	0.03
ChEMBL	8	0.73	0.03
NatProd	2	0.64	0.04
NatProd	3	0.63	0.04
NatProd	4	0.65	0.03
NatProd	5	0.65	0.03

Tables

Table 4: Correlation of Chebyshev–distances and SK–rank distances.

Dataset	Reference Scaffold Set	z	Correlation
DrugBank	ChEMBL	2	0.70
DrugBank	ChEMBL	3	0.72
DrugBank	ChEMBL	4	0.72
DrugBank	ChEMBL	5	0.73
DrugBank	ChEMBL	6	0.73
DrugBank	ChEMBL	7	0.73
DrugBank	ChEMBL	8	0.73
DrugBank	NatProd	2	0.63
DrugBank	NatProd	3	0.63
DrugBank	NatProd	4	0.65
DrugBank	NatProd	5	0.64
CANVASS	ChEMBL	2	0.65
CANVASS	ChEMBL	3	0.72
CANVASS	ChEMBL	4	0.72
CANVASS	ChEMBL	5	0.72
CANVASS	ChEMBL	6	0.72
CANVASS	ChEMBL	7	0.72
CANVASS	ChEMBL	8	0.72
CANVASS	NatProd	2	0.60
CANVASS	NatProd	3	0.56
CANVASS	NatProd	4	0.57
CANVASS	NatProd	5	0.58

Table 5: Space overlap between DrugBank and CANVASS libraries in different chemical spaces.

Reference Scaffold Set	z	θ
ChEMBL	2	0.20
ChEMBL	3	0.16
ChEMBL	4	0.11
ChEMBL	5	0.09
ChEMBL	6	0.06
ChEMBL	7	0.05
ChEMBL	8	0.05
NatProd	2	0.19
NatProd	3	0.13
NatProd	4	0.09
NatProd	5	0.07

Figures

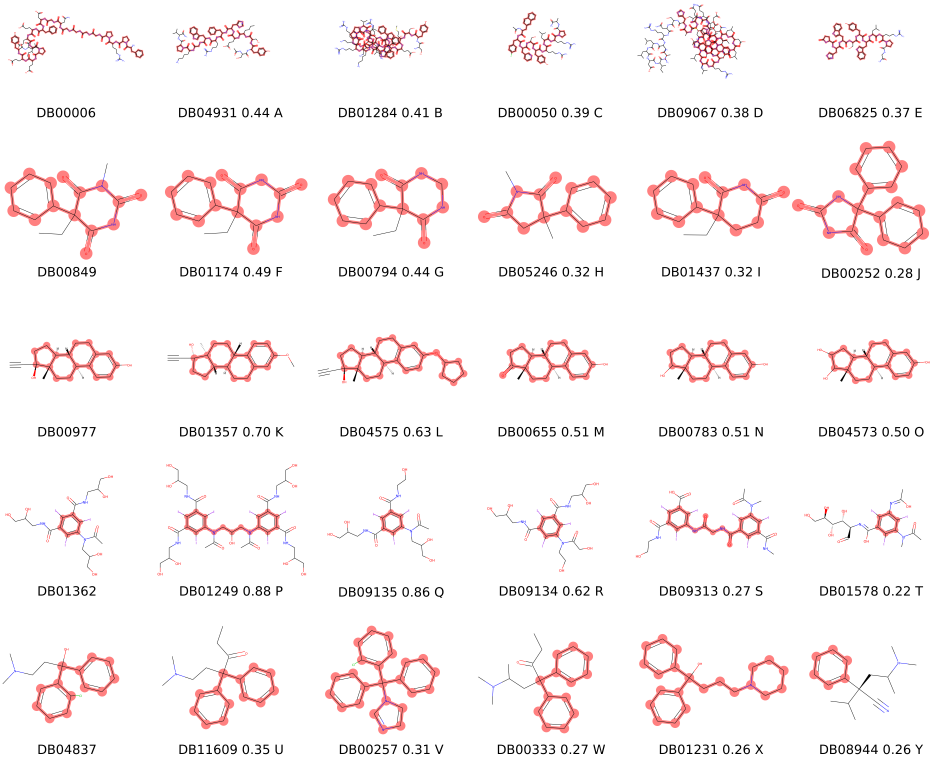

Figure 1: **KNNs of randomly selected molecules.** First column contains the query structures and subsequent columns contain the $k = 5$ NNs in decreasing order of similarity. Tanimoto-similarity was computed using Morgan-fingerprints, radius = 3, length = 2048. The value of Tanimoto-similarity coefficient and the label of compounds are shown after the compound IDs for NNs. The BMSs of compounds are highlighted by red.

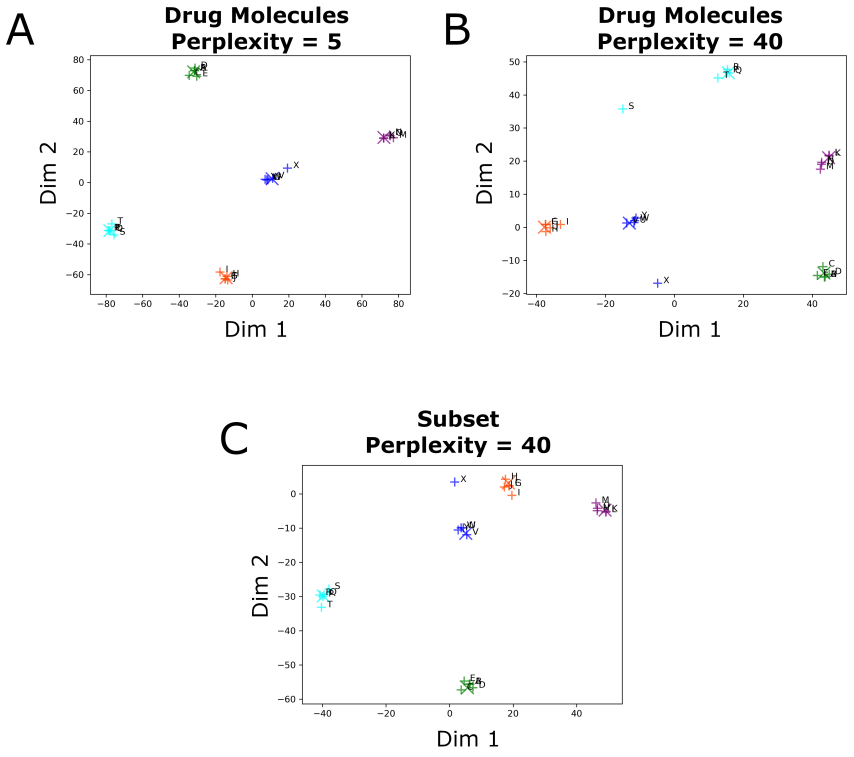

Figure 2: **Maps generated by *t*-SNE Analysis of Drug Molecules.** Embedding of DrugBank molecules performed by the original *t*-SNE algorithm at various perplexity values and repeating the embedding with a subset of drug molecules. The randomly selected five molecules are marked by enlarged (x) symbol. Green: DB00006, orange: DB00849, purple: DB00977, aqua: DB01362, blue: DB04837. The NNs of each molecule are indicated by (+) symbol with matching color. Molecules are labeled according to *Fig. 1.* **A**) Drug molecules, perplexity = 5. **B**) Drug molecules, perplexity = 40. **C**) Subset, perplexity = 40.

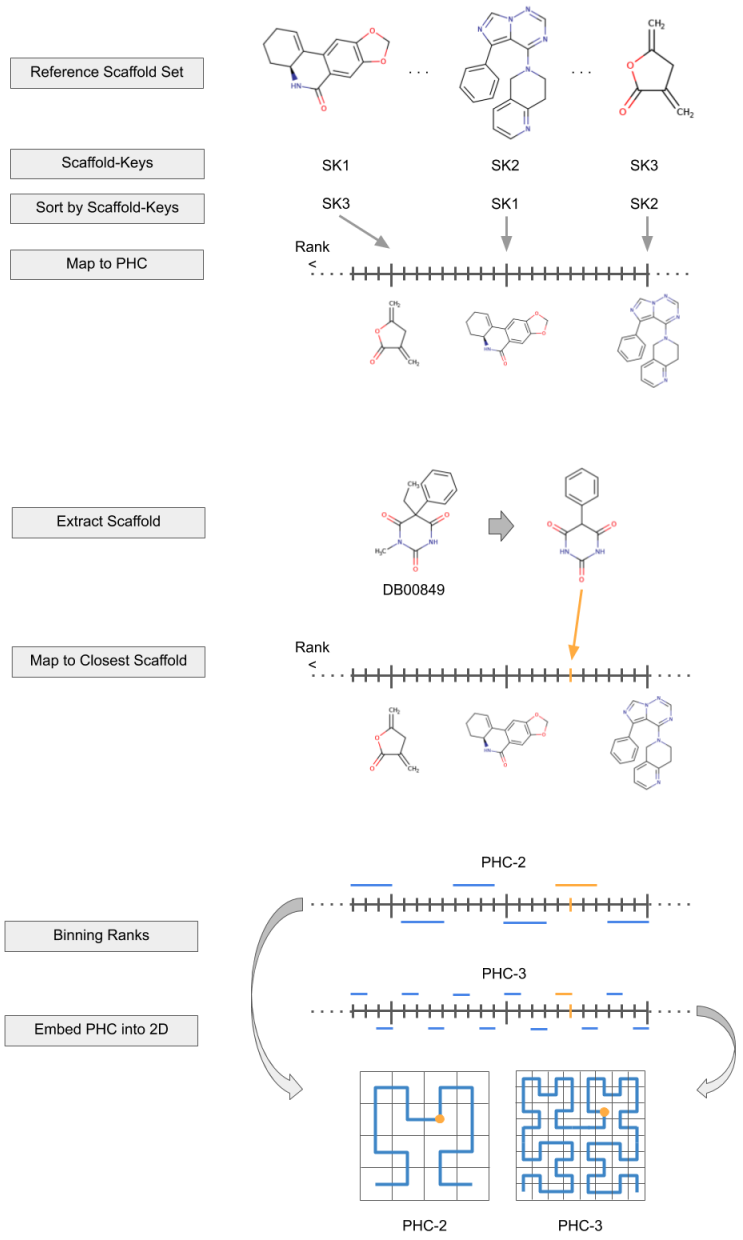

Figure 3: **HCASE method.** The process of embedding compounds into a chemical space with the HCASE method is demonstrated. The chemical space is defined by a reference scaffolds which are ordered based on their Scaffold-Keys (SK). The HCASE method maps the reference scaffolds on a series of PHCs of increasing order. Then, a compound of the library to be embedded are mapped to its closest scaffold based on their Scaffold-Key distances (d_{SK}). A binning step is also included in the process to make sure that each of the reference scaffolds, hence each compound, can be mapped to one of the possible coordinates in the higher dimension space. The number of possible coordinates is influenced by the order of the PHC the scaffolds are mapped to. A compound highlighted by *yellow* is tracked in this process. As it can be seen, the position of the compound in a $2D$ space is function of the order of the PHC it was mapped to. Due to the nature of PHCs the position of compounds converges to a "stable" position when increasing the order of PHCs.

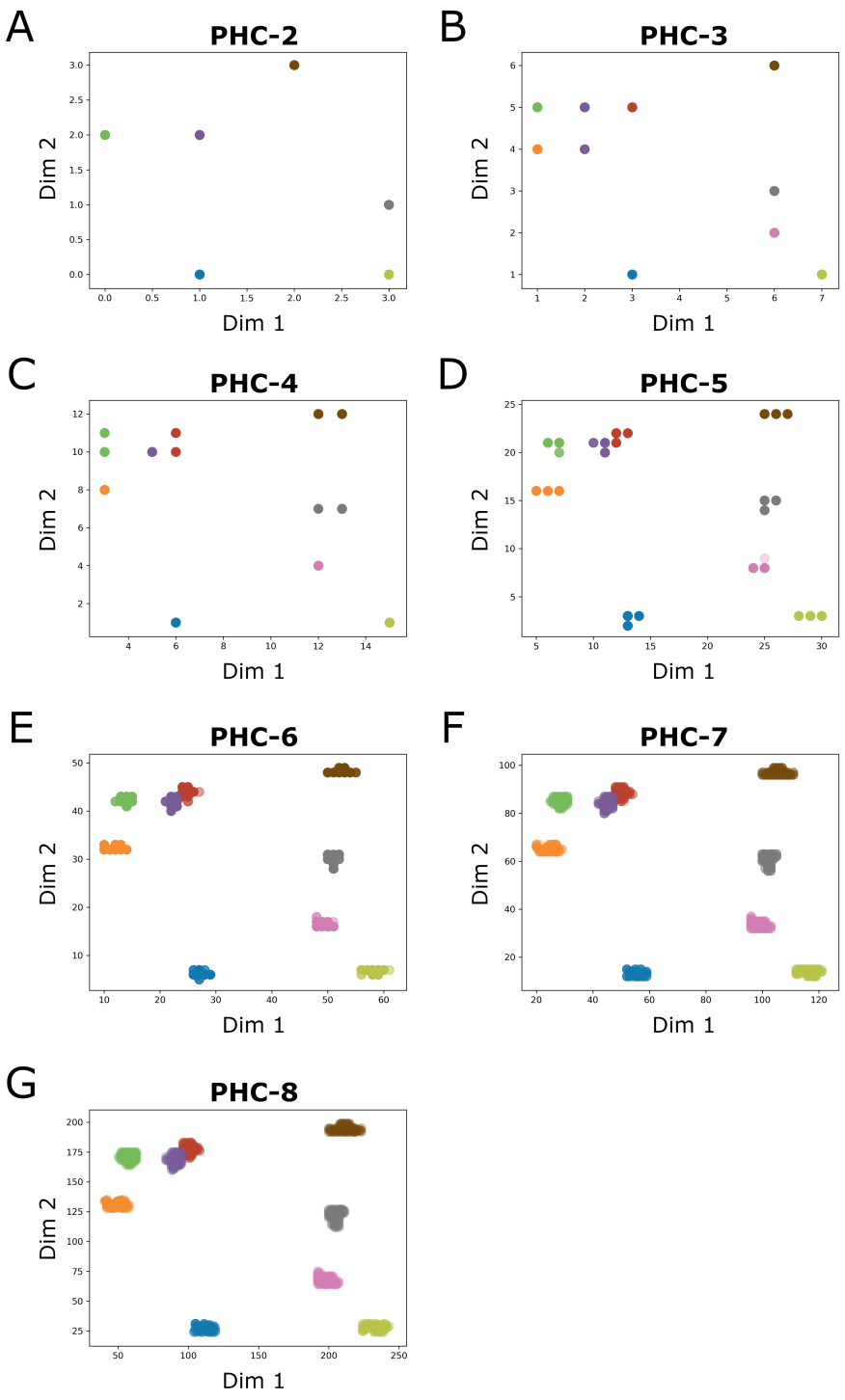

Figure 4: **Embedding of Cherry-Picked Scaffolds.** The embedding of several BMSs and their neighborhood is tracked. The cherry-picked scaffolds and their respective colors are provided in *Tab.1.*

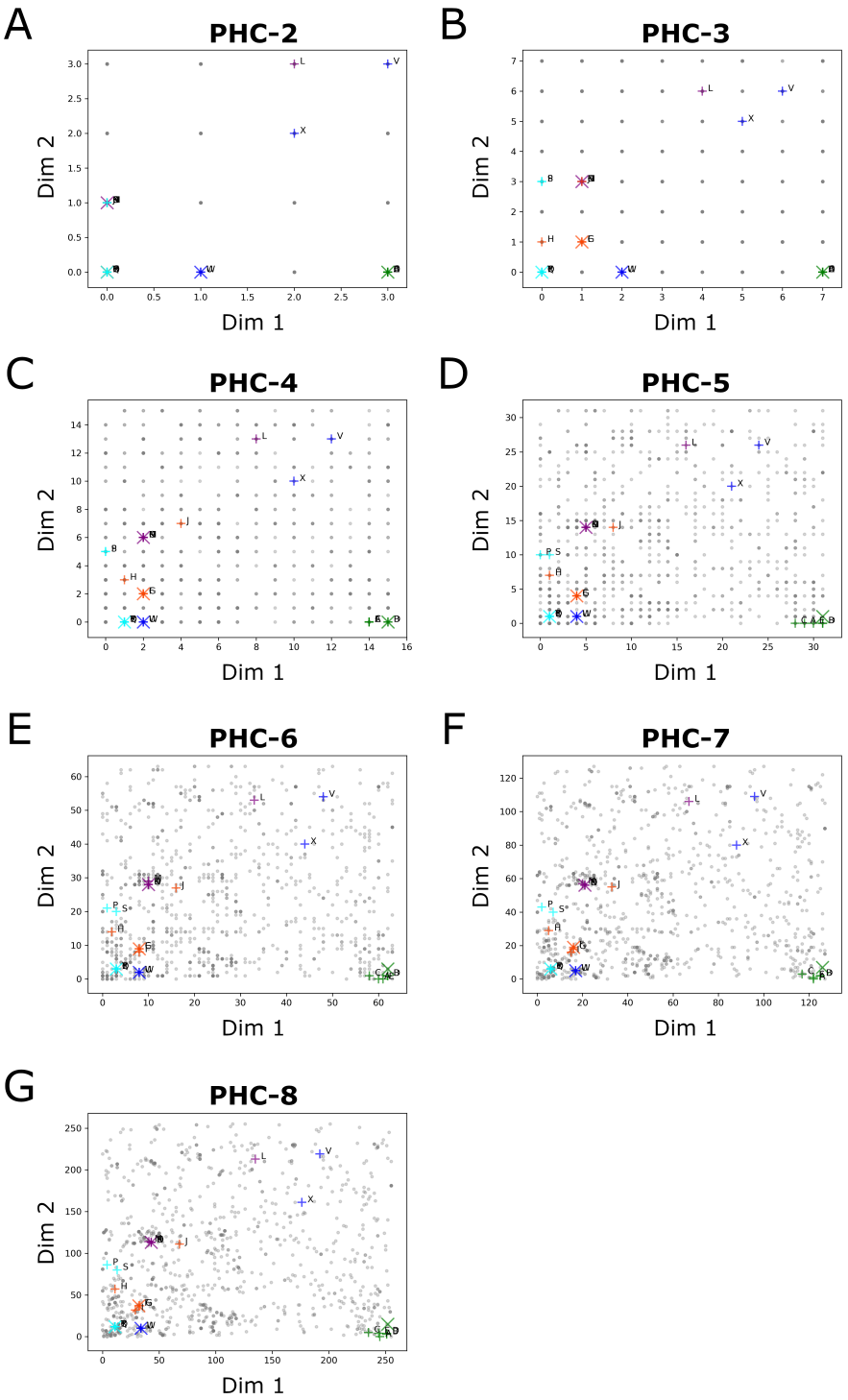

Figure 5: **KNN Analysis.** Shown is the embedding of $k = 5$ nearest neighbors of 5 randomly selected compounds from the DrugBank dataset. The order of PHC utilized for structure embedding is indicated by suffix. Enlarged (x) signs indicate the query compound of KNN analysis; green: DB00006, orange: DB00849, purple: DB00977, aqua: DB01362, blue: DB04837. (+) signs indicate the NNs of a query compound with identical color. Gray circles indicate other DrugBank compounds. Compounds are labeled according to *Fig. 1.*

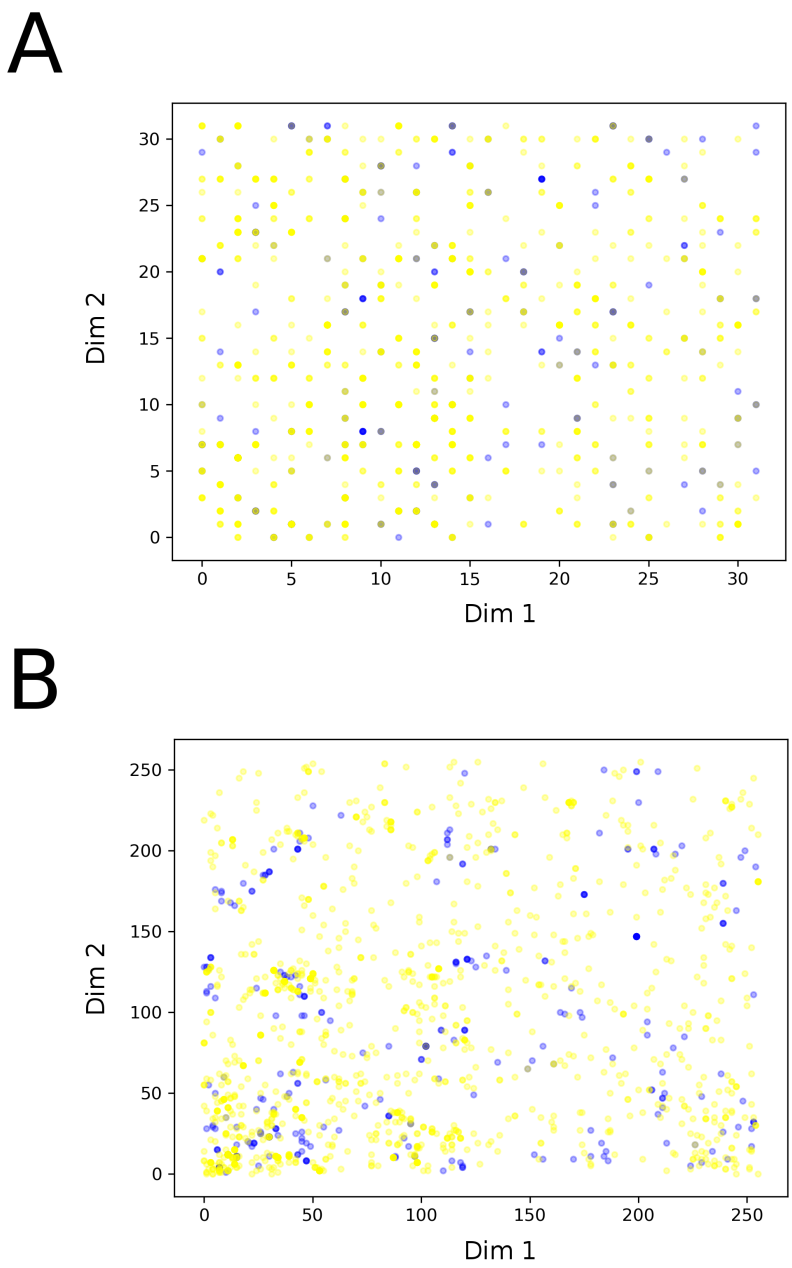

Figure 6: **Comparing Embeddings in Natural Product and ChEMBL Scaffold Space.** Blue: CANVASS compounds, yellow: drugs. Overlapping datapoints are colored by green-brown color due to the transparency of the datapoints. **A**) NatProd Scaffold Space, PHC-5. **B**) ChEMBL Scaffold Space, PHC-8.

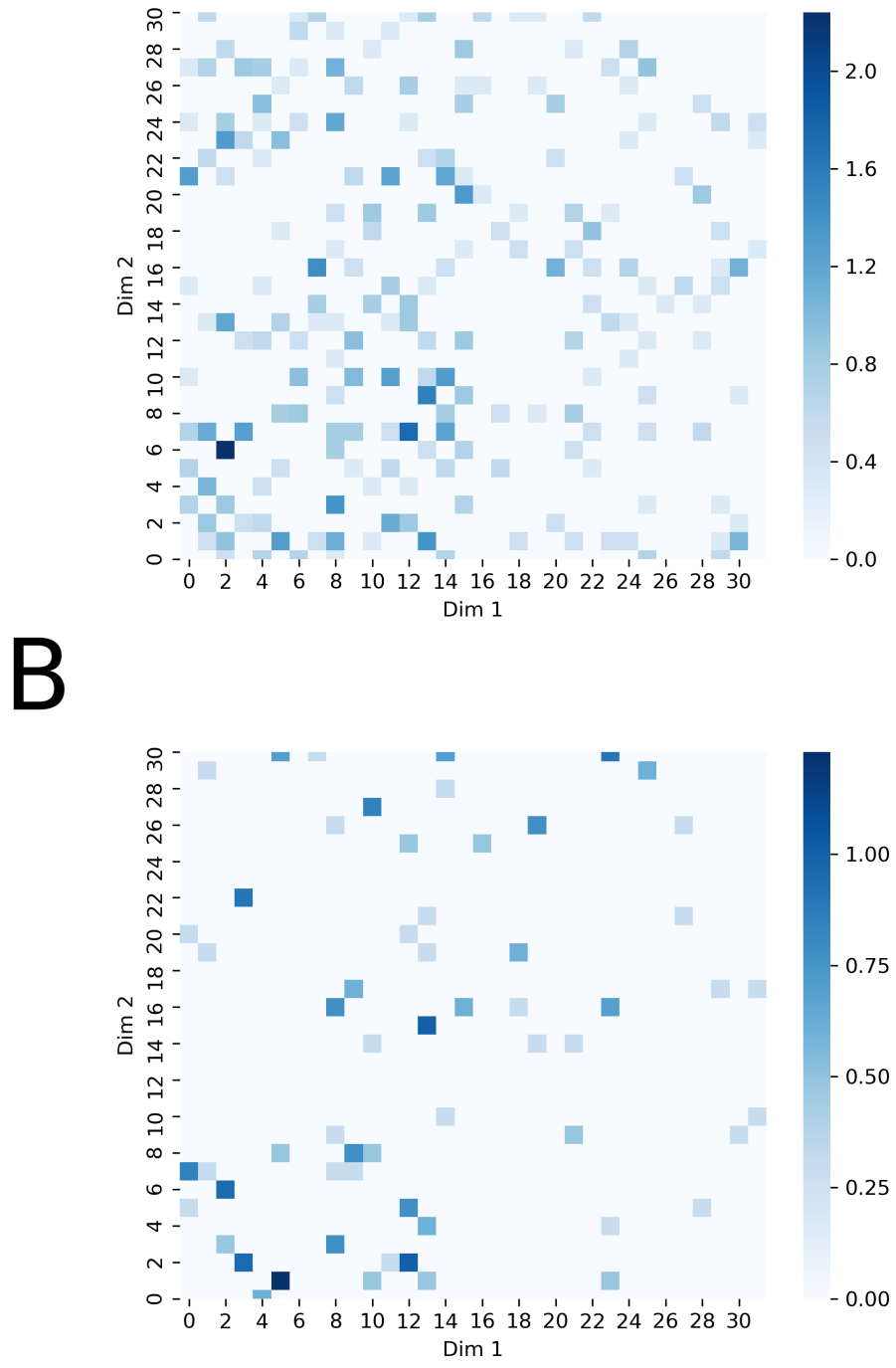

Figure 7: **Distribution of Compounds Embedded with the HCASE Method**. Compounds were embedded into the NatProd scaffold space with the help of HCASE method. DrugBank and CANVASS datasets into a space defined by NatProd reference scaffolds. The intensity of each cell of the heatmaps is proportional to the log_{10} of the number of compounds assigned to each cell, i.e. position in the embedded space. **A**) Embedding of DrugBank dataset. **B**) Embedding of CANVASS dataset.

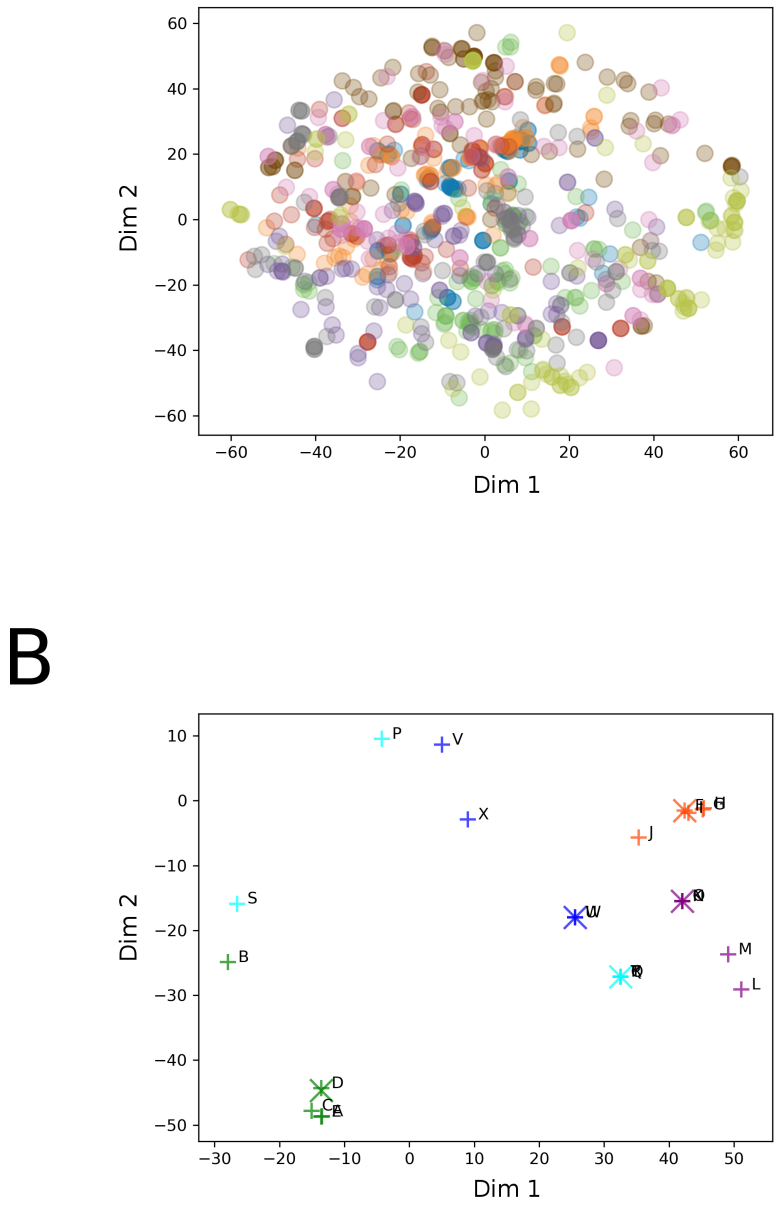

Figure 8: **Embedding of ChEMBL Reference Scaffolds with and Drug Molecules with Scaffold *t*-SNE.** The parameters of t-SNE embedding were set to default values, except for perplexity, i.e. learning rate = 200, iteration number $1,000$. **A**) t-SNE embedding of ChEMBL reference scaffolds at perplexity = 40. Highlighted are the scaffolds (see: *Tab.1*) cherry-picked from the ChEMBL reference scaffold set. **B**) Embedding of $k = 5$ Nearest Neighbors of selected DrugBank Molecules with Scaffold t-SNE. Enlarged (x) signs indicate the query compound of KNN analysis; green: DB00006, orange: DB00849, purple: DB00977, aqua: DB01362, blue: DB04837. (+) signs indicate the NNs of a query compound with identical color. Compounds are labeled according to *Fig. 1*