Universiteit
Leiden
The Netherlands

Inhibitor selectivity: profiling and prediction

Janssen, A.P.A.

Citation
Janssen, A. P. A. (2019, May 1). *Inhibitor selectivity: profiling and prediction*. Retrieved from https://hdl.handle.net/1887/71808

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/71808

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The following handle holds various files of this Leiden University dissertation:
http://hdl.handle.net/1887/71808

Author: Janssen, A.P.A.
Title: Inhibitor selectivity: profiling and prediction
Issue Date: 2019-05-01

Machines take me by surprise
with great frequency.
Alan Turing

7 t-Distributed Stochastic Neighbour Embeddings applied to drug discovery

Part of this research was published in A.P.A. Janssen *et al., J. Chem. Inf. Model.,* acs.jcim.8b00640 (2018).

Introduction

Chemical space is vast and can only be explored to a small extent by experimental methods to find suitable hits for drug discovery programs.[1,2] The search for new chemical starting points to modulate therapeutic targets is essential for the development of novel drugs. Due to the difficulties in finding suitable hits, it has been postulated that the best way to find a new drug is to start with an old drug.[3] This is in line with the central paradigm in medicinal chemistry that similar structures exert similar biological activities.[4] This paradigm can be mirrored to state that similar biological entities will be subject to modulation by similar molecular structures. It is this latter formulation that explains many

of the off-target activities observed when studying the inhibitory activity of a hit molecule against a subset of enzymes performing a similar function, which are part of the same cascade, or belong to a closely related family. An example from the endocannabinoid field is LEI104, which was found to be active against fatty acid amide hydrolase (FAAH) and diacylglycerol lipase (DAGL) α and DAGL-β. FAAH is responsible for the hydrolysis of *N*-arachidonoylethanolamine (anandamide), whereas the DAGLs produce 2-arachidonoyl-glycerol (2-AG).[5] The high similarity in structure of the substrate and product, respectively, of these enzymes implies that the binding pockets of the enzymes are similar, allowing LEI104 to bind to both.

Since most serine hydrolase inhibitors interact with the catalytically active amino acid, often in a covalent manner, off-target activity across this family is commonly observed for these compounds.[6] Such off-target activity is usually unwanted, and may sometimes lead to profound toxic effects as discussed in Chapter 5.[7] Despite the efficiency of activity-based protein profiling and other techniques to determine the selectivity of lead inhibitors, extensive screening campaigns are still expensive and time consuming. In the last decade, several large datasets with structure-activity relationships (SAR) of serine hydrolase inhibitors spanning a number of enzymes have been published.[6,8] These empirical datasets may serve as guides to explore chemical space around this drug target family and predict (off-)target activity using advanced computational chemistry methods, such as quantitative SAR models (QSAR), similarity ensemble approach (SEA), support vector machines, k-nearest neighbour, random forest, naïve bayes, (deep learning) neural networks (NN) and principal component analysis (PCA).[9–12]

Advanced machine learning models promise to revolutionize the field of drug discovery. Employing large high-dimensional datasets, these models are used to predict a wider range of biological activities for a compound, compared to traditional drug design methods (e.g. molecular modelling, docking and early QSAR-models, such as Hansch and Free-Wilson analysis[13]). Yet, many machine learning models are hampered in their applicability by lack of a clear interpretation and tendency to overfit on the high-dimensional data. Many of the best performing machine learning models are a black box in which it is unclear in what way the data are used to generate novel hypotheses. They also require in-depth knowledge of advanced cheminformatics and highly specialised or purpose-build software. These technical requirements slow down the implementation of the tools in the daily practice of drug discovery and consequently prevent the research community to take full advantage of the wealth of data becoming available. Therefore, there is a clear need for better tools to interpret and visualize complex, high-dimensional SAR datasets in an easy and intuitive manner and to predict the biological activity profile of novel hits for drug discovery programs.

t-Distributed Stochastic Neighbour Embedding

t-SNE was first published by van Maaten and Hinton in 2008 as the third variant of the Stochastic Neighbour Embedding concept.[14,15] The aim of the algorithm is to generate visualizations of high dimensional data by generating a two or three dimensional embedding thereof (see Box 7.1 for more detail). This makes t-SNE foremost a

dimensionality reduction algorithm like the more well-known principal component analysis (PCA)[16] and multidimensional scaling (MDS)[17]. The main difference, conceptually, between t-SNE, PCA and MDS is that the former is non-linear whereas the latter two are linear algorithms. This means that in general PCA and MDS predominantly preserve the large distances (differences) between points far apart in the high dimensional space, because these obtain a larger weight in the optimization algorithm than small differences. The result is that global structure is preserved well, but small variations in local structure are poorly mapped to the low dimensional embedding. t-SNE, like other algorithms before, aims to overcome these shortcomings by utilizing a non-linear cost-function, which weighs both the local and global structures. The success and general applicability of the t-SNE algorithm in dimensionality reduction is shown by the spur of publications utilizing it to visualize complex high-dimensional data sets in diverse experimental settings.[18–22] Here, the applicability of t-SNE to drug discovery is investigated. The applicability as a similarity metric is studied, with the ultimate goal to apply it in a general workflow to predict target-ligand interactions naïvely based on the similarity dogma.

Box 7.1 | A brief introduction to the t-SNE algorithm

t-Distributed Stochastic Neighbour Embedding is a non-linear dimensionality reduction algorithm, or 'manifold learning' algorithm. This branch of (unsupervised) machine learning is based on the premise that many real-world high-dimensional datasets are intrinsically low-dimensional. To be interpretable by humans, a low dimensional view of data is required. The process of dimensionality reduction is therefore aimed at removing redundant dimensions and deflating it to two or three dimensions, primarily to be visually interpretable by the analyst.

Mathematical framework

The original high-dimensional data-space is defined as $\mathbb{X}^D$ and the mapped space $\mathbb{Y}^2$, where *D* is the dimensionality. Original data points are denoted as x_i, mapped points y_i. The similarity of data point x_j to datapoint x_i was defined by Maaten and Hinton as the conditional probability, $p_{j|i}$, that x_j would be picked as its neighbour by x_i. If neighbours are picked in proportion to their probability density under a Gaussian centred at x_i the mathematical definition for $p_{j|i}$ is given by:

$$p_{j|i} = \frac{\exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)}{\sum_{k \neq l} \exp\left(-\frac{|x_k - x_l|^2}{2\sigma^2}\right)} \tag{7.1}$$

The *perplexity* variable, which is set by the user, is directly coupled to the σ in (7.1), and provides a measure of the smoothness of the embedding.

In the original SNE, a very similar definition was used for the probability condition for the embedding, $q_{j|i}$, where for convenience the σ_i is set to $\frac{1}{\sqrt{2}}$ such that:

$$q_{j|i} = \frac{\exp\left(-|y_i - y_j|^2\right)}{\sum_{k \neq l} \exp\left(-|y_k - y_l|^2\right)} \tag{7.2}$$

This definition led to various problems in the optimization procedure and subsequent implementations, so in t-SNE a Student t's distribution is used as the probability condition for the low-dimensional map. This leads to a definition of q_{ij} as follows:

$$q_{ij} = \frac{\left(1 + |y_i - y_j|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + |y_k - y_l|^2\right)^{-1}} \tag{7.3}$$

Where p_{ij} is the symmetrized version of $p_{j|i}$ according to $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}$, and likewise for q_{ij}. In a perfect embedding, p_{ij} and q_{ij} will be equivalent for all possibilities of ij, so the difference between these two quantities needs to be minimized in the embedding process. To achieve this, the Kullback-Leibler divergence is employed by t-SNE as a measure of the faithfulness of the embedding. The function to minimize, the cost function, is then given by:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{7.4}$$

in which P and Q represent the joint conditional probability distribution summed over all points where $j \neq i$. The values of p_{ii} and q_{ii} are set to zero. The cost function is optimized by a gradient descent which is given by:

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - p_{ij})(y_i - y_j) \left(1 + |y_i - y_j|^2\right)^{-1} \tag{7.5}$$

Physical interpretation

The optimization generating the low-dimensional embedding can be interpreted as a mass-spring system reaching its equilibrium. The magnitude of the Kullback-Leibler divergence for a given j will scale linearly with $\left(y_i - y_j\right)$ (7.5), analogous to Hooke's law, which scales linearly with the displacement of points. Intuitively the stiffness of the spring connecting i and j is determined by their distance in high-dimensional space, and also follows for t-SNE, where p_{ij} is strongly influenced by $\left(x_i - x_j\right)$.

t-SNE maps molecular similarity of drugs in drug-like chemical space

Based on the principle that the chemical structure of a compound determines its biological and chemical properties, a machine learning algorithm that predicts target-ligand interaction landscapes should be able to recognize molecular similarity between different molecules. Traditionally, chemical similarity is measured by the Tanimoto coefficient (Tc).[23] To calculate the distance between compounds the Tc uses a molecular fingerprint, which is a high-dimensional bit vector that captures the presence or absence of chemical groups in a molecule. As a similarity metric the Tc has its limitations, predominantly because it averages differences over all bits, thereby losing information.[24] Thus, we envisioned that the data contained in the molecular fingerprint should be used more efficiently by a machine learning algorithm to determine molecular similarity. Moreover, the Tanimoto coefficient is principally a distance metric between two fingerprints and does not allow

visualization of similarity within a set of molecules, making interpretation of this similarity metric nontrivial. t-SNE can be readily applied to bit vectors of any length and as such is easily applicable to chemical structures represented by molecular fingerprints. The algorithm could thus be used to find and cluster the most similar molecules in a large dataset and visualize that similarity clustering in a two- or three-dimensional space.

The applicability of the t-SNE algorithm to this problem was initially tested by visualizing the molecular similarity of molecules from the Drug Repurposing Hub, an online repository containing compounds that have been clinically tested or used in humans.[25] Only the launched drugs (2774) were selected and partially attributed to 27 chemotypes by hand. Morgan fingerprints (RD-Kit, 4096-bits, radius = 2, equivalent to ECFP4[26]) were generated for each of these 2774 clinical compounds using KNIME, an open source software package.[27,28] The fingerprints were fed into the Python implementation of the Barnes-Hut t-SNE algorithm to generate a map of the drug-like chemical space.[29] The resulting map (Figure 7.1) shows remarkable co-localization of most of the chemotypes. As an example, the family of penicillin-like structures at the far right of the plot (cyan) is completely separated from all other chemical matter. Some unannotated molecules (in grey) are visible in this cluster, but upon detailed inspection they all constitute β-lactams in which the sulfur is either substituted or omitted. In addition, many other highly dense unannotated clusters are visible at the boundaries of the map, corresponding to highly defined chemotypes, such as the rapamycin, conazole and oxytocin analogues. Of note, even in the apparently less defined centre of the map, clear co-localization of similar molecules can be observed, for example a cluster of aspirin-like molecules (orange, near origin).

Quantifying the quality of the generated embedding is not trivial. Despite the broad interest in machine learning algorithms and their application in numerous fields, few objective quality measures for embeddings are available.[30] Often co-ranking histograms are constructed, comparing the original high-dimensional space (H-space) to the constructed low-dimensional space (here: t-SNE, PCA or MDS space). Briefly, co-ranking histograms are constructed by counting the occurrences of co-ranking for each sample against all other samples. As this generates a two dimensional plot, they are often visualized as a heat-map. For the current purposes, such co-ranking histograms were constructed for t-SNE, PCA and MDS (Euclidian distances calculated in 2D embedding) and that of the Tanimoto distance (Figure 7.2A-D). Generally, the co-ranking histogram of t-SNE fanned out somewhat quicker than the Tanimoto plots, but for the lower ranked neighbours it corresponded well to the H-space. Compared to the two linear dimensionality reduction algorithms, t-SNE looks to be considerably better based on the heat-maps. There are several ways to summarize the information contained in a co-ranking histogram, of which the Q-score is conceptually one of the easiest to interpret.[30] It assesses the overall quality of the dimensionality reduction where 1 is a perfect score, and 0 is the lowest. The Q-score neglects information regarding the kind of ranking error (too high or too low) but summarizes the quality up until a certain value of the rank, k. The higher rate of co-ranking for low ranks that was visible in the histograms is immediately evident from the Q-score graphs (Figure 7.2E and F), which start

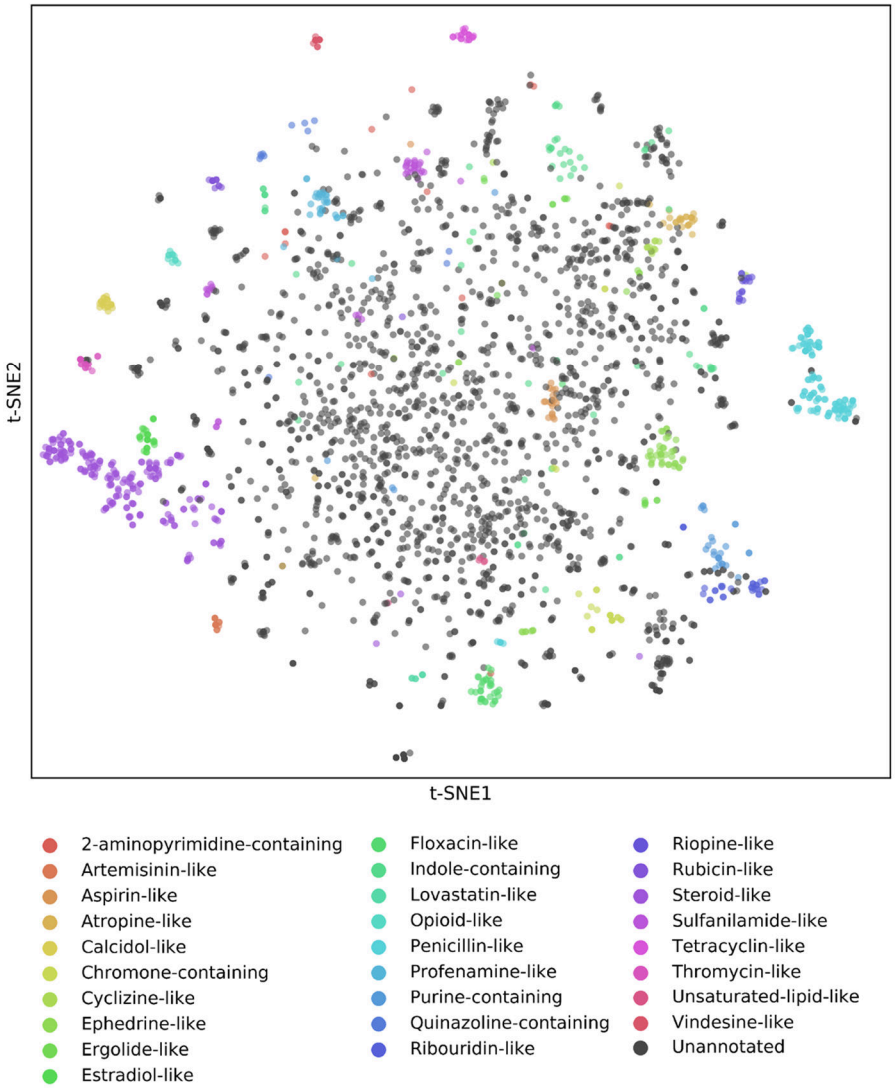

Figure 7.1 | t-SNE visualization of chemical space. t-SNE embedding of the 'launched' drugs in the Drug Repurposing Hub.[25] Embedding is based on the 4096-bit Morgan fingerprint. t-SNE settings: perplexity = 25, learning rate = 50, iterations = 10,000. Markers are coloured according to 27 manually attributed chemotypes.

out high but quickly fall upon increasing k. This behaviour is observed for all three dimensionality reductions and the Tanimoto distance. t-SNE however performs much better than the two linear algorithms, even for small values of k. When comparing t-SNE for the more direct neighbours in the histogram and plot (Figure 7.2A and E) it is clear that t-SNE follows the Tanimoto distance quite closely. Of note, this comparison is technically a bit skewed, as the Tc is only a different distance measure derived from the same H-space, not the result of a dimensionality reduction as is the case for t-SNE. A benefit of the three dimensionality reduction algorithms is that the 2D embedding allows for interrogation of the directionality of similarity, whereas when using the Tc, it is impossible to assess the similarity between two molecules when their Tc distance to a third is known. It can be concluded that t-SNE is able to map the local chemical space of approved drugs well, much better than linear dimensionality reduction approaches, whilst following a chemist's intuition. It recognizes molecular similarity in a broad set of diverse drug-like molecules and is able to present the results in a visually attractive and interpretable manner.

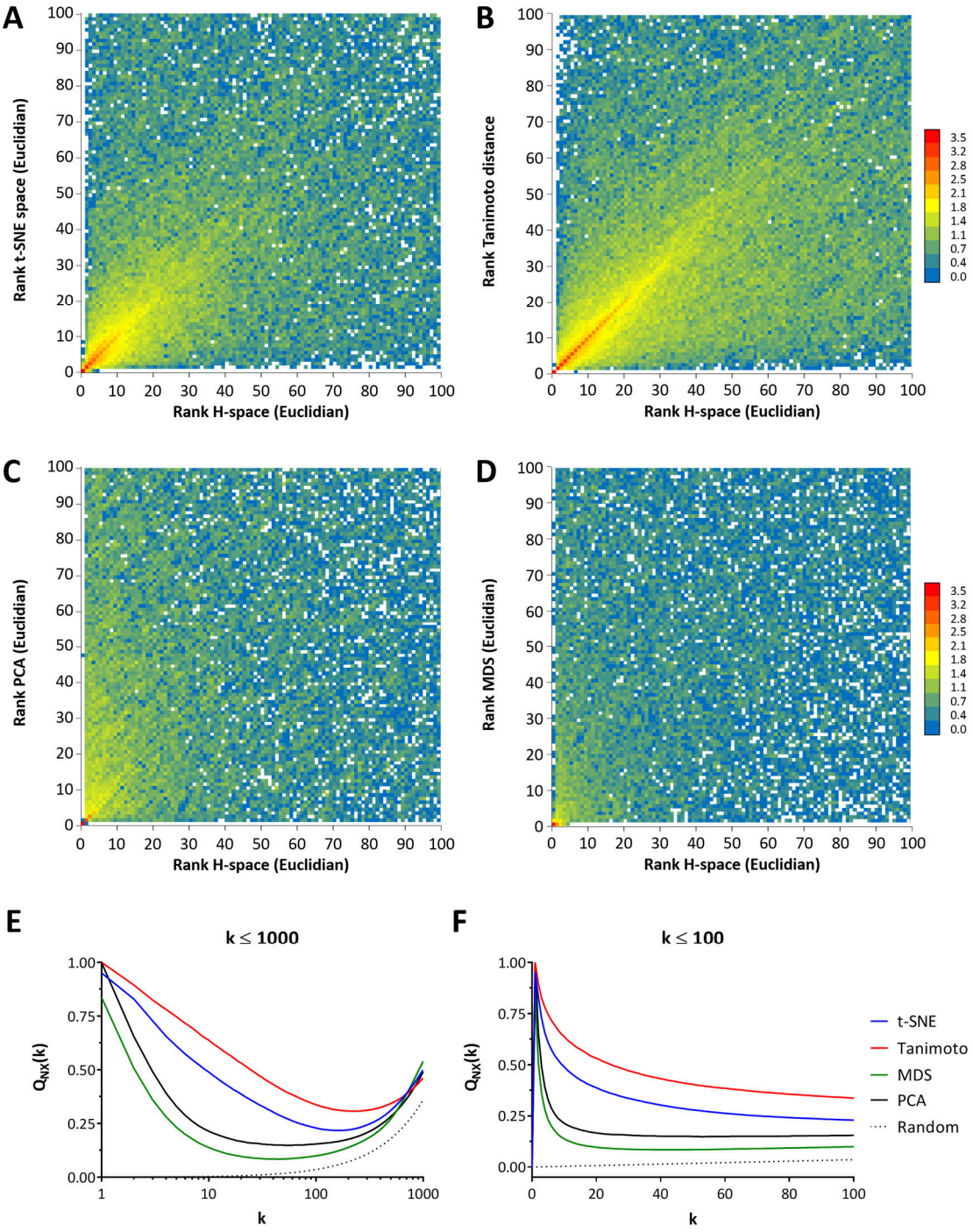

Figure 7.2 | Assessment of the quality of the dimensionality reduction as performed by t-SNE, PCA and MDS on the Drug Repurposing Hub molecules, based on the t-SNE embedding in Figure 7.1, the PCA plot in Supplementary Figure 7.1A and the MDS plot in Supplementary Figure 7.1B. Panels A-D show a $\log_{10}$-transformed co-ranking histogram for the top 100 ranked molecules for t-SNE (A), the Tanimoto coefficient (B), PCA (C) and MDS (D). The bottom panels show Q-score[30] graphs corresponding to the histograms above for the top 1000 (E) and top 100 (F) ranked molecules. H-space: the original high-dimensional space.

t-SNE map of serine hydrolases recapitulates phylogenetic information

Since closely related proteins typically bind similar endogenous molecules and (experimental) drugs, it was investigated whether the t-SNE algorithm is also capable of clustering proteins based on the similarity of their amino acids. Conceptually, this approach is analogous to proteochemometric modelling.[31] To quantify the similarity of serine hydrolases, first a manually curated list containing all serine endopeptidases and metabolic serine hydrolases (mSHs) was constructed. This was done in a similar manner as reported previously.[6] The amino acid sequences of the enzymes were then aligned using Clustal Omega.[32] This alignment was transformed to a fingerprint based on the physicochemical properties of the amino acids.[33] These fingerprints were used to create a two dimensional map of the target space by the t-SNE algorithm. The resulting map (Figure 7.3A) is capable of clustering most of the peptidases (triangles) and the mSHs (squares) in separate groups.

The embedding was compared with a more traditional phylogenetic tree (Figure 7.3B). As an illustration, several groups of proteins have been highlighted and annotated in the embedding. In purple all phospholipase A2 group IV proteins, except PLA2G4C, are clustered together. This separation is also observed in the phylogenetic tree. The small group of hepatocyte growth factor proteins (MST1, MST1L and HGF, in red) are related to the F2 protein, according to the phylogenetic tree, and contain kringle domains. In the t-SNE embedding however, F2 is quite far away, placed directly next to PRSS51 (not annotated), to which it is also quite similar. Intriguingly, the HGF related proteins are all expected to be catalytically inactive as they lack one or more of the typical catalytic triad residues, whereas F2 is confirmed to be a functional peptidase. This could indicate that t-SNE is able to separate similar proteins (global properties) on the absence or presence of key amino acids (local structure). It might also be one of the reasons why the red group is isolated in the plot. Other well defined protein clusters include the carboxylic esterases (CES, blue) and the products of genes *LIPA*, *LIPF* and *LIPM* (yellow). The centre of the embedding where less clusters are observed, still holds large co-localized families of proteins, e.g. the kallikreins (KLK, light green) or the acyl-coenzyme A thioesterases (ACOT, not annotated). It appears that t-SNE, when applied to binarized sequence alignments, is able to analyze and cluster these entities, maintaining the separation of large groups (peptidases and mSHs), whilst also keeping subclasses of proteins closely associated.

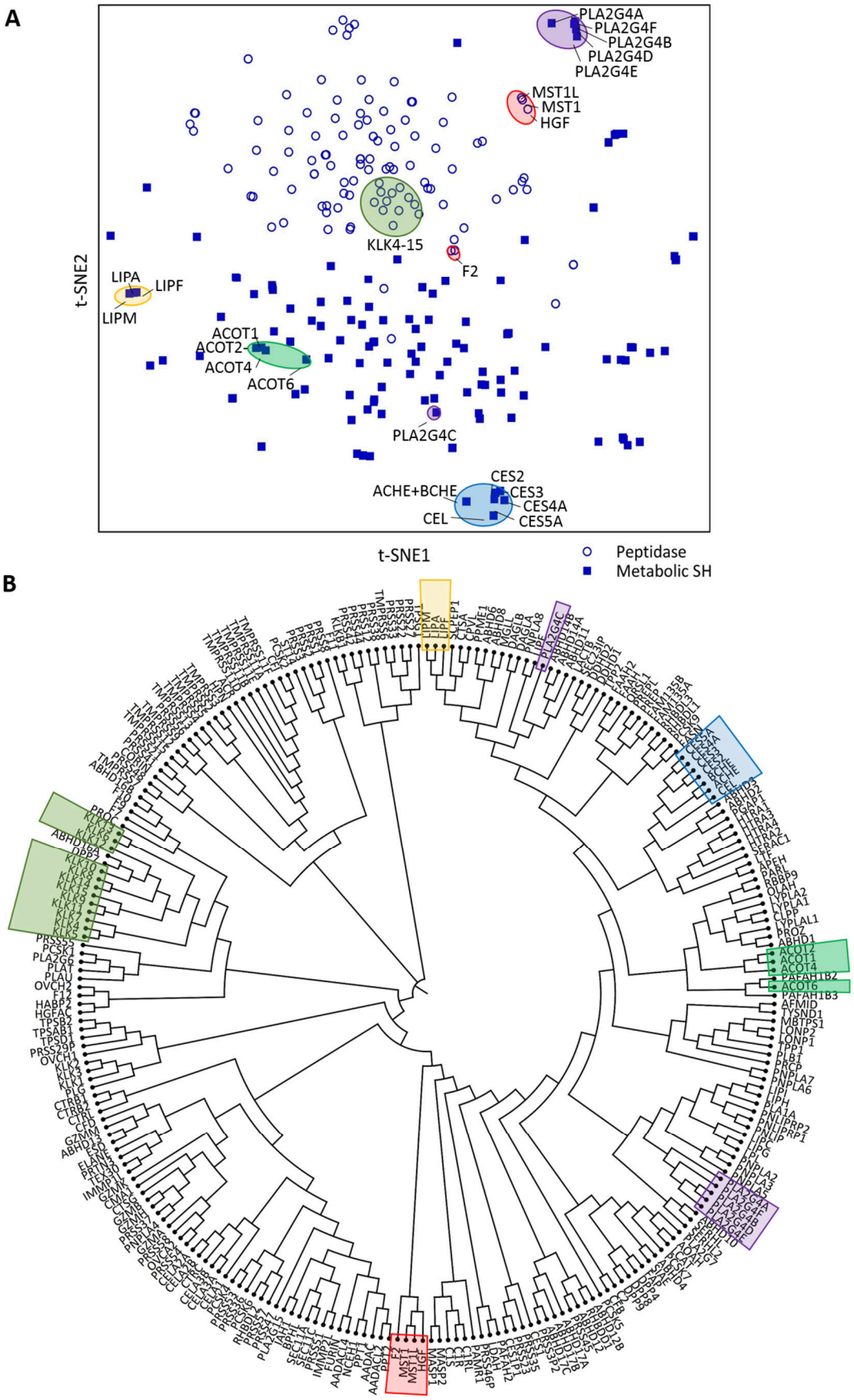

Figure 7.3 | Comparison of t-SNE embedding (A) and phylogeny tree (B) of the serine hydrolase (SH) family. t-SNE settings: perplexity = 30, learning rate = 50, iterations = 10,000.

t-SNE analysis of serine hydrolase inhibition

t-SNE thus presents itself as a similarity visualization technique able to find and quantify similarities between molecules and proteins. With this technique a system could be envisioned where a newly designed molecule is compared to a large set of available molecules of which the bio-activity data is known. If one now finds the most similar known compounds utilizing t-SNE, the bio-activity data of those molecules could be used as a prediction for the bio-activity of the newly designed inhibitor. The t-SNE embedding of the protein space could then be used to append enzymes, for which no bio-activity data is available, as potential targets to the prediction based on the fact that they are closely related to known targets. Through this intuitive process, targets, or off-targets, of novel molecules may be predicted in an efficient way. This approach is comparable to PASS and various other techniques, reviewed by Bender *et al*.[34,35]

Currently, the largest dataset reported for serine hydrolases is that of Bachovchin *et al*.[6], which screens a library of 140 inhibitors against 72 SHs (10,080 data points). This dataset contains bio-activity at one concentration (50 μM) and all inhibitors have a carbamate warhead. The predictive utility of this dataset is, therefore, restricted to similar carbamate-based inhibitors. Another comprehensive serine hydrolase inhibitor dataset was generated using the EnPlex assay.[8] This dataset has bio-activity data for 55 quite diverse inhibitors (Figure 7.4) and 94 serine hydrolases (Figure 7.5). This dataset was processed using t-SNE to visualize the inhibition profiles projected on the embedding of molecules and proteins and thereby assess the co-localization.

When the 55 inhibitors were transformed in their respective fingerprints and clustered by t-SNE a quite disperse embedding was obtained (Figure 7.4A), in line with the diversity in the molecular set. Some more defined groups were observed, which upon inspection indeed corresponded to more similar molecules. When the bio-activity data was projected on this visualization by colouring the markers according to the pIC_{50}s for specific targets (Figure 7.4B-D) very distinct patterns could be observed. All DPP4 inhibitors were positioned in the far right of the embedding, with the exception of methoxyarachidonoyl fluorophosphonate (MAFP) which is in the centre of the plot and, by its reactive nature, inhibits virtually all serine hydrolases tested in this assay. For the endocannabinoid system related enzyme FAAH a similar observation can be made, with all inhibitors co-localized to the left of the embedding, barring some reactive molecules in the centre. Most of the tested FAAH-inhibitors are carbamate based, and this chemotype typically shows off-target activity against carboxylic esterases (CES). This is exemplified by Figure 7.4D where the potency for CES1 is annotated. Only the more chemically distinct urea based PF-3845 (topmost marker) is inactive for CES1.

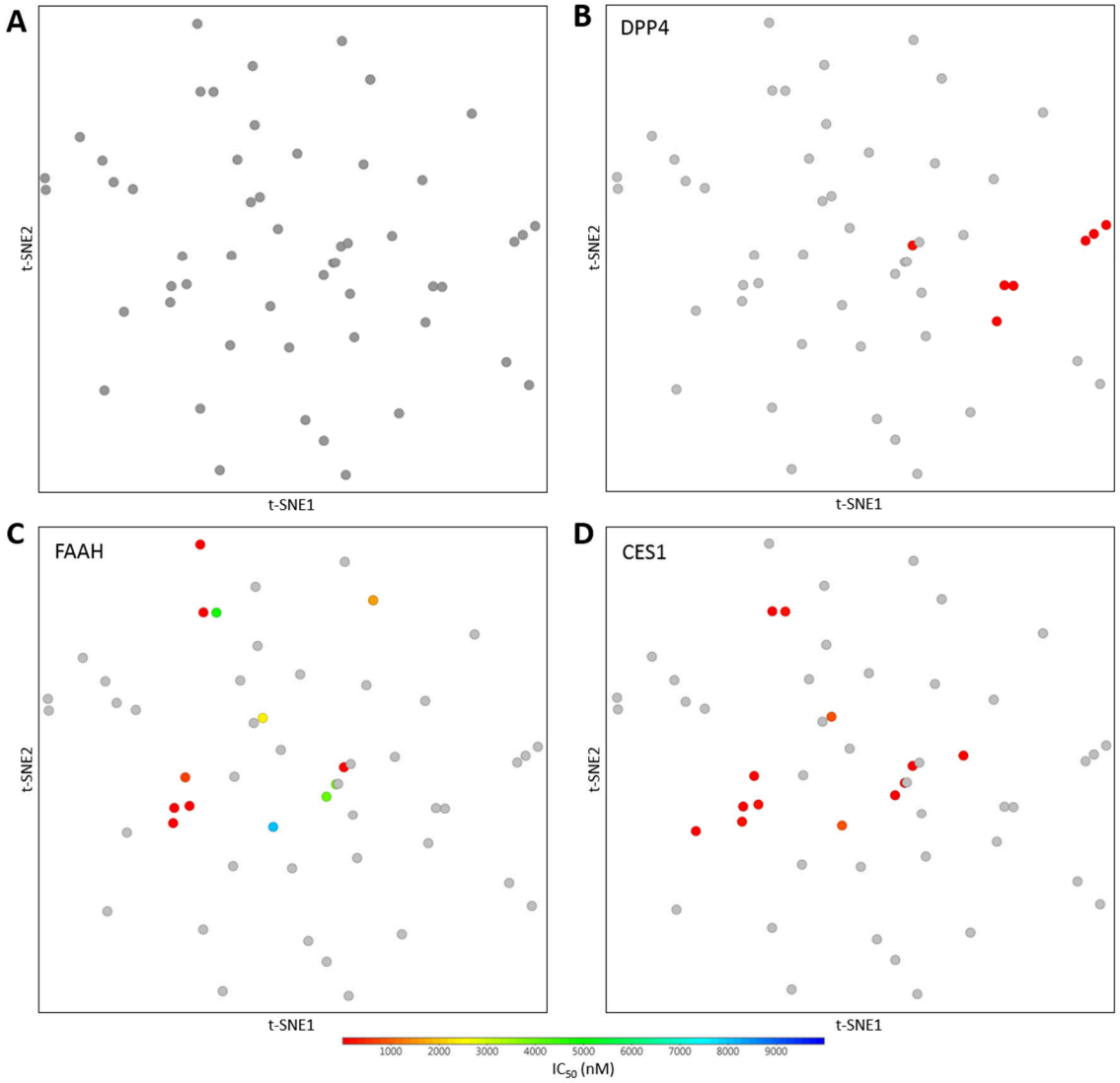

Figure 7.4 | t-SNE embeddings of the serine hydrolase inhibitors screened in the EnPlex assay.[8] A) Unannotated t-SNE clustering. B-D) All reported inhibition interactions of title enzymes are denoted in a gradient colouring. Only measured interactions < 10.000 nM are coloured in the embedding.

Next, the protein-target space was analysed. To aid in this analysis, the serine hydrolases for which no bio-activity data is available were filtered out (Figure 7.5A), and the markers were coloured according to the activity of three dissimilar inhibitors (Figure 7.5B-E). These inhibitors vary not only in structure, but also in reactive group (warhead), targeted subfamily, and selectivity. The localization of the inhibited SHs in the plots shows clear co-localization in pairs or groups. For example, JP104 (Figure 7.5B) has six measured IC_{50}s, which are arranged in the embedding as three duos. The active metabolite of lactacystin is, as might be expected from its small molecular structure containing a reactive β-lactone, less selective. Several clear groups of targets can be seen, most notably the seemingly unrelated cluster of retinoid inducible serine carboxypeptidase (SCPEP1), probable serine carboxypeptidase CPVL (CPVL), lysosomal protective protein (CTSA) and α/β-hydrolase domain containing protein 11 (ABHD11). Of note, the only enzyme classified as peptidase targeted by lactacystin is located close to several targeted mSHs. Finally, chymostatin is

quite selective, with only 2 observed interactions in this enzyme panel. The two peptidases targeted, chymase and chymotrypsin-C, are indeed closely related according to the t-SNE embedding, which is not directly evident from the phylogeny tree.

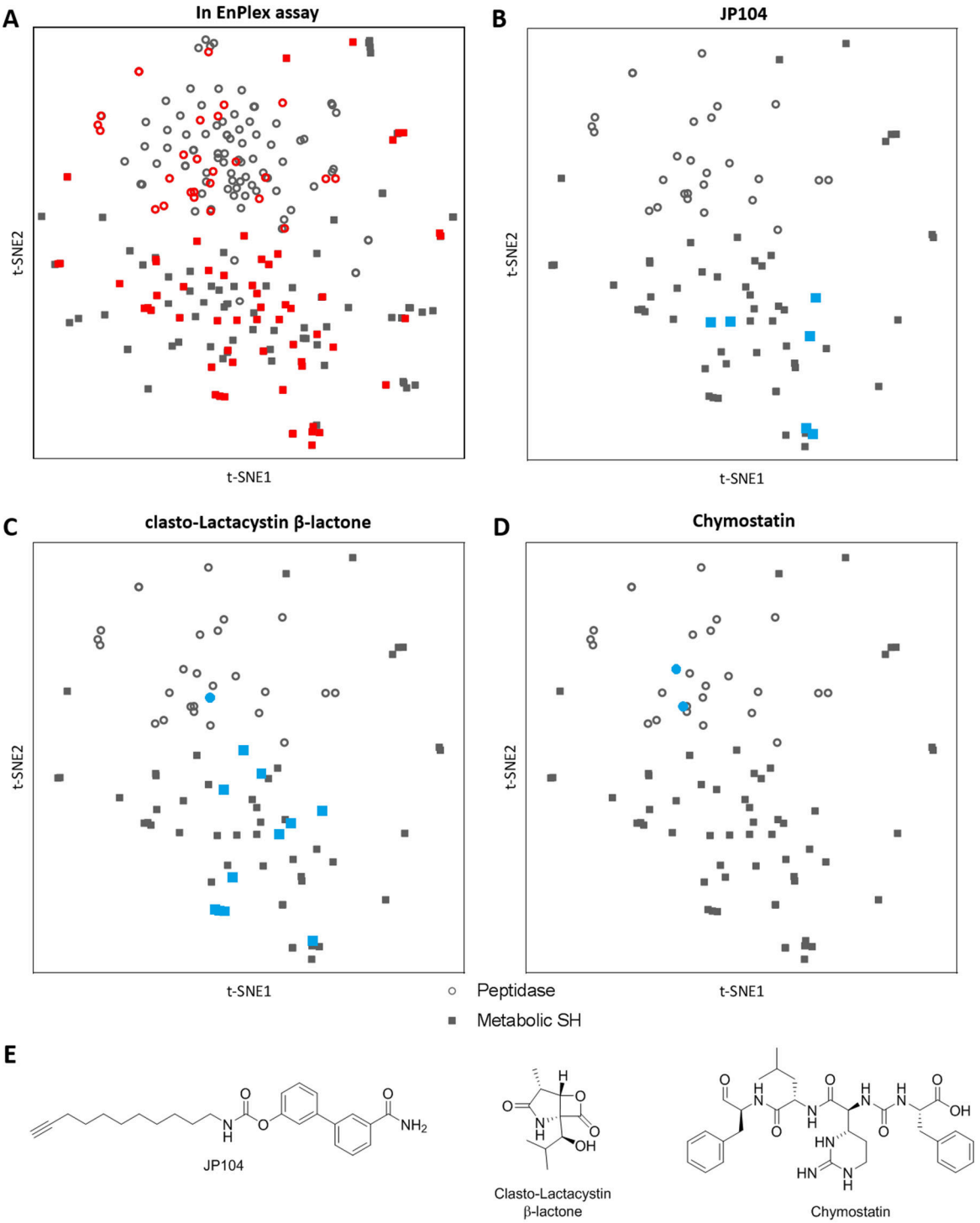

Figure 7.5 | t-SNE embeddings of the serine hydrolase superfamily. A) Red colouring denotes the inclusion in the EnPlex assay. B-D) All reported inhibition interactions of title inhibitor are denoted in blue. Only measured serine hydrolases are included in the embedding (those coloured red in A). E) Structures of inhibitors in B-D.

Conclusion

Taken together, it was shown that the dimensionality reduction algorithm t-Distributed Stochastic Neighbour Embedding can successfully be applied to molecular fingerprints, and that this generates a new, visually interpretable and quantitative similarity metric. Annotating a similarity visualization of a small inhibitor set with serine hydrolase inhibition data showed very distinct co-localization of activities in the embedding. t-SNE was also able to generate similarity maps for the serine hydrolase enzyme family when applied to a binarised multiple sequence alignment. This map revealed similarities between proteins not necessarily directly related in phylogeny. However, there is significant overlap between the traditionally assigned subfamilies and the co-localization in the t-SNE embeddings. By superimposing the available bio-activity data as a third dimension in colour on top of the t-SNE embeddings, definitive hotspots were shown where co-localized proteins were inhibited by the same inhibitor, which not always corresponded to the phylogenetic information. Given more (biological) data, it is expected that these mappings combined, in principle, could be applied to predict (off-)targets that are in these hotspots, or even predict (off-)targets for new molecules based on their molecular similarity to known compounds. It is expected that the target space visualization could be enhanced by improving the alignment of the enzymes. One option would be to centre the alignment around the catalytic serines, to improve the definition of the active site similarities, or truncate the sequences such that the alignment is more biased towards the actual binding site of the enzyme.

The concept presented here can easily be adapted to work with any dataset available. Because all data, algorithms, and data processing tools used are in the public domain or open source, it is highly adaptable and extensible. Concrete examples include different druggable protein classes, such as kinases (Chapter 8), G-protein coupled receptors, ion channels or nuclear hormones. It could also be trained on a different molecular set altogether, e.g. solubility, membrane permeability, metabolic stability, pharmacokinetics or toxicological data.

Methods

Cheminformatics tools

All molecular descriptors, molecular representations (SMILES, InChIKey) and fingerprints were generated using the RDKit software, either using the KNIME extensions or as the Python implementation.[28] Morgan fingerprints (4096-bits, radius 2) were used for all molecular fingerprints.

t-SNE algorithm

All t-SNE embeddings were generated with the Python Scikit-learn (v. 0.19) implementation of the Barnes-Hut t-SNE algorithm, either implemented in a 'Python for KNIME' node or as part of a Python script.[29]

Principal component analysis

Principal component analysis was performed using the Python Scikit-learn decomposition module (v. 0.19) based on the same 4096-bit fingerprints used for the other analyses.

Multidimensional scaling

The multidimensional scaling was performed using the Python Scikit-learn manifold module (v. 0.19), based on the 4096-bit fingerprints used for other analyses.

Co-ranking histogram and Q-scoring

The co-ranking histograms were constructed in Python, using SciPy's Rankdata functionality, using ordinal ranking.[36] Generated histograms were exported to Microsoft Excel 2016 where they were $\log_{10}$ transformed and heatmaps were generated.
The Q-score was calculated as described previously.[37] In brief, Q-scores were calculated by summarizing mild k intrusions and mild k extrusions for each k and normalizing to the sample size according to the following formula:

$$Q_{NX}(k) = \frac{1}{kN} \sum_{k,l} q_{kl} \tag{7.6}$$

where *N* is the total number of identities (2774, the number of molecules) and

$$q_{kl} = |\{(i,j) : \rho_{ij} = k,\ r_{ij} = l\}| \tag{7.7}$$

where ρ_{ij} denotes the rank in the low-dimensional manifold or the rank according to Tanimoto distance, and r_{ij} the rank in high dimensional space.

Serine hydrolase sequence information and bitstring

Sequence information of the serine hydrolases was retrieved from Uniprot.[38] The proteins were aligned using the online Clustal Omega tool provided by the EMBL-EBI using the default settings.[32] The standard "Clustal w/o numbers" output generated was transformed to a bitstring using the amino acid fingerprints as provided in Heil *et al.*[33] with the following additions: alignment dashes (-), stops (*) and blanks (X) were all considered empty, and were represented by 23 0's.

Bioactivity used

Activity data for the serine hydrolase superfamily was used without adaptations from Bachovchin *et al.*[8].

Supplementary Figures

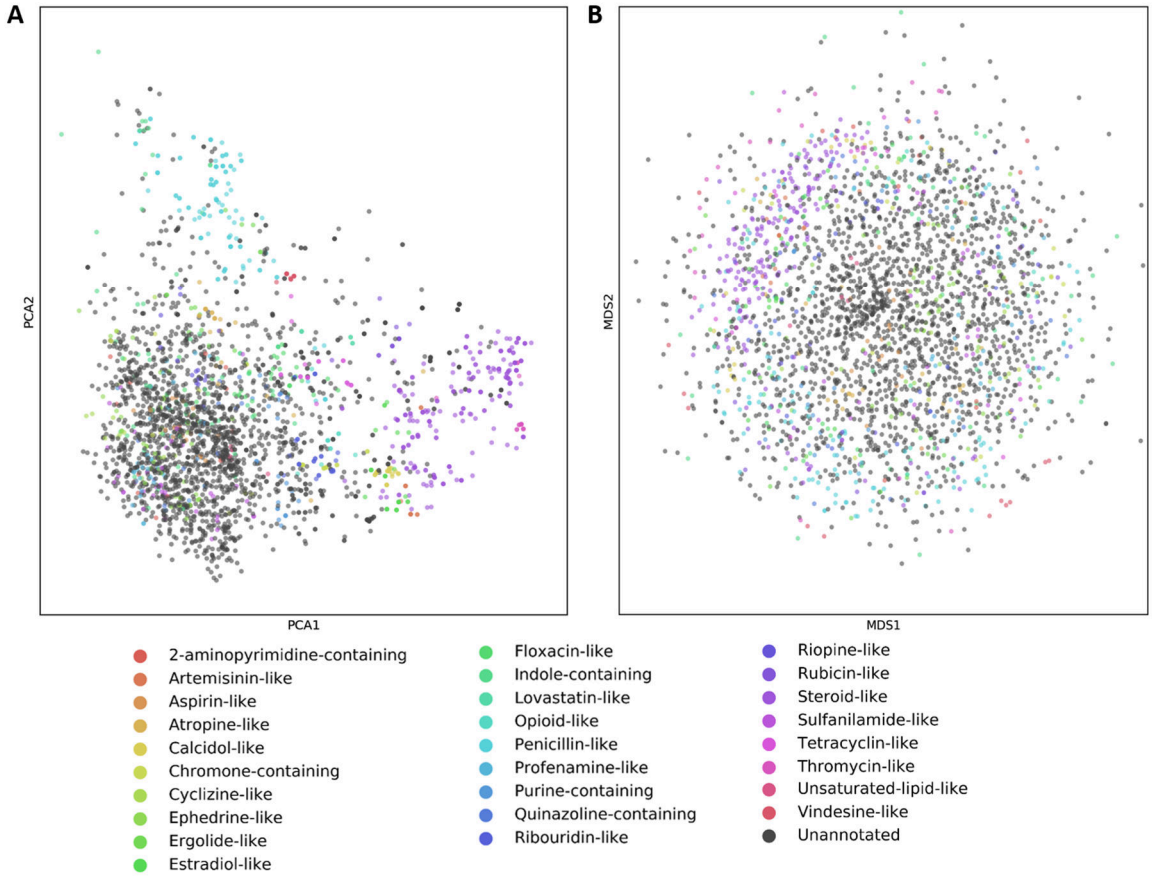

Figure S7.1 | Principal component analysis (A) and multidimensional scaling analysis (B) of the drug repurposing hub molecules with the status "Launched". Embeddings are based on the 4096-bit Morgan fingerprint (RD-Kit). Markers are coloured according to 27 manually attributed chemotypes.

References

1. Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **48**, 722–730 (2015).
2. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–823 (2004).
3. Imming, P. Medicinal Chemistry: Definitions and Objectives, Drug Activity Phases, Drug Classification Systems. in *The Practice of Medicinal Chemistry* (ed. Wermuth, C. G.) 3–13 (2011).
4. Bender, A. & Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204 (2004).
5. Baggelaar, M. P. *et al.* Development of an Activity-Based Probe and In Silico Design Reveal Highly Selective Inhibitors for Diacylglycerol Lipase-α in Brain. *Angew. Chemie Int. Ed.* **52**, 12081–12085 (2013).
6. Bachovchin, D. A. *et al.* Superfamily-wide portrait of serine hydrolase inhibition achieved by library-versus-library screening. *Proc. Natl. Acad. Sci.* **107**, 20941–20946 (2010).
7. van Esbroeck, A. C. M. *et al.* Activity-based protein profiling reveals off-target proteins of the FAAH inhibitor BIA 10-2474. *Science* **356**, 1084–1087 (2017).
8. Bachovchin, D. A. *et al.* A high-throughput, multiplexed assay for superfamily-wide profiling of enzyme activity. *Nat. Chem. Biol.* **10**, 656–663 (2014).
9. Christmann-Franck, S. *et al.* Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **56**, 1654–1675 (2016).
10. Sorgenfrei, F. A., Fulle, S. & Merget, B. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem* **13**, 495–499 (2018).
11. Merget, B., Turk, S., Eid, S., Rippmann, F. & Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J. Med. Chem.* **60**, 474–485 (2017).
12. Cichonska, A. *et al.* Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. *PLOS Comput. Biol.* **13**, e1005678 (2017).
13. Hansch, C. & Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **86**, 1616–1626 (1964).
14. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
15. Hinton, G. & Roweis, S. Stochastic Neighbor Embedding. in *Proceedings of the 15th International Conference on Neural Information Processing Systems* 857–864 (MIT Press, 2002).
16. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
17. Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401–419 (1952).
18. Mahfouz, A. *et al.* Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods* **73**, 79–89 (2015).
19. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–52 (2013).
20. Abdelmoula, W. M. *et al.* Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proc. Natl. Acad. Sci.* **113**, 12244–12249 (2016).
21. Bushati, N., Smith, J., Briscoe, J. & Watkins, C. An intuitive graphical visualization technique for the interrogation of transcriptome data. *Nucleic Acids Res.* **39**, 7380–7389 (2011).
22. Taskesen, E. & Reinders, M. J. T. 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *PLoS One* **11**, e0149853 (2016).
23. Willett, P. & Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity. *Quant. Struct. Relationships* **5**, 18–25 (1986).
24. Baldi, P. & Nasr, R. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values. *J. Chem. Inf. Model.* **50**, 1205–1222 (2010).
25. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
26. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
27. Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. in *Data Analysis, Machine Learning and Applications* (ed. Preisach C., Burkhardt H., Schmidt-Thieme L., D. R.) 319–326 (Springer, Berlin, Heidelberg, 2008).
28. Landrum, G. RDKit: Open-source cheminformatics; http://www.rdkit.org.
29. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
30. Lee, J. A. & Verleysen, M. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognit. Lett.* **31**, 2248–2257 (2010).
31. van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W. T. & Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.* **2**, 16–30 (2011).

32. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
33. Heil, B., Ludwig, J., Lichtenberg-Frate, H. & Lengauer, T. Computational recognition of potassium channel sequences. *Bioinformatics* **22**, 1562–1568 (2006).
34. Lagunin, A., Stepanchikova, A., Filimonov, D. & Poroikov, V. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* **16**, 747–8 (2000).
35. Bender, A. *et al.* Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screen.* **10**, 719–31 (2007).
36. Jones, E., Oliphant, T., Peterson, P. & others. SciPy: Open source scientific tools for Python.
37. Lee, J. A. & Verleysen, M. *Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods*. **4**,
38. Consortium, T. U. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).