Chemography of Natural Product Space

Authors Tomoyuki Miyao[1,2]*, Daniel Reker[2]*, Petra Schneider[2], Kimito Funatsu[1], Gisbert Schneider[2]

Affiliations
[1] Department of Chemical Systems Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan
[2] Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, ETH Zurich, Zurich, Switzerland

Key words
- computational chemistry
- drug discovery
- generative topographic map
- machine learning
- polypharmacology

received July 8, 2014
revised Nov. 28, 2014
accepted January 12, 2015

Bibliography
DOI http://dx.doi.org/10.1055/s-0034-1396322
Published online February 26, 2015
Planta Med 2015; 81: 429–435

ISSN 0032-0943

Correspondence
Prof. Dr. Gisbert Schneider
Institute of Pharmaceutical Sciences
Swiss Federal Institute of Technology (ETH) Zürich
Department of Chemistry and Applied Biosciences
Vladimir-Prelog-Weg 4
8093 Zürich
Switzerland
Phone: + 41 44 633 73 27
Fax: + 41 44 633 13 79
gisbert.schneider@pharma.ethz.ch

Abstract

We present the application of the generative topographic map algorithm to visualize the chemical space populated by natural products and synthetic drugs. Generative topographic maps may be used for nonlinear dimensionality reduction and probabilistic modeling. For compound mapping, we represented the molecules by two-dimensional pharmacophore features (chemically advanced template search descriptor). The results obtained suggest a close resemblance of synthetic drugs with natural products in terms of their pharmacophore features, despite pronounced differences in chemical structure. Generative topographic map-based cluster analysis revealed both known and new potential activities of natural products and drug-like compounds. We conclude that the generative topographic map method is suitable for inferring functional similarities between these two classes of compounds and predicting macromolecular targets of natural products.

Abbreviations

CATS:	chemically advanced template search
COBRA:	Collection of Bioactive Reference Analogs
DI:	deoxydihydroisoflindissol
DNP:	Dictionary of Natural Products
EM:	expectation-maximization
GMM:	Gaussian mixture model
GPCR:	G-protein coupled receptor
GTM:	generative topographic map
5-HT:	5-hydroxytryptamine
MACCS:	molecular access system
MOE:	Molecular Operating Environment
NCE:	new chemical entity
PCA:	principal component analysis
RBF:	radial basis function
RMSE:	root mean square error
SOM:	self-organizing map
WOMBAT:	World of Molecular Bioactivity

Introduction

Natural products have a long-standing history as a source for innovative compounds in drug discovery [1–3]. A rationale for their success is the historic evolutionary exploration of chemical modifications leading to compounds containing privileged structural motifs with biophoric properties [4, 5]. It has been estimated that for more than half of the published NCEs for therapeutic use, natural products served as an inspiration [6], with a particular emphasis on anticancer agents [7]. First studies have reported computer-assisted natural product deorphaning and also presented algorithmic advances for automated structure elucidation [8–11], suggesting that pharmacophore models derived from bioactive natural products may guide the development of drug-like, chemically tractable natural product mimetics. Here, we show how a global perspective on the pharmacophores found in natural products may be exploited for drug discovery by visualizing landscapes of pharmacophoric traits. These give insightful hints for relating natural products to synthetic compounds and assessing their potential polypharmacology. The approach also provides a means for identifying sparsely populated biophoric regions of chemical space.

Chemography is an umbrella term describing computational methods for the visual inspection of typically two-dimensional representations of

* These authors contributed equally to this work.

chemical space [12]. Dimensionality reduction is required for generating these chemical space maps, because most of the computed molecular representations are high-dimensional "descriptors" [13], typically real-numbered vectors, or binary fingerprints. Various dimension reduction algorithms may be used for this purpose [14]. Chemographic techniques have been mainly applied to assess the structural diversity of compound libraries and the selection of molecular subsets in drug design [15,16]. Both aspects play an important role in the comparison of natural products to synthetic compound libraries to assess interesting regions in terms of naturally optimized scaffolds and investigate the quality of compound libraries for chemotype and scaffold diversity [17]. Accordingly, maps of natural products and drug-like compounds have been constructed using properties or structure-based descriptions with various dimension reduction techniques [18]. These studies report a pronounced discrepancy of structural features between man-made compounds and natural products. For example, Grabowski et al. trained an SOM on physicochemical properties of natural products and drug-like compounds [19]. They observed clearly separated clusters of drugs and natural products with little overlap. Lee et al. used the same dimension reduction technique but on a pharmacophore pattern representation [20]. These maps were constructed to grasp pharmacophoric traits of small molecules and capture their interaction potential with large biomolecules. Intriguingly, in contrast to the maps solely based on physicochemical properties, pharmacophore-based chemical space projections indicate a strong mixing of natural products and synthetic compounds. This observation suggests that, in spite of their structural differences, the members of these two chemical universes are related through their pharmacophoric features and, therefore, their interaction potential with macromolecular targets. Consequently, we hypothesize that chemography could help identify synthetically neglected or unexplored pharmacophoric patterns of natural products, and at the same time relate natural products to target-specific regions in pharmacophore space. In a related study, Oprea and coworkers compared properties of natural products with synthetic drugs from WOMBAT [21] by linear PCA [22] and concluded that there are, in fact, natural product clusters that lack representation in the set of synthetic drugs. Here, we extend this view on chemical space by nonlinear pharmacophore mapping.

Results and Discussion

Motivated by the fact that pharmacophore representations allow for a meaningful mixing of natural products and synthetic drugs when mapped with the help of SOMs, here we explored GTMs [23] for their ability to generate meaningful visualizations of pharmacophoric natural product space [24–26]. GTMs follow a similar theoretical concept as SOMs and are therefore often considered a probabilistic extension of the SOM algorithm [27]. Both methods distinguish themselves from other projection methods by being nonlinear: They create a manifold in the original data space (here: chemical space) that is used for the mapping [28]. However, GTMs extend this concept by adding a GMM onto the manifold, which is assumed to reflect the underlying distribution of the data. Thereby, the creation of the map is an instance of the problem of positioning the Gaussian functions appropriately. There are methods like the EM algorithm that allow for a fast convergence to local optima, in contrast to heuristic SOM training which stops after a user-defined number of iterations [29]. The

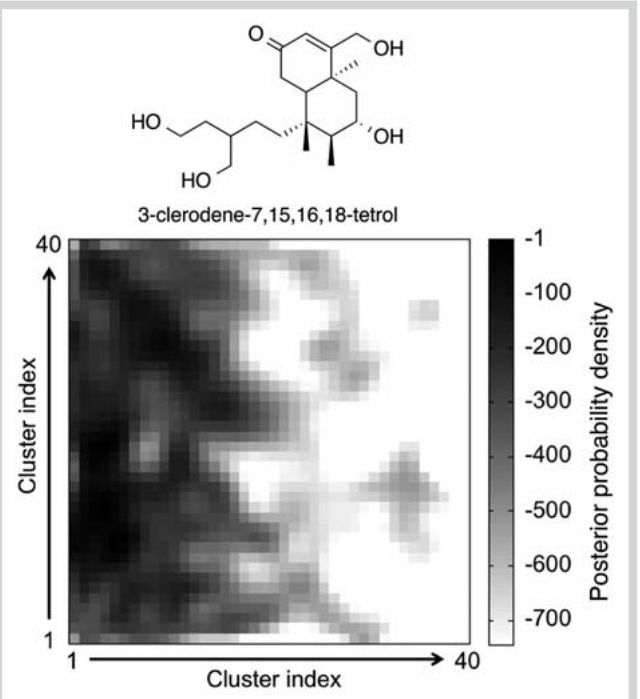

Fig. 1 Logarithmized posterior probability density on the final generative topographic map for 3-clerodene-7,15,16,18-tetrol.

GTM's probabilistic character allows for projecting a data instance (here: a molecule) not only to one point on the map but instead calculates the activation for every single Gaussian, so that we obtain a fuzzy location of every compound on the map (**Fig. 1**). The quality of the map can be assessed through calculating the RMSE (Eq. 1) of back-projected data points.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| y(I_i, W) - x_i \|^2} \quad (1)$$

where N is the number of compounds, $\mathbf{l}_i$ is the position of compound i on the map (also referred to as the *latent space*), $\mathbf{x}_i$ is the descriptor vector of compound i, and y is the back-projection function from the latent space to the original data space. In simple terms, Eq. 1 corresponds to the loss of information by projecting from the high-dimensional original data space spanned by the chemical descriptors to the lower-dimensional latent space. The number of RBFs, the variance (width) of the basis functions, the regularization parameters, and the number of latent points (map size) govern GTM training. The parameters related to the RBFs control the linearity (smoothness) of the map, and the regularization parameter helps avoid overfitting. The number of Gaussian distribution functions in the high-dimensional original data space gives the number of latent points in the projection. Each latent point is connected to the mean value of a Gaussian distribution in the original data space, and the sum of the Gaussian distributions captures the underlying data distribution.

We generated a GTM for a total of 157 929 natural products and a small but carefully curated collection of 12 644 drug-like synthetic compounds for which macromolecular targets are annotated (COBRA [30]). The molecules were represented by 210-dimensional CATS (version 2) pharmacophore descriptors [31,32]. After GTM training, the RMSE (Eq. 1) for all compounds in both data collections had a value of 0.99. For the RMSE calculation,

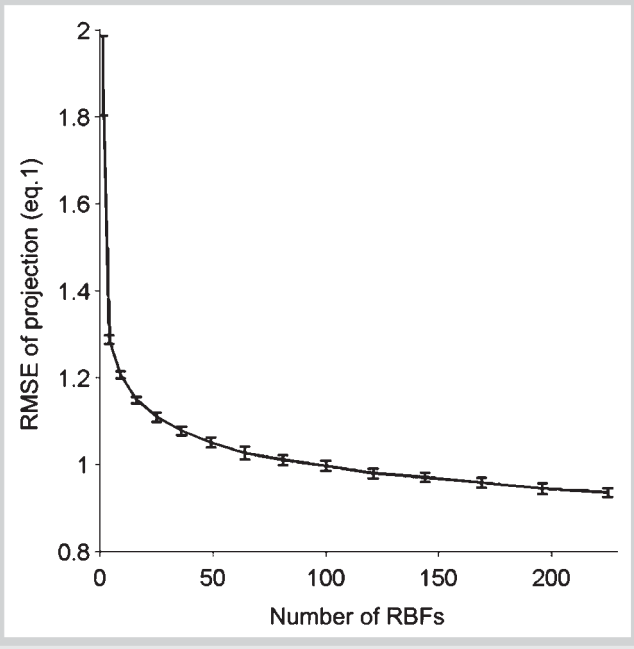

Fig. 2 Mapping error of the GTM projection (RMSE, Eq. 1) with respect to the number of RBFs used to define the map's resolution. The error bars present the standard deviations computed from 30 experiments with randomly initialized mapping parameters (**W**, β). The curve gives the RMSE of models initialized through PCA analysis of the original data.

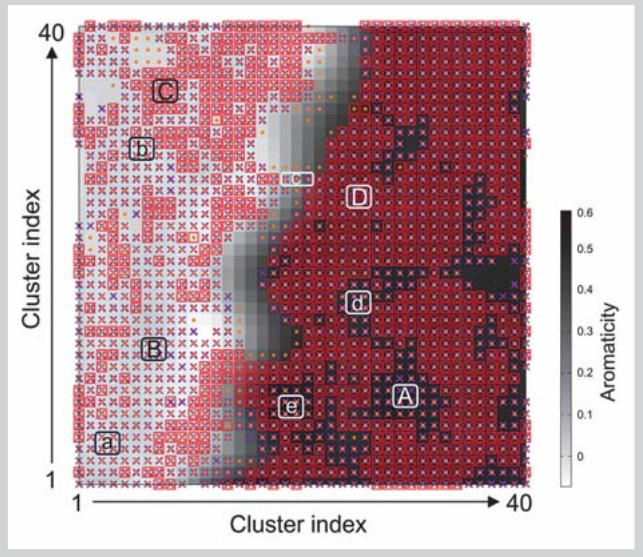

Fig. 3 GTM projection of pharmacophore space. The map shows 99.5% of the training data lacking extreme outliers. A magenta cross marks clusters that contain natural products (157 265 compounds from DNP), red open squares mark clusters that contain drugs or leads from COBRA (12 455 compounds), and orange dots show the location of the projected ChEMBL library (1 351 370 compounds). Note that ChEMBL data were not used for GTM training. Background shading reflects the position of the cluster centroids according to aromaticity, indicating the average aromaticity of the clustered compounds. The coloring ranges from white (no compound aromatic) to black (all compounds aromatic). Upper-case labels (**A**)–(**D**) highlight areas populated by natural products and drugs from both ChEMBL and COBRA (*cf.* **Fig. 5**), while lower-case labels (**a**)–(**e**) point to regions of chemical space containing natural products and ChEMBL compounds only (*cf.* **Fig. 4**). (Color figure available online only.)

every compound was projected at the posterior mean on the map. A higher number of RBFs resulted in a better fit to the original data, but at the same time increased the complexity of the model (**Fig. 2**). We decided that 100 RBFs represented an acceptable compromise to achieve a generalization of the compound distribution while still leading to an appropriate quality of the mapping. **Fig. 2** shows that our choice was reasonable in terms of the change of the RMSE against the number of RBFs.

To identify positions (clusters) on the GTM that are populated with natural products and drug-like molecules, we projected each of the compounds onto exactly one point on the map using the maximum posterior probability criterion. In agreement with an earlier study using SOM projections [20], the natural products and drug-like compounds intermixed strongly (**Fig. 3**). In fact, more than 60% (856 of 1424) of the natural product clusters also contained synthetic drugs. Importantly, these co-clustered regions contained more than 70% (110 788) of all the natural products. Only 10 drug-like compounds were projected to a total of six clusters that were not occupied by any natural product. While we observed mostly intermixed clusters of the COBRA drug set and natural products sharing pharmacophoric traits, natural products exclusively populated 568 other clusters. These clusters might point to pharmacophores that have only rarely been explored for synthetic drug design. In an attempt to further analyze this observation and clarify whether it is caused by the biased sample size, i.e., the small set of synthetic drugs compared to the much larger DNP set, we projected the ChEMBL compound collection [33] onto the map (the large number of ChEMBL entries prevented GTM retraining). We considered only 98.8% of the ChEMBL data (n = 1 351 370 without duplicate DNP entries) by avoiding projecting compounds that clearly reside outside of the applicability domain. We defined the model's applicability domain as the region in which the probability density of a molecule was larger than a threshold value corresponding to the 99.5 percentile of the training data (Eq. 3). The ChEMBL compounds spread almost over the complete spanned chemical product space, including many of the clusters not populated by COBRA compounds. It might therefore be a worthwhile exercise to systematically analyze the ChEMBL entries for activity annotations and hypothesize related activities for the co-located natural products (ongoing). Nevertheless, it is important to keep in mind that there are many natural products and derivatives in ChEMBL which are not part of the DNP (**Fig. 4**), and that the ChEMBL data were forced on the GTM trained only with DNP and COBRA. We consequently did not consider ChEMBL data for subsequent GTM analysis. This decision is supported by the statistically insignificant difference of compound properties between COBRA and ChEMBL (**Table 1**), which motivates the use of small, curated compound sets as surrogates for much larger collections.

We investigated representative natural products located in regions not populated by COBRA drugs (**Fig. 3 A, B**). In region A, we observed highly hydrophilic compounds [e.g., triaspidin (**1**, **Fig. 5**) in cluster (29/9)]. While this may be an undesired property in many drug discovery projects, pronounced lipophilicity is not necessarily a requirement for natural products as their secretion and uptake might be governed by different mechanisms or they might act intracellularly. GTM region B is populated by guaianolide derivatives [e.g., vestenolide (**2**) in cluster (7/13)]. This natural product scaffold has recently gained attention for multiple indications but is still only scarcely studied [34, 35]. These select examples reveal the potential of the map to suggest

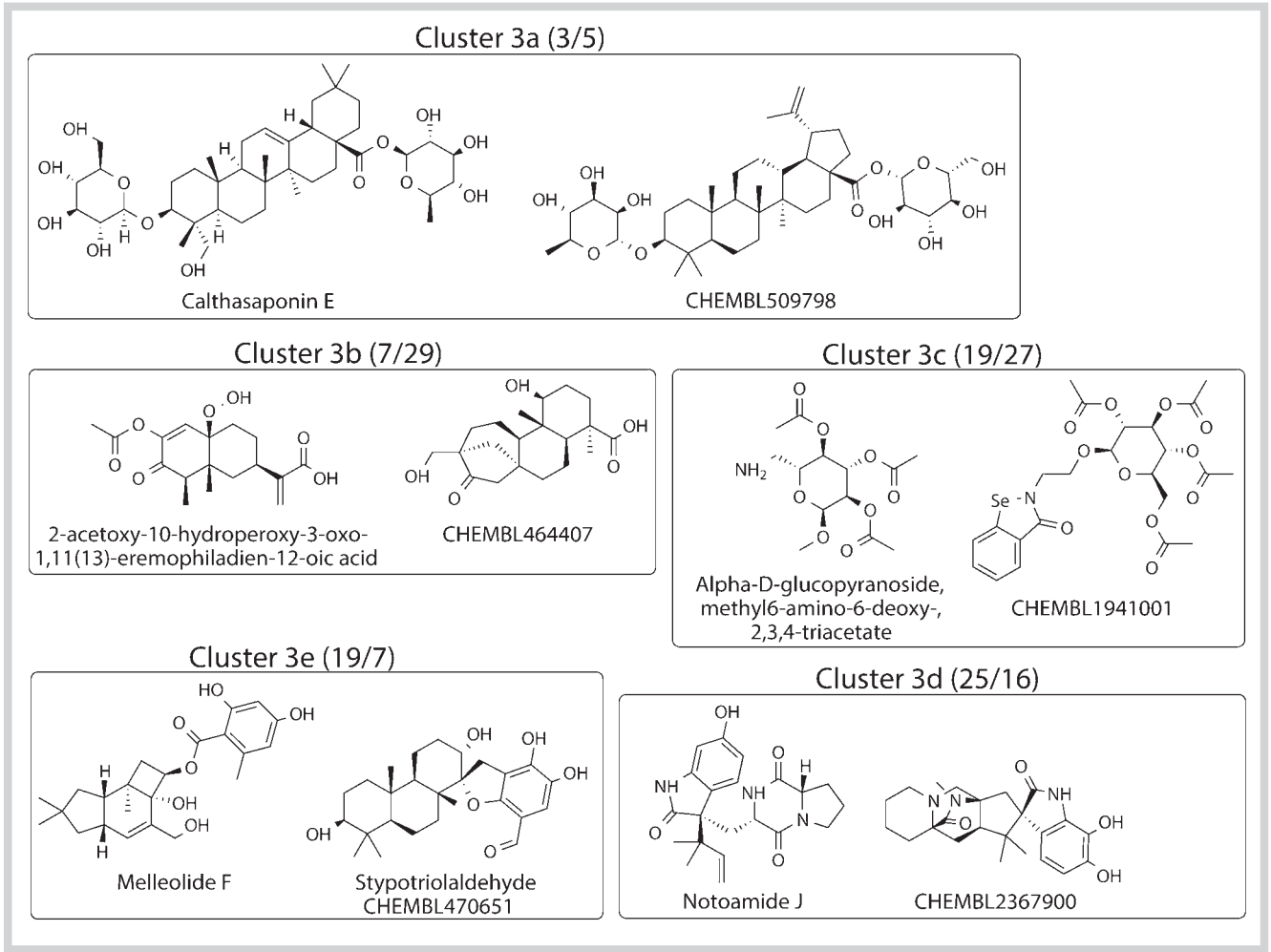

Fig. 4 Representative examples of natural products from the DNP co-located on the GTM with molecules from ChEMBL but not with COBRA compounds (*cf.* **Fig. 3a** to **e**). Most of these ChEMBL entries are also natural products or natural product derivatives that apparently have no similar relatives among the synthetic drugs from COBRA in terms of their pharmacophore feature patterns. Note that although the configuration of the chiral centers is shown, this information was not considered for computational analysis. Numbers in parentheses are (x/y) coordinates of the GTM clusters highlighted in **Fig. 3a** to **e**.

Table 1 Substructure counts and properties of the compound collections used in this study.

Data set (unique compounds)	H-bond acceptors	H-bond donors	Rotatable bonds	Rings	Slog*P*	MW
DNP (n = 157 929)	5.9 ± 5.8	2.9 ± 3.6	6.6 ± 6.9	3.5 ± 2.4	2.5 ± 3.3	447 ± 256
COBRA (n = 12 644)	3.3 ± 2.0	1.5 ± 1.5*	7.4 ± 5.1*	3.4 ± 1.4*	2.7 ± 2.7	417 ± 136*
ChEMBL (n = 1 368 655)	3.6 ± 3.5	1.5 ± 2.6*	7.7 ± 8.9*	3.4 ± 1.6*	2.8 ± 2.9	423 ± 245*

* No significant difference of the means (Welch's two-sided t-test, p value > 10^{-6})

new routes for filling pharmacophoric holes in synthetic compound libraries. We also compared natural products and drug-like compounds that are actually co-clustered. This analysis allowed us to assess whether functionally similar compounds were grouped together. GTM region C consists of prostaglandin derivatives from both the natural product collection [e.g., prostaglandin J2 (**3**) at position (9/35)] and the COBRA drug database [e.g., gemeprost (**4**) in cluster (8/34)]. Region D also features compounds that possess convincingly similar pharmacophore patterns in spite of apparent differences in their chemical structures. Two representative examples are the natural coumarin derivative mexoticin (**5**), which has been shown to inhibit platelet aggregation [36], and synthetic compound **6** [cluster (26/26)]. The coumarin-derived scaffold of mexoticin (**5**) contains the pharmacophore of psoralen, which is known to induce insomnia as a side effect in patients [37]. The co-clustered compound **6** is a synthetic orexin receptor antagonist that was developed as a treatment for sleep disorders [38]. Despite their apparent difference in chemical structure (MACCS-key based structural Tanimoto index = 0.27), the GTM suggests that the two compounds share a pharmacophore pattern. These observations motivate the testing of mexoticin for orexin receptor binding, and compound **6** for effects on platelet aggregation. In this way, the GTM may be used for predicting the macromolecular targets of natural products, in analogy to the SOM [11, 39] and property-based methods [22].

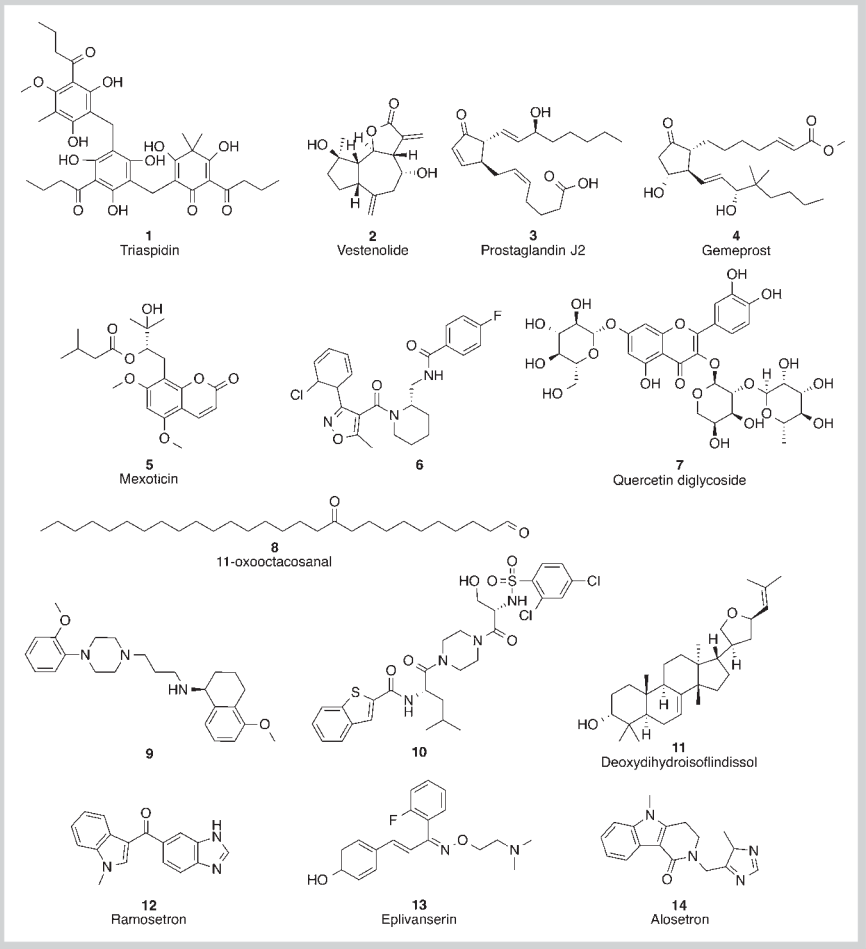

Fig. 5 The chemical structures mentioned in the text. Note that although the configuration of chiral centers is shown, this information was not considered for computational analysis.

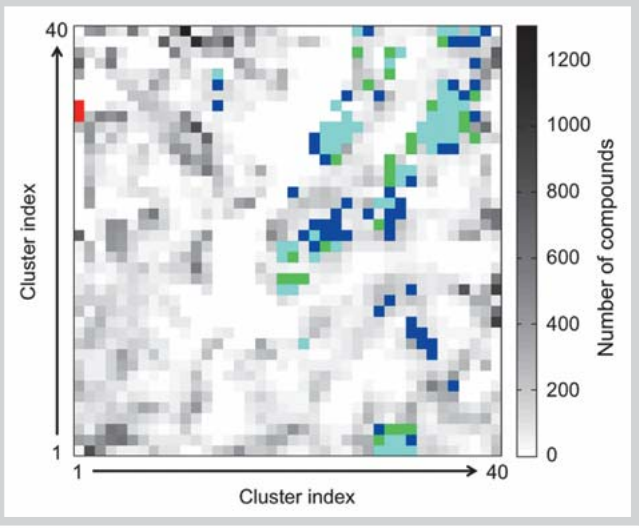

Fig. 6 GTM projection of compounds for selected targets (red: vitamin D receptor; blue: serotonin receptors; green: adrenergic receptors; cyan: overlap between serotonin and adrenergic receptors). Background coloring corresponds to the total number of compounds in each cluster. (Color figure available online only.)

We investigated whether the GTM appropriately spans the chemical space of synthetic COBRA compounds and natural products, and analyzed representative structures at regions that are heavily populated. We observed pronounced structural differences in these clusters. For example, a group of carbohydrate-containing natural products is represented by quercetin diglycoside (**7**) located at position (40/13) on the map. We found 11-oxooctacosanal (**8**) at position (11/40) representing a cluster of fatty acids and derivatives. In contrast to the natural products, the cluster representatives of synthetic drugs are not as easily structurally distinguishable. However, they represent ligands for different classes of biomolecules, for example GPCRs (e.g., compound **9**) or nuclear receptors (e.g., compound **10**). We aim at connecting the structural separation of natural products and target-specific regions on the map for the identification of meaningful relationships of natural product structures with synthetic ligands of pharmacologically relevant biomolecules. For example, we found DI [40] [position (1/32)] as a representative natural product in the region populated by vitamin D receptor ligands. The pharmacophore of DI closely resembles the one of vitamin D itself but constitutes a scaffold-hop from the secosteroids to the closed steroidal form. While this structural modification does not change the relative positioning of the pharmacophores, as correctly recognized by the CATS descriptor, it locks the hydrogen bond donor function in the 6-s-*cis* conformation [41]. It has been suggested that the 6-s-*cis* conformation is associated with the immediate response of vitamin D [42]. Further analysis of DI might reveal its role in different cellular processes and help to explain these effects. Chemographic methods can help in formulating motivated hypotheses for these experiments.

Ligands of serotonin (5-HT) receptors occur in six distinctive clusters on different positions on the map (**Fig. 6**). Interestingly, five of those are adjacent or intermixed with ligands of the adrenergic receptor. This is in line with their known pharmacological cross-activity [43]. However, one of the clusters is completely disconnected from any area containing adrenergic receptor ligands [around position (32/12)]. We investigated this area further and in fact found several compounds known to be selective for certain 5-HT receptor subtypes without activity on the adrenergic receptor family, for example alosetron [position (32/12)] [44], eplivanserin [position (33/11)] [45], and ramosetron [position (34/11)] [44]. Apparently, the resolution of the map is high enough to distinguish such subtle changes in pharmacophores.

GTMs are gaining increasing attention in cheminformatics [25, 46]. In natural product-related studies, GTMs have been applied to applications outside of drug discovery, for example, in the analysis of the content of fish oil extracts [47]. An exception is the study by Owen et al. that used structural MACCs key fingerprints to distinguish drugs, combinatorial synthetic compounds, and natural products [48]. Here, we have introduced GTMs as a technique for dimensionality reduction and target prediction for natural products based on molecular pharmacophore representations. We show that this concept allows for relating natural products and synthetic compounds in spite of clearly observable structural differences, fully in line with results previously acquired with SOMs [11, 20]. Results suggest that the resolution of a GTM is sufficient to identify functional relationships between natural products and synthetic drugs. The concept of analyzing regions of natural product space for a lack of synthetic compounds or the presence of compounds with a desired polypharmacological profile will make such chemographic methods a helpful tool for natural product-inspired drug discovery.

Materials and Methods

Data

For GTM analysis, we compiled the natural products contained in the Chapman & Hall/CRC Dictionary of Natural Products (DNP v20.1 DNP, 210 273 compounds; http://dnp.chemnetbase.com) [49]. Drugs and drug-like bioactive compounds were taken from COBRA (v12.6, 13,702 compounds; inSili.com LLC) [30]. ChEMBL compound data were compiled from database version 19 (https://www.ebi.ac.uk/chembl/) [33]. We removed all duplicates (5088 compounds) and structures also present in the DNP (31 009 compounds) from the ChEMBL collection, which resulted in 1 368 655 remaining ChEMBL entries. All molecules were preprocessed with the MOE wash node (v2011.10, Chemical Computing Group) as implemented in KNIME v2.9.4 [50] using the options "protonate strong bases", "deprotonate strong acids", "remove minor components", "disconnect salts", and "remove lone pairs". Duplicate structures were removed by grouping according to canonical SMILES representations. This procedure resulted in 157 929 compounds from the DNP and 12 644 compounds from the COBRA database. All molecules were described in terms of pharmacophore patterns using our in-house CATS2 descriptor implementation with a correlation distance of 0–9 bonds and type-sensitive scaling [31]. Consequently, each molecule was represented by a 210-dimensional topological pharmacophore representation.

Generative topographic map

In GTMs, every latent point is connected to the point in the original data space according to Eq. 2. Simultaneously, the projected point is the mean value of a Gaussian and the sum of these Gaussians describes the data distribution (Eq. 3).

$$y(I_i, W) = W\Phi(I_i) \quad (2)$$

$$p(x|I, W, \beta) = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}}\exp\left\{-\frac{\beta}{2}\| x - y(I_i, W)\|^2\right\} \quad (3)$$

In Eq. 2, $\mathbf{I}_i$ is a latent point, K is the number of latent points, and $\mathbf{\Phi}$ is an M-dimensional vector consisting of RBFs evaluated at $\mathbf{I}_i$. The matrix parameter $\mathbf{W}$ ($D \times M$) governs the projection from a point in the latent space to the point in the original data space, where D is the dimensionality (descriptor vector cardinality) of the original data space. In Eq. 3, the sum of Gaussian distributions gives the probability distribution in the data space. Each Gaussian has the mean value $y(\mathbf{I}_i,\mathbf{W})$ with variance β^{-1}. By using the EM algorithm, locally optimized parameters ($\mathbf{W}$, β) were obtained. We trained a GTM with $(40 \times 40) = 1600$ latent points that served as Gaussian cluster centers. The maps were constructed using the GTM toolbox v1.0 [51] from Matlab 2014a (The MathWorks, Inc.). During the training, we used 100 RBFs that were aligned on the lattice in latent space. The width of the RBFs was set to the distance between neighboring RBF centers. The regularization parameter was set to 0.001. We initialized the map through PCA of the original data. Note that for the estimation of the standard deviation of the mapping error (RMSE, Eq. 1), we randomly initialized the mapping parameters ($\mathbf{W}$, β). Applicability domain was determined based on Eq. 3 with optimized parameters. We set a 99.5 percentile of the values of descending ordered training data as the threshold value.

Acknowledgements

ETH Zürich and the OPO Foundation Zurich funded this research study.

Conflict of Interest

G.S. and P.S. are cofounders of inSili.com LLC, Zürich, and act as consultants in the pharmaceutical and chemical industry.

References

1 *Koehn FE, Carter GT.* The evolving role of natural products in drug discovery. Nat Rev Drug Discov 2005; 4: 206–220
2 *Harvey AL.* Natural products in drug discovery. Drug Discov Today 2008; 13: 894–901
3 *Ganesan A.* The impact of natural products upon modern drug discovery. Curr Opin Chem Biol 2008; 12: 306–317
4 *Paterson I, Anderson EA.* The renaissance of natural products as drug candidates. Science 2005; 310: 451–453
5 *Picket S.* The biophore concept. In: Böhm HJ, Schneider G, editors. Protein-ligand interactions: from molecular recognition to drug design. Weinheim: Wiley-VCH; 2003: 73–105
6 *Newman DJ, Cragg GM, Snader KM.* Natural products as sources of new drugs over the period 1981–2002. J Nat Prod 2003; 66: 1022–1037
7 *Basmadjian C, Zhao Q, Bentouhami E, Djehal A, Nebigil CG, Johnson RA, Serova M, de Gramont A, Faivre S, Raymond E, Désaubry LG.* Cancer wars: natural products strike back. Front Chem 2014; 2: 20

8 *Grienke U, Mihály-Bison J, Schuster D, Afonyushkin T, Binder M, Guan SH, Cheng CR, Wolber G, Stuppner H, Guo DA, Bochkov VN, Rollinger JM.* Pharmacophore-based discovery of FXR-agonists. Part II: identification of bioactive triterpenes from *Ganoderma lucidum*. Bioorg Med Chem 2011; 19: 6779–6791

9 *Rollinger JM, Schuster D, Danzl B, Schwaiger S, Markt P, Schmidtke M, Gertsch J, Raduner S, Wolber G, Langer T, Stuppner H. In silico* target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. Planta Med 2009; 75: 195–204

10 *Hufsky F, Scheubert K, Böcker S.* New kids on the block: novel informatics methods for natural product discovery. Nat Prod Rep 2014; 31: 807–817

11 *Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Mönch B, Koeberle A, Lamers C, Gabler M, Steinmetz H, Müller R, Schubert-Zsilavecz M, Werz O, Schneider G.* Revealing the macromolecular targets of complex natural products. Nat Chem 2014; 6: 1072–1078

12 *Oprea TI, Gottfries J.* Chemography: the art of navigating in chemical space. J Comb Chem 2001; 3: 157–166

13 *Todeschini R, Consonni V.* Molecular descriptors for chemoinformatics. Weinheim: Wiley-VCH; 2009

14 *Reutlinger M, Schneider G.* Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. J Mol Graph Model 2012; 34: 108–117

15 *Ivanenkov YA, Savchuk NP, Ekins S, Balakin KV.* Computational mapping tools for drug discovery. Drug Discov Today 2009; 14: 767–775

16 *Reutlinger M, Rodrigues T, Schneider P, Schneider G.* Multi-objective molecular *de novo* design by adaptive fragment prioritization. Angew Chem Int Ed Engl 2014; 53: 4244–4248

17 *Lachance H, Wetzel S, Kumar K, Waldmann H.* Charting, navigating, and populating natural product chemical space for drug discovery. J Med Chem 2012; 55: 5989–6001

18 *Wetzel S, Schuffenhauer A, Roggo S, Ertl P, Waldmann H.* Cheminformatic analysis of natural products and their chemical space. Chimia 2007; 61: 355–360

19 *Grabowski K, Baringhaus KH, Schneider G.* Scaffold diversity of natural products: inspiration for combinatorial library design. Nat Prod Rep 2008; 25: 892–904

20 *Lee ML, Schneider G.* Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. J Comb Chem 2001; 3: 284–289

21 *Olah M, Mracec M, Ostopovici L, Rad RF, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI.* WOMBAT: World of Molecular Bioactivity. In: Oprea TI, editor. Cheminformatics in drug discovery. New York: Wiley -VCH; 2005: 223–239

22 *Rosen J, Gottfries J, Muresan S, Backlund A, Oprea TI.* Novel chemical space exploration via natural products. J Med Chem 2009; 52: 1953–1962

23 *Bishop CM, Svensen M, Williams CKI.* GTM: The generative topographic mapping. Neural Comput 1998; 10: 215–234

24 *Hasegawa K, Funatsu K.* Prediction of protein-protein interaction pocket using L-shaped PLS approach and its visualizations by generative topographic mapping. Mol Inf 2014; 33: 65–72

25 *Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A.* Generative Topographic Mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. Mol Inf 2012; 31: 301–312

26 *Ovchinnikova SI, Bykov AA, Tsivadze AY, Dyachkov EP, Kireeva NV.* Supervised extensions of chemography approaches: case studies of chemical liabilities assessment. J Cheminform 2014; 6: 20

27 *Bishop CM, Svensen M, Williams CKI.* GTM: a principal alternative to the self-organizing map. In: Mozer MC, Jordan MI, Petsche T, editors. Advances in neural Information Processing Systems 9: Proceedings of the 1996 Conference. Cambridge (MA): MIT Press; 1997: 354–360

28 *Kruskal JB.* Multidimensional-scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964; 29: 1–27

29 *Frey BJ, Jojic N.* A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Trans Pattern Anal Mach Intell 2005; 27: 1392–1416

30 *Schneider P, Schneider G.* Collection of bioactive reference compounds for focused library design. QSAR Comb Sci 2003; 22: 713–718

31 *Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G.* Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. Mol Inf 2013; 32: 133–138

32 *Schneider G, Neidhart W, Giller T, Schmid G.* "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. Angew Chem Int Ed Engl 1999; 38: 2894–2896

33 *Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP.* ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012; 40: 1100–1107

34 *Wesołowska A, Nikiforuk A, Michalska K, Kisiel W, Chojnacka-Wójcik E.* Analgesic and sedative activities of lactucin and some lactucin-like guaianolides in mice. J Ethnopharmacol 2006; 107: 254–258

35 *Zhang Q, Lu YX, Ding YH, Zhai JD, Ji Q, Ma WW, Yang M, Fan HX, Long J, Tong ZS, Shi YH, Jia YS, Han B, Zhang WP, Qiu CJ, Ma XY, Li QY, Shi QQ, Zhang HL, Li DM, Zhang J, Lin JP, Li LY, Gao YD, Chen Y.* Guaianolide sesquiterpene lactones, a source to discover agents that selectively inhibit acute myelogenous leukemia stem and progenitor cells. J Med Chem 2012; 55: 8757–8769

36 *Chen KS, Wu CC, Chang FR, Chia YC, Chiang MY, Wang WY, Wu YC.* Bioactive coumarins from the leaves of *Murraya omphalocarpa*. Planta Med 2003; 69: 654–657

37 *Koo JYM, Maloney J.* Drug therapy of psoriasis. In: Millikan LE, editor. Drug therapy in dermatology. New York: Marcel Dekker; 2000: 305–322

38 *Palomer A, Princep M, Guglietta A.* Recent advances in the treatment of insomnia. In: Macor JE, editor. Annual reports in medicinal chemistry, Vol 42. San Diego: Elsevier Academic Press Inc.; 2007: 63–80

39 *Reker D, Rodrigues T, Schneider P, Schneider G.* Identifying the macromolecular targets of *de novo*-designed chemical entities through self-organizing map consensus. Proc Natl Acad Sci U S A 2014; 111: 4067–4072

40 *Birch AJ, Collins DJ, Muhammad S, Turnbull J.* The structure of flindissol. Some remarks on the elemi acids. J Chem Soc 1963; 2762–2772

41 *Norman AW.* Minireview: vitamin D receptor: new assignments for an already busy receptor. Endocrinology 2006; 147: 5542–5548

42 *Bikle DD.* Vitamin D metabolism, mechanism of action, and clinical applications. Chem Biol 2014; 21: 319–329

43 *Guan XM, Peroutka SJ, Kobilka BK.* Identification of a single amino acid residue responsible for the binding of a class of beta-adrenergic receptor antagonists to 5-hydroxytryptamine 1A receptors. Mol Pharmacol 1992; 41: 695–698

44 *Hirata T, Keto Y, Funatsu T, Akuzawa S, Sasamata M.* Evaluation of the pharmacological profile of ramosetron, a novel therapeutic agent for irritable bowel syndrome. J Pharmacol Sci 2007; 104: 263–273

45 *Rinaldi-Carmona M, Congy C, Santucci V, Simiand J, Gautret B, Neliat G, Labeeuw B, Le Fur G, Soubrie P, Breliere JC.* Biochemical and pharmacological properties of SR 46349B, a new potent and selective 5-hydroxytryptamine2 receptor antagonist. J Pharmacol Exp Ther 1992; 262: 759–768

46 *Maniyar DM, Nabney IT, Williams BS, Sewing A.* Data visualization during the early stages of drug discovery. J Chem Inf Model 2006; 46: 1806–1818

47 *Aursand M, Standal IB, Axelson DE.* High-resolution 13C nuclear magnetic resonance spectroscopy pattern recognition of fish oil capsules. J Agric Food Chem 2007; 55: 38–47

48 *Owen JR, Nabney IT, Medina-Franco JL, López-Vallejo F.* Visualization of molecular fingerprints. J Chem Inf Model 2011; 51: 1552–1563

49 The Chapmann & Hall/CRC Chemical Database Dictionary of Natural Products. Chapman and Hall/CRC. Available at http://dnp.chemnetbase.com/.

50 *Berthold MR, Cebron N, Dill F, Gabriel TR, Koetter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B.* KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data analysis, machine learning and applications. Berlin: Springer; 2008: 319–326

51 *Svensen M.* The GTM Toolbox – User's Guide, Neural Computing Research Group. Birmingham: Aston University; 1999