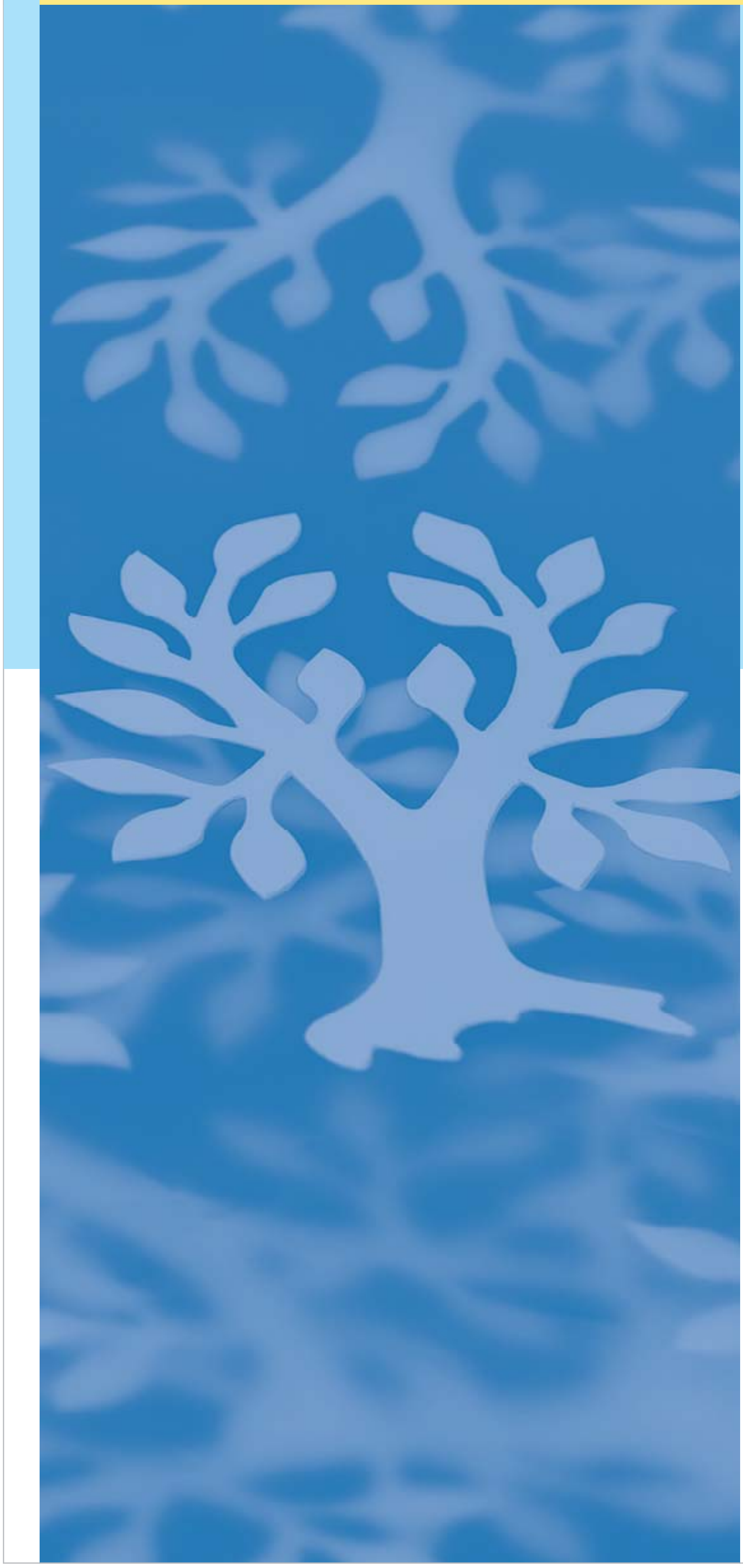

Unique Macrocycles in the Taiwan Traditional Chinese Medicine Database

DOI 10.1055/s-0035-1545881
Planta Med 2015; 81: 459–466

Publisher and Copyright:
© 2015 by
Georg Thieme Verlag KG
Rüdigerstraße 14
70469 Stuttgart
ISSN 0032-0943

Unique Macrocycles in the Taiwan Traditional Chinese Medicine Database

Authors
Christian Kramer[1], Maren Podewitz[1], Peter Ertl[2], Klaus R. Liedl[1]

Affiliations
[1] Center for Molecular Biosciences Innsbruck (CMBI), University of Innsbruck, Innsbruck, Austria
[2] Novartis Institutes for BioMedical Research, Basel, Switzerland

Key words
- chemical space
- natural product
- macrocycle
- TCM database@Taiwan

received August 11, 2014
revised Dec. 20, 2014
accepted February 23, 2015

Bibliography
DOI http://dx.doi.org/10.1055/s-0035-1545881
Published online April 9, 2015
Planta Med 2015; 81: 459–466
© Georg Thieme Verlag KG
Stuttgart · New York ·
ISSN 0032-0943

Correspondence
Dr. Christian Kramer
Institute of General, Inorganic and Theoretical Chemistry
Center for Molecular Biosciences Innsbruck (CMBI)
University of Innsbruck
Innrain 80/82
6020 Innsbruck
Austria
Phone: + 43 5 12 50 75 71 03
Christian.Kramer@uibk.ac.at

Correspondence
Dr. Peter Ertl
Novartis Institutes for BioMedical Research
Novartis Campus
4056 Basel
Switzerland
Peter.Ertl@novartis.com

Abstract

Chemical space coverage within natural product databases is an important criterion for the selection of databases for virtual screening. Chemical space analysis can also provide valuable hints towards chemical substructures that are responsible for medical activity. We therefore developed a protocol for structurally characterizing the chemical space covered in specific natural product databases by comparing it to a "standard" natural product scaffold distribution. In this contribution, we analyzed the structural characteristics of the traditional Chinese medicine database@Taiwan as an example. While we did not find classes of very characteristic small molecule scaffolds, we found that there are a number of specific macrocyclic scaffolds that are highly enriched in the traditional Chinese medicine database@Taiwan and not documented elsewhere. This surprising finding points towards underused regions in chemical space with a big potential for biological impact.

Abbreviations

JNP:	Journal of Natural Products
NP:	natural product
SCONP:	structural classification of natural products
Taiwan TCM:	traditional Chinese medicine database@Taiwan
TCM:	traditional Chinese medicine

Introduction

NPs are attractive starting points for drug development, since they evolutionarily developed to influence biological systems. A large number of approved small molecule drugs have directly been developed from or inspired by NPs [1,2]. Since NPs play such a pivotal role, medical, biological, and chemical knowledge about NPs from various organisms has been collected in different databases. The chemical space represented in the databases can be broad or focused, depending on the source organism, geographical region, or cultural background that has been used to guide the collection. The compounds contained in the databases will also differ depending on the database focus and the extraction and chemical workup procedures applied.

Taiwan TCM is one of the most prominent NP databases that is freely available [3]. On the 22nd of July 2014, the publication describing the Taiwan TCM that had been published in January 2011 had already been cited 160 times according to Google Scholar. The purpose of the Taiwan TCM database is to link the medical knowledge from TCM to the chemical structures that are known to exist within the plants, fungi, animals, and mineral materials used as TCM drugs. Since TCM represents a potentially rich source for linking chemical structures to medical activity, we set out to identify the specific characteristics of the chemical space covered in the Taiwan TCM database.

There are two principally different approaches for characterizing chemical space: (i) descriptor- and fragment-based and (ii) scaffold-based. The historically older approach is based on molecular descriptors such as physicochemical properties and substructure counts and the analysis of the distribution of compounds that belong to different classes within the space spanned by these descriptors. Many databases have been analyzed by such an approach, including the Taiwan TCM database [4] and the TCM Compound database [5].

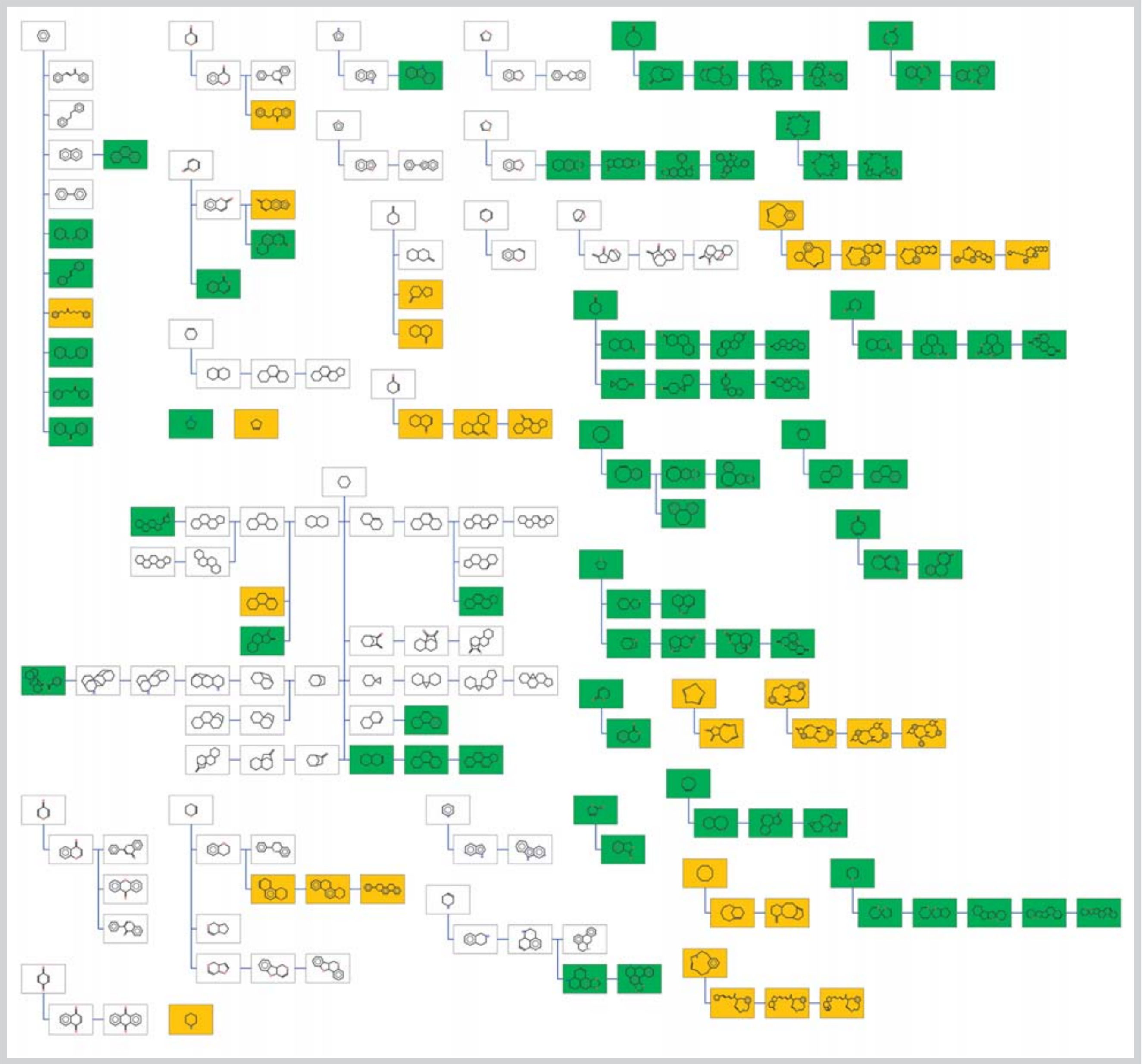

Fig. 1 Scaffold tree analysis of the joint JNP and Taiwan TCM databases. Only scaffolds are shown that occur in at least 0.05% of the joint databases. The scaffolds that are enriched by at least a factor of ten in one of the databases are shown in yellow (light grey) (Taiwan TCM) and green (dark grey) (JNP). Uncolored scaffolds appear in both databases with roughly equal proportions. (Color figure available online only.)

Ertl et al. have used this type of chemical space description to create an NP likeness score that can be used to prioritize compound libraries [6,7]. The second approach is based on a comparison of the chemical scaffolds found in different databases. Schuffenhauer et al. have developed the scaffold tree method for mapping and visualizing the chemical scaffolds contained in chemical databases [8]. Based on the scaffold tree, Koch et al. developed the SCONP approach to chart the NP scaffold space [9]. Advantages, disadvantages, and findings from these two complementary approaches have been discussed in a review by Wetzel et al. [10].

Here we use the SCONP approach to compare the scaffold space of the Taiwan TCM database to a general distribution of NP scaffolds. The goal of this analysis initially was to identify specific small molecule scaffolds that are enriched in the Taiwan TCM database as a starting point for further investigations into scaffolds with a high propensity for medical impact. Surprisingly, we did not find enriched small molecule scaffolds that are characteristic for the TCM chemical space. However, we did find a number of exclusive macrocyclic scaffolds with unique structural motifs.

Results

The results of the scaffold tree analysis are shown in ▶ **Fig. 1**, in which only scaffolds that occur in at least 0.05% of the overall dataset are presented. These are rather small scaffolds. If the threshold for showing scaffolds is lowered, the scaffold tree grows quickly and is hardly visualizable any more. From ▶ **Fig. 1**, it can be seen that there are only a few scaffolds that appear to be specific for the Taiwan TCM database. In a next step, we searched

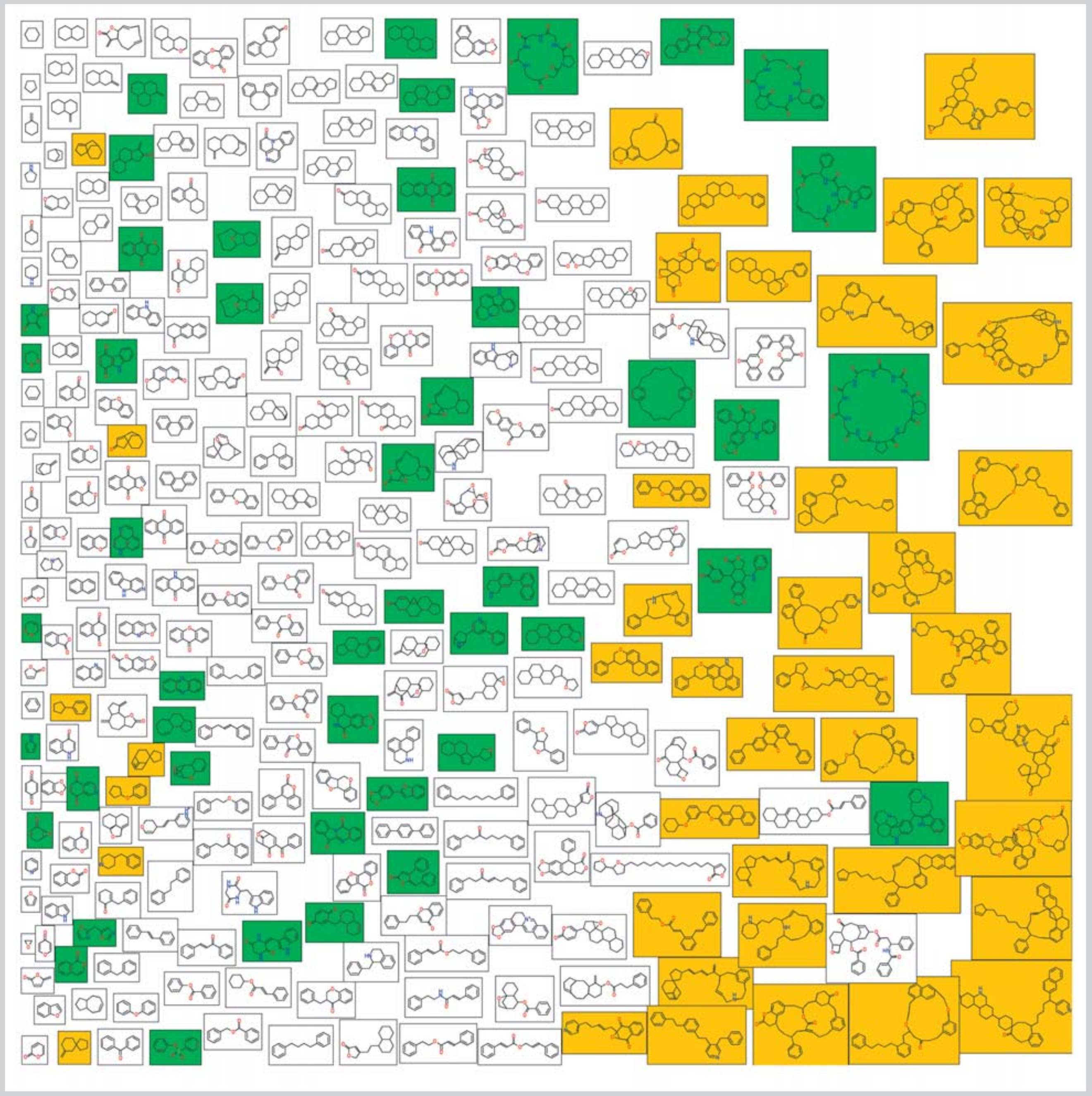

Fig. 2 Scaffold map of the most frequent scaffolds for the joint JNP and Taiwan TCM databases. Scaffolds are sorted according to size (x-axis) and combination of other scaffold keys that roughly represent molecular complexity (y-axis). The scaffolds that are enriched by at least a factor of ten in one of the databases are shown in yellow (light grey) (Taiwan TCM) and green (JNP). Uncolored scaffolds appear in both databases with roughly equal proportions. (Color figure available online only.)

public databases for the small scaffolds identified in the first step. All of the NPs containing these scaffolds also frequently occur in ChEMBL18, however, they were often not published in the JNP, but in the *Journal of Medicinal Chemistry* or other medicinal chemistry journals. Also, those scaffolds come from rather diverse backgrounds. Therefore, those scaffolds have to be considered as not characteristic for the Taiwan TCM database.

Next, we turned to scaffold maps ordered by scaffold keys [11] for visualizing the scaffold universe. ▶ **Fig. 2** shows a scaffold map for the joint databases, colored by the difference of occurrences of the scaffolds in both Taiwan TCM and JNP databases.

▶ **Fig. 2** shows that scaffold enrichment is highly size dependent. For the very small scaffolds on the left of ▶ **Fig. 2**, there are seven Taiwan TCM database scaffolds that are enriched. Upon closer inspection, all of those turned out to also frequently occur in ChEMBL18, however, they were published in different journals than JNP. For scaffolds with a median size in the middle of ▶ **Fig. 2**, there is no scaffold that is enriched within the Taiwan TCM database. However, for the large scaffolds towards the right side of ▶ **Fig. 2**, there are a lot of scaffolds that are enriched within the Taiwan TCM database. In order to analyze those further, we created a scaffold map based on simple ring systems. It is shown in ▶ **Fig. 3**.

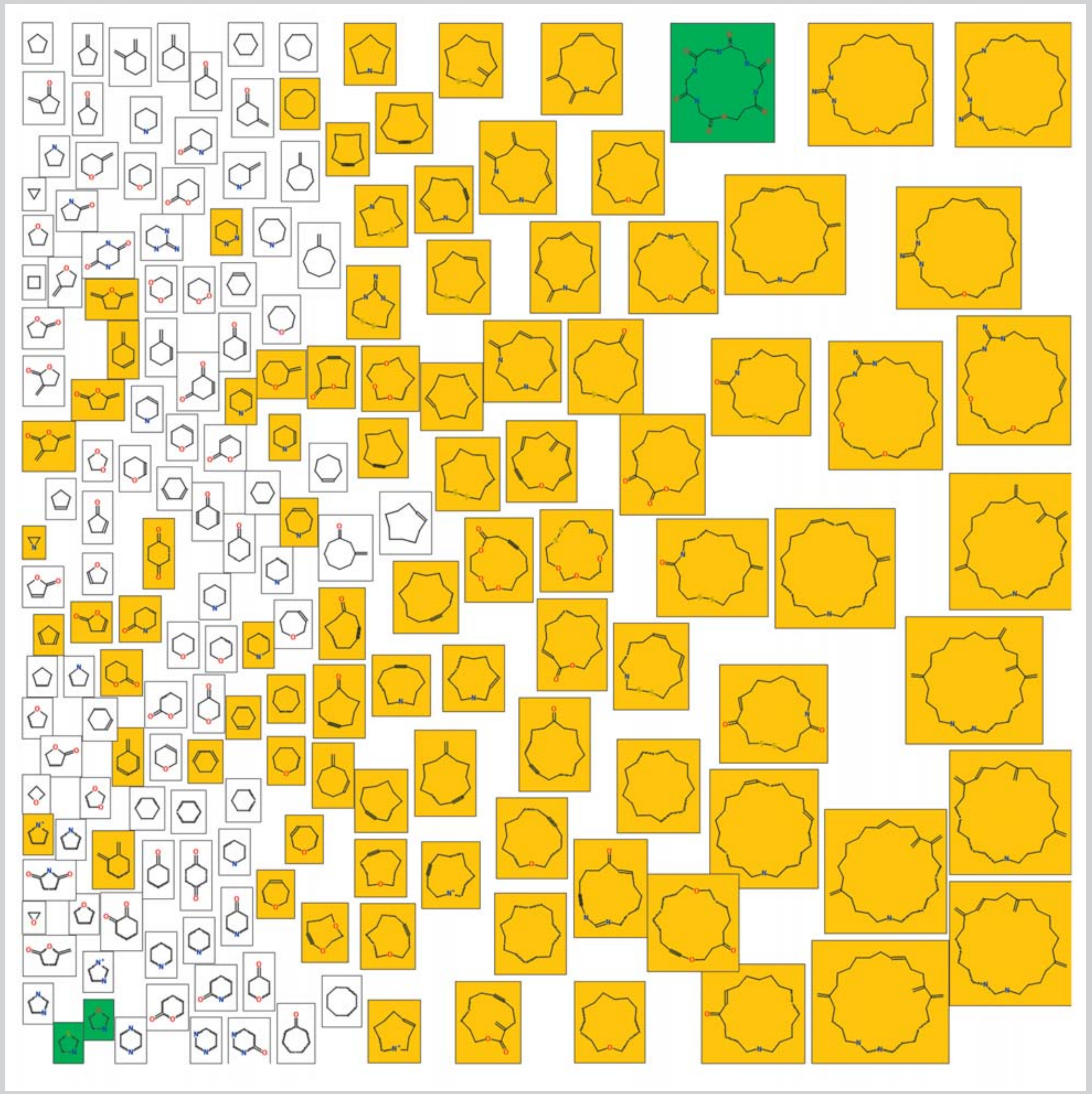

Fig. 3 Scaffold map based on simple ring systems for the joint Taiwan TCM and JNP compound databases. Rings are sorted according to size (x-axis) and combination of other scaffold keys representing roughly molecular complexity (y-axis). The scaffolds that are enriched by at least a factor of ten in one of the databases are shown in yellow (Taiwan TCM) and green (dark grey) (JNP). Uncolored scaffolds appear in both databases with roughly equal proportions. (Color figure available online only.)

▶ **Fig. 3** shows that there are a lot of highly enriched macrocycles within the Taiwan TCM database. In order to further analyze those, we extracted the most discriminative macrocycles with the strongest enrichments from both databases. They are shown in ▶ **Fig. 4**.

▶ **Fig. 4 a** shows that 11 out of the 16 most discriminative ring systems of the JNP compounds are peptidic macrocycles, whereas none of the most discriminative Taiwan TCM macrocycles is peptidic. Also, there are a lot more samples for the Taiwan TCM macrocycles (400 : 0 to 167 : 0 for the 16 most discriminative macrocycles) than for the JNP macrocycles (110 : 1 to 14 : 0 for the 16 most discriminative macrocycles). Moreover, the Taiwan TCM macrocycles are quite special, since they contain triple bonds in the ring system. Three compounds from the Taiwan TCM that contain such macrocycles are shown in ▶ **Fig. 5**.

Since the macrocycles occurred as a special feature of the Taiwan TCM database, we searched other databases (Scifinder, PubChem, ChEMBL) for these features. For many macrocycles, we did not find a specific hit. We then concentrated on the macrocycles with triple bonds to find out whether this rather special feature occurs in other NP compounds. We found few compounds that also contain triple bonds in the macrocyle, including namenamicin, nostocyclyne, shishijimicin, calicheamicin, dynemicin, esperamicin,

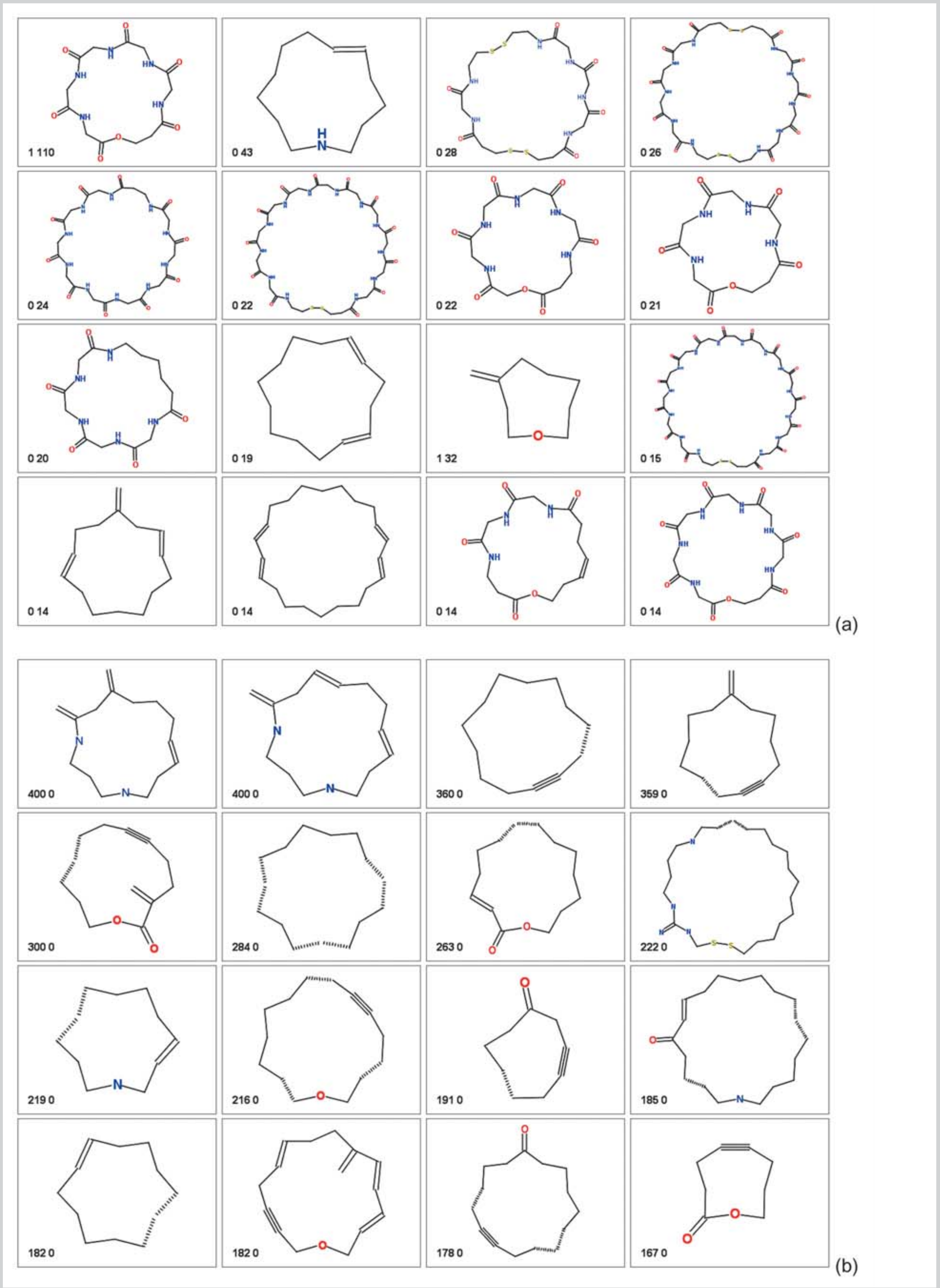

Fig. 4 The 16 most enriched macrocycles in the JNP (**a**) and Taiwan TCM (**b**) databases. The first number indicates the number of Taiwan TCM compounds that contain this substructure; the second number indicates the number of JNP compounds with this substructure. The dashed bonds indicate that this bond is part of an aromatic system. (Color figure available online only.)

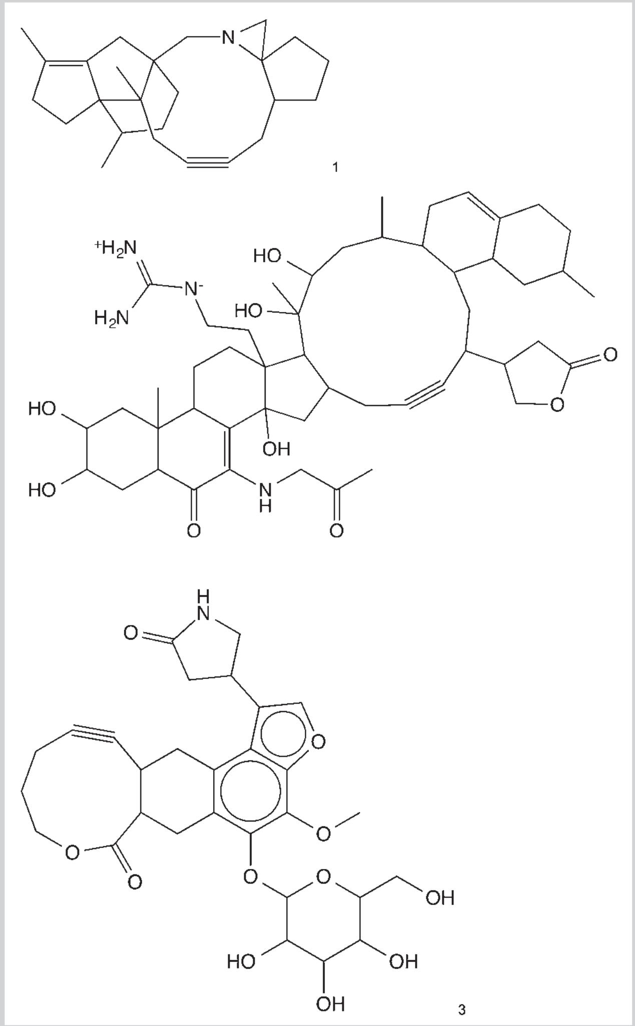

Fig. 5 Three sample structures from the Taiwan TCM database that contain macrocycles with triple bonds in the ring system.

kedarcidin, and neocarzinostatin. Structural formulae for calicheamicin and neocarzinostatin are shown in ▶ **Fig. 6**.

However, there are a lot more structures with non-peptidic macrocycles and triple bonds in the Taiwan TCM database that are not documented in any other database we searched.

Discussion

In this study, we searched for scaffolds that are characteristic for the Taiwan TCM database in contrast to a background distribution taken from JNP. We only found very few small molecule scaffolds that at first glance appear to be characteristic for Taiwan TCM, and comparison with the ChEMBL database showed that this is because the NPs that contain these substructures have been published in other journals than the JNP.

However, we also found that there are a lot of macrocyclic substructures that are highly enriched in the Taiwan TCM database. A substructure search for these macrocycles in the ChEMBL, PubChem, and SciFinder databases gave zero results. Thus, to the best of our knowledge, these macrocycles are only documented in the Taiwan TCM database, and potentially in some very specific local sources.

Since this is such a significant finding, we first tried to verify the structures. The compounds in the downloadable Taiwan TCM database are given as *.mol2* files, however, without an identifier. Thus, it is not possible to trace back where the structure comes from – neither the organism nor the medical indication. We next tried to use the substructure search function on the Taiwan TCM database homepage and contact the author. However, this was also without success. Then we checked whether there are other known NPs that contain macrocycles with these specific substructures, in particular with the triple bonds in the ring system. While there is no example with exactly the same substructure, we did find macrocycles that contain triple bonds, as shown in ▶ **Fig. 6**. Therefore, we consider it possible that the structures from the Taiwan TCM database really exist; yet we are not able to verify them. It would be extremely helpful for such verification if the source of the chemical structures was linked within the structure file.

If the structures given in the Taiwan TCM really exist, they point to a part of NP space that has not been charted yet and that might be very relevant in terms of compounds that mediate biological activity. Properties of macrocycles as drug candidates and synthetic challenges have been reviewed before [12–15]. The best-characterized macrocycles that contain triple bonds in the ring are enediynes. The enediyne system is highly reactive and can cleave DNA via hydrogen abstraction from the backbone sugar atoms [16]. However, the structures that are characteristic for the Taiwan TCM database are not enediynes. In the Taiwan TCM macrocycles we found, the triple bond could have a structural function; the sp-bond dictates the local geometry and could also drastically constrain the number of conformations possible, thereby introducing some non-common ring conformations. Initial test calculations on the structures that are unique to the Taiwan TCM database indicate that indeed there are very few conformations possible for the ring systems, giving rise to very well-defined pharmacophores. However, without being able to verify the structures and the stereochemistry of the Taiwan TCM database, we refrain from detailing the analysis at this point in time.

In conclusion, we compared the scaffold distribution of the Taiwan TCM database to a generic NP scaffold distribution represented by all compounds published in the JNP. While we did not find any characteristic "small molecule" scaffold in the Taiwan TCM database, we found that the Taiwan TCM contains a large number of unique macrocycles, many of them with triple bonds in the ring system. Attempts to verify the scaffolds in other databases such as SciFinder, PubChem or ChEMBL failed. Most probably, the structures are documented in specific local sources.

If the structures really exist, they point to a part of chemical space that may have strong biological relevance but is completely undocumented in the chemistry literature indexed in standard chemistry databases. Since secondary metabolite NPs have evolutionarily been designed to cause biological effects, the specific macrocyclic compounds with their potentially well-defined pharmacophores are an interesting field for further study.

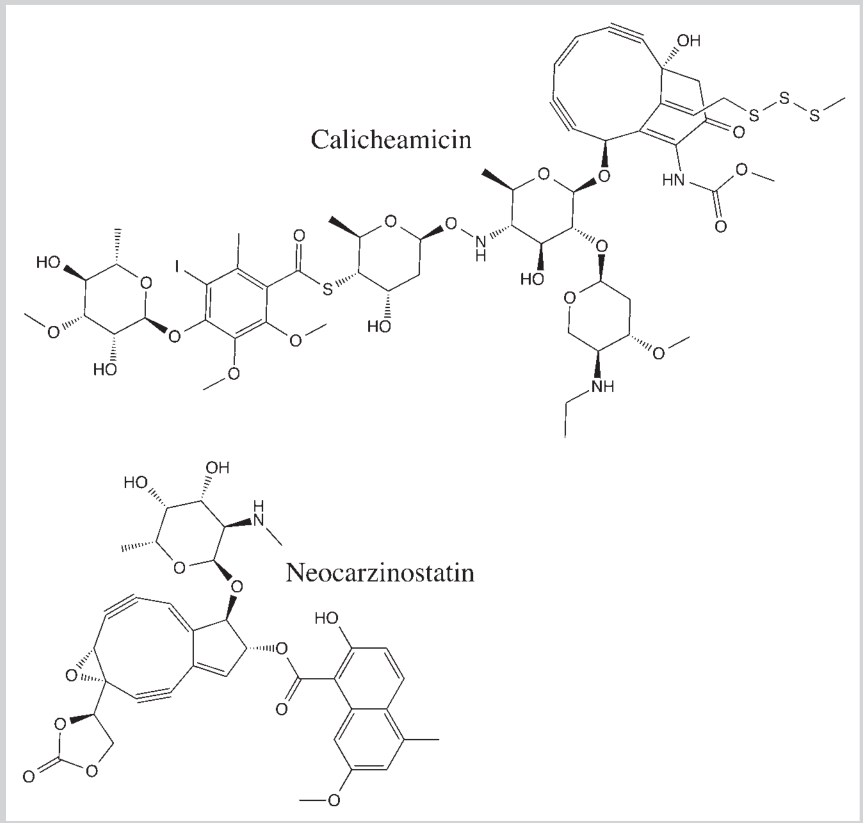

Fig. 6 Well-documented macrocyclic compounds that contain triple bonds in the macrocycle.

Materials and Methods

Databases for comparison

We downloaded the Taiwan TCM database on the 2nd of July 2014. It was assembled by Calvin Yu-Chian Chen and coworkers [3] and contains 57 423 molecule entries, each as a single *.mol2* file. The Taiwan TCM database can be browsed online, but the files, unfortunately, do not contain any identifier that would allow linking the individual structures back to a chemical name, organism, or medical indication.

As an NP reference distribution, we extracted all of the compounds that were originally published in the JNP from the ChEMBL18 database [17]. ChEMBL contains all compounds that were published in the JNP, and where at least one bioactivity measurement is available (even if there was no activity measurable). Overall, we extracted 39 534 JNP compounds. Since the JNP does not have a focus on specific organisms, this dataset can be considered a standard reference distribution.

Database preparation

Molecules were first converted from *.mol2* files into SMILES representation, and all subsequent operations were performed with the SMILES. During this step, the structures were also cleaned, neutralized, and molecules with valence errors were removed. Also, stereo features from molecules were removed. Although rich stereochemistry is typical for NPs, distinguishing them clearly from standard synthetic molecules, it is also a feature that causes many problems in database comparison, since stereochemistry is often only partially or not at all defined and sometimes even incorrect. Since our major interest was the analysis of scaffolds without substituents (removal of substituents in most cases destroys stereocenters), this step was justified.

The last step in molecule processing was *in silico* deglycosylation, i.e., removal of sugar rings from the molecules. The presence of sugar rings which are often multiply connected into a complex tree pattern is one of the most typical structural features of NPs. The main role of sugar moieties in NPs is to affect pharmacokinetic properties of parent molecules and make them more soluble. In most cases, sugar units do not affect the biological activity of the aglycon directly (although several notable exceptions to this rule exist). Because we did not want numerous sugars to surpass other more interesting structural elements of NPs, particularly the structural characteristics of central scaffolds, the sugar units were removed before the actual scaffold analysis. The removal was done by a recursive deglycosylation procedure described in detail in [9]. In the end, the cleaning procedure described above generated 35 482 unique aglycons for the Taiwan TCM database and 24 836 aglycons for the JNP database.

Scaffold analysis

The major part of our analysis was comparison of scaffolds between the two databases. The scaffold is clearly the most important part of the molecule giving it shape, determining whether the molecule is rigid or flexible, and keeping substituents in their proper positions. The structure of the scaffold considerably influences global molecular properties and in many cases also the biological activity of the parent molecule. Many important medicinal chemistry techniques are based on scaffolds; typical examples are scaffold hopping or combinatorial chemistry. The term "scaffold" is therefore used very often in the medicinal chemistry literature. However, it is used rather freely, without a clearly defined meaning. The exact interpretation of this term varies from publication to publication and depends also on the particular application area. This nomenclature ambiguity therefore necessitates a clear definition of the term "scaffold" in every application. Throughout this article, the term "scaffold" is used to describe the part of the molecule that remains after removal of all non-ring substituents, however, keeping exocyclic and exochain multiple bonds.

Scaffold analysis of both databases produced 12 785 scaffolds for the TCM database and 8726 scaffolds for the JNP database. At the same time, we also analyzed the distribution of what we call "simple rings", i.e., substructures consisting of a single ring system only, including as well exocyclic double bonds. The simple rings might originally be part of a larger fused system. In this case, all unique rings from a complex fused system were extracted. The analysis generated 1404 unique rings for the TCM database and 1672 unique rings for the JNP database.

We were interested in identifying scaffolds and rings that exhibited the most significant differences in their distribution between the two databases. In order to analyze the data as well as rationalize the results, we used two methods for visualization of the chemical space, namely the scaffold tree [8] and the recently published scaffold map [11].

The scaffold tree methodology has been described in detail in [8], therefore, here we only provide a brief overview. The method is based on the generation of a scaffold hierarchy for every molecule in the dataset. First the molecule is reduced to its scaffold. Then rings from the scaffolds are removed one by one starting from the less important rings at the periphery until finally a single "root" ring is retained. Which ring is less important and will be removed first is decided based on a set of simple rules, fully described in [8]. Basically, ring systems containing more heteroatoms or nonstandard structural features (like spiro centers or nonstandard fusing) are considered more important and are retained. After all molecules are processed, they are grouped together based on their "root" rings. All molecules having the same root form a single branch of a tree and all branches together form a "scaffold tree". If the database is very large, not all scaffolds can be displayed in the graph (the scaffolds are too numerous). In this study, we only display scaffolds that are present in at least 0.05 % of the molecules from the original data set.

The scaffold map [11] method visualizes molecules in a form of two-dimensional graph. Molecules (in our case scaffolds and rings) are sorted based on set of simple descriptors termed scaffold keys. These descriptors are designed to mimic the intuitive classification of scaffolds by medicinal chemists.

In both scaffold tree and scaffold map diagrams, molecules can be colored to add an additional level of information. In our analysis, the scaffolds and rings are color coded according to their propensity in the original databases. Substructures that are at least ten times more frequent in the JNP database than in the TCM database are highlighted in green, and the structures that are at least ten times more frequent in the TCM database are highlighted in yellow.

Acknowledgements

We thank Judith M. Rollinger for helpful discussions.

Conflict of Interest

The authors declare there is no conflict of interest regarding this study.

References

1 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the last 25 years. J Nat Prod 2007; 70: 461–477

2 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod 2012; 75: 311–335

3 *Chen CY.* TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening *in silico*. PLoS One 2011; 6: e15939

4 *López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL.* Expanding the medicinally relevant chemical space with compound libraries. Drug Discov Today 2012; 17: 718–726

5 *Shen M, Tian S, Li Y, Li Q, Xu X, Wang J, Hou T.* Drug-likeness analysis of traditional Chinese medicines: 1. property distributions of drug-like compounds, non-drug-like compounds and natural compounds from traditional Chinese medicines. J Cheminform 2012; 4: 31

6 *Ertl P, Roggo S, Schuffenhauer A.* Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 2008; 48: 68–74

7 *Vanii K, Moreno P, Truszkowski A, Ertl P, Steinbeck C.* Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinform 2012; 13: 106

8 *Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H.* The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 2007; 47: 47–58

9 *Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H.* Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci U S A 2005; 102: 17272–17277

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

11 *Ertl P.* Intuitive ordering of scaffolds and scaffold similarity searching using scaffold keys. J Chem Inf Model 2014; 54: 1617–1622

12 *Driggers EM, Hale SP, Lee J, Terrett NK.* The exploration of macrocycles for drug discovery–an underexploited structural class. Nat Rev Drug Discov 2008; 7: 608–624

13 *Giordanetto F, Kihlberg J.* Macrocyclic drugs and clinical candidates: what can medicinal chemists learn from their properties? J Med Chem 2014; 57: 278–295

14 *Marsault E, Peterson ML.* Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. J Med Chem 2011; 54: 1961–2004

15 *Wessjohann LA, Ruijter E, Garcia-Rivera D, Brandt W.* What can a chemist learn from nature's macrocycles? – A brief, conceptual view. Mol Divers 2005; 9: 171–186

16 *Walker S, Landovitz R, Ding WD, Ellestad GA, Kahne D.* Cleavage behavior of calicheamicin gamma 1 and calicheamicin T. Proc Natl Acad Sci U S A 1992; 89: 4608–4612

17 *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 2011; 40: D1100–D1107